УДК УДК 614.2:004:303.4
МРНТИ 76.75.75, 12.41.55

МНОЖЕСТВЕННАЯ ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ ДЛЯ БИНАРНОЙ ПЕРЕМЕННОЙ ОТКЛИКА В ПРОГРАММНОЙ СРЕДЕ R – КРАТКИЕ 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

Егошин В.Л. – https://orcid.org/0000-0002-8407-3789
Саввина Н.В. – https://orcid.org/0000-0003-2441-6193
Гржибовский А.М. – https://orcid.org/000-0002-5464-0498

Citation/
Библиографиялық сілтеме/
Библиографическая ссылка:

Egoshin VL, Savvina NV, Grjibovski AM. Multivariable logistic regression in r: brief guidelines for master and doctoral students in medicine and public health. West Kazakhstan Medical Journal 2019 June; 61(2):83–90.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. R бағдарламалық ортасында бинарлы құбылмалы жауап үшін бірнеше логистикалық регрессия – «медицина» және «қоғамдық денсаулық сақтау» мамандықтары бойынша магистранттар мен докоранттарға арналған қысқаша ұсыныстар. West Kazakhstan Medical Journal 2019 June; 61(2):83–90.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. Множественная логистическая регрессия для бинарной переменной отклика в программной среде r – краткие специальности «медицина» и «общественное здравоохранение». West Kazakhstan Medical Journal 2019 June; 61(2):83–90.

Multivariable logistic 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]
[1]Northern State Medical University, Arkhangelsk, Russia
[2]North-Eastern Federal University, Yakutsk, Russia
[3]Al-Farabi Kazakh National Medical University, Almaty, Kazakhstan

In this paper we describe basic principles of using R package for multivariable logistic regression analysis for binary dependent variable. We present step-by-step guidelines and syntax for creation and evaluation of regression models using practical example with real data from an earlier published study on cervical cancer screening. In addition to the syntax we present original R outputs and their interpretation.
Keywords: *R, logistic regression, syntax, listing, modeling, validation.*

R бағдарламалық ортасында бинарлы құбылмалы жауап үшін бірнеше логистикалық регрессия – «Медицина» және «Қоғамдық денсаулық сақтау» мамандықтары бойынша магистранттар мен докоранттарға арналған қысқаша ұсыныстар

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 мақалада бірнеше логистикалық регрессиялы анализді жүзеге асыру үшін R бағдарламалық ортасын қолданудың негізгі принциптері ұсынылған. Жатыр мойны қатерлі ісігіне скрининг мәліметтерін қолдана отырып тәжірибелік мысал түрінде регрессиялық моделдерді жасау және бағалау үшін R-де қадамдық алгоритм мен синтаксис ұсынылды. Синтаксистен бөлек нәтижелер оларды R, сондай-ақ олардың интерпретациясы бергеніндей түрде ұсынылған.
Негізгі сөздер: *R, логистикалық регрессия, синтаксис, листинг, моделдеу, валидация.*

Множественная логистическая регрессия для бинарной переменной отклика 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 логистического регрессионного

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

Received/
Келіп түсті/
Поступила:
17.06.2019

Accepted/
Басылымға қабылданды/
Принята к публикации:
28.06.2019

ISSN 1814-5620 (Print)

Published by West Kazakhstan Marat Ospanov Medical University

анализа. Представлен пошаговый алгоритм и синтаксис в R для создания и оценки регрессионных моделей в виде практического примера с использованием данных скрининга на рак шейки матки. Помимо синтаксиса представлены результаты в том виде, как их выдает R, а также их интерпретация.
Ключевые слова: *R, логистическая регрессия, синтаксис, листинг, моделирование, валидация.*

Математическое моделирование является важной частью аналитического процесса в медицине и общественном здравоохранении. Среди применяемых статистических техник широко используется метод логистической регрессии, а умение применять этот метод рассматривается как важная компетенция исследователя [1-4]. Интересно отметить, что во многих странах подразумевается, что владеть им должны не только соискатели степени доктора философии (PhD), но и будущие магистры общественного здоровья (MPH). В казахстанской медицинской науке применение многомерных методов статистики распространено значительно реже, но доля публикаций, в которых применяются многомерные методы статистики, в последние годы стабильно увеличивается.

Логистический регрессионный анализ наиболее часто используется при создании многомерных моделей для бинарных исходов. На основании этих моделей делается прогноз для данных, не включенных в первоначальную модель. Логистическая регрессия используется исследователями для 1) предсказания вероятности равенства 1 (единице) значения зависимой переменной (вероятность принадлежности к группе с произошедшим событием или изучаемым признаком); 2) классификации переменных; 3) изучения отношения шансов возникновения исхода, связанных с независимыми переменными модели [5]. Метод предполагает использование в качестве зависимой переменной только категориальной величины (со значениями 1 или 0 в случае логистической бинарной регрессии) с оценкой ее вероятности и классификацией получаемых результатов. Использование метода предполагает построение логистической регрессионной модели, её оценку и проведение классификационного анализа, включая построение ROC кривой, определение наилучшего порогового значения, создание и оценку матрицы неточностей.

К основным условиям (ограничениям) применения логистического регрессионного анализа относятся следующие: наличие категориальной зависимой переменной, независимость наблюдений, отсутствие мультиколлинеарности (тесной корреляционной связи между независимыми переменными), а также линейная зависимость между независимой переменной и натуральным логарифмом отношения шансов. Параметры, оценивающие модель логистической регрессии, включают информационные критерии Акайке (AIC – Akaike's information criterion) и Байеса (BIC – Bayesian information criterion).

Модель бинарной логистической регрессии использует логистическую функцию для получения значения между 0 и 1.

Логистическая функция определяется как

$$\text{logistic}(x) = \frac{1}{1 + \exp(-x)}$$

И выглядит так (рис.1)

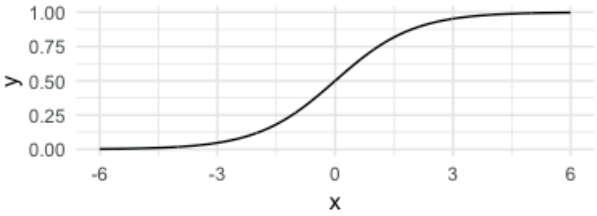


Рис.1 График функции логистической регрессии

В линейной регрессионной модели связь между предикторами и переменной отклика определяется линейным уравнением:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Логистическая бинарная регрессия, решая задачу классификации, оценивает вероятность равенства зависимой переменной 1 (единице):

$$P(y_i = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))}$$

Обычно исследователей интересует вопрос: какой объем выборки должен быть для применения того или иного метода статистической обработки данных? Работающие с регрессионными моделями используют термин "количество событий на переменную" (Events Per Variable (EPV). Процедура расчета размера выборки для логистической регрессии достаточно трудоемка и большинство исследователей применяет статистические пакеты, например, Sample Power, PASS и другие для расчета выборки, но есть одно простое правило, которое может быть рекомендовано начинающим исследователям. Многие пособия по статистике рекомендуют иметь не менее 10 исходов на одну зависимую переменную, однако, некоторые авторы рекомендуют более консервативный подход и предлагают не менее 20 исходов на каждую переменную [6].

Интерпретация коэффициентов логистической регрессии отличается от оценки их при линейной регрессии, поскольку оценивается вероятность

отклика, которая может принимать значения только в диапазоне между 0 и 1. Коэффициенты не влияют на вероятность линейно. Преобразуем уравнение логистической функции:

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \ln\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Модель логистической регрессии является линейной моделью для натурального логарифма отношения шансов:

$$\frac{P(y=1)}{1-P(y=1)} = odds = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

Одним из результатов логистического регрессионного анализа является расчёт регрессионных коэффициентов. Регрессионный коэффициент показывает изменение натурального логарифма отношения шансов изучаемой зависимой переменной при изменении независимой переменной на единицу при постоянных значениях всех остальных независимых переменных. Поскольку логарифм отношения шансов сложно интерпретировать, можно потенцировать коэффициенты, чтобы представить их в виде отношения шансов. Потенцированный коэффициент показывает, во сколько раз увеличивается отношение шансов при увеличении независимой переменной на единицу.

При оценке доверительного интервала для коэффициентов логистической регрессии необходимо учитывать наличие единицы в интервале. Если доверительный интервал включает в себя единицу, то это указывает на отсутствие статистически значимого влияния данного коэффициента на исход (на выбранном уровне доверительной вероятности). Обычно выбирают уровень доверительной вероятности 95% и рассчитывают 95% доверительный интервал.

Оценка модели логистической регрессии выполняется с точки зрения её пригодности для классификации. Модель логистической регрессии оценивает вероятность признака, что позволяет построить матрицу неточностей (confusion matrix), в которой в виде таблицы сопряженности будут представлены переменные с реальным значением признака (референтные значения) и с предполагаемым значением (таб.1). В SPSS матрица неточностей носит название классификационной таблицы.

Та.1 Матрица неточностей

Спрогнозированные значения	Фактические значения	
	Есть исход	Нет исхода
Есть исход	TP	FP
Нет исхода	FN	TN

- **TP** истинные положительные решения
- **TN** истинные отрицательные решения
- **FP** ложные положительные решения
- **FN** ложные отрицательные решения

Для оценки выбранной модели в классификационном анализе используются такие метрики как безошибочность, точность, чувствительность, специфичность, F-мера и другие.

Безошибочность (accuracy) – отношение суммы всех правильных результатов к количеству всех случаев $(TP + TN)/(TP + TN + FP + FN)$.

Точность (precision, прогностическая ценность положительного решения) определяется как доля истинно-положительных решений среди всех положительных решений $TP/(TP + FP)$.

Чувствительность (sensitivity, recall, полнота) – это доля правильно определенных истинно-положительных решений $TP/(TP + FN)$.

Специфичность (specificity) – это доля правильно определенных истинно-отрицательных решений $TN/(TN + FP)$.

Показатель F-меры может быть рассчитан на основании показателей точности и полноты по формуле

$$2 \cdot \frac{Precision \times Recall}{Precision + Recall}$$

Этот показатель (с теоретическим максимальным значением равным 1) стремится к нулю, если точность или полнота стремится к нулю.

Отнесение признака к тому или иному классу выполняется на основании вероятности признака, значение величины вероятности, по которому проводится разделение, понимается как порог (threshold) решающего правила. Порог решающего правила, называемый также Youden's J statistic или Youden's index – это значение вероятности, при котором одна часть участников исследования имеет изучаемый признак, представленный зависимой переменной, а другая – не имеет. При использовании ROC-кривой этот показатель рассчитывается для каждой её точки. Оптимальным значением порога считается та его величина, при котором достигается наилучший баланс между чувствительностью и специфичностью модели.

Качество бинарной классификации позволяет оценить график ROC (receiver operating characteristic) кривой, отображающий соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак, (true positive rate, TPR, называемой чувствительностью алгоритма классификации) и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак (false positive rate, FPR, величина 1-FPR называется специфичностью алгоритма классификации) при изменении порога решающего правила (рис 2).

ROC (receiver operating characteristic) кривая, которую в некоторых пособиях называют характеристической кривой, отображает соотношение долей объектов от общего количества носителей признака, верно классифицированных как несущий признак

(true positive rate, TPR) и долей объектов от общего количества объектов, не несущих признака (false positive rate, FPR), при изменении порога решающего правила. Количественной характеристикой ROC-кривой является AUC (area under the curve) – площадь под кривой. Чем ближе значение AUC к единице, тем лучше модель классифицирует данные. AUC для ROC-кривой на рисунке 2 равна 0,918, что является хорошим результатом для медицинских исследований.

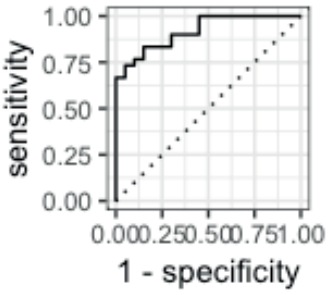


Рис.2 ROC кривая

Практический пример: оценка факторов риска развития рака шейки матки с применением метода логистической регрессии.

Рак шейки матки является важной проблемой общественного здравоохранения. Во многих странах работают программы скрининга на предмет раннего обнаружения заболевания. Особый интерес представляет оценка факторов риска данной патологии. Далее будет показано, как в программной среде R создать модель логистической регрессии, оценить эту модель, улучшить её и на основании улучшенной модели выполнить классификационный анализ. В работе используются базовые пакеты R 3.5.3, а также пакеты dplyr, ggplot2 и другие, входящие в пакет tidyverse, а также пакеты broom, pROC, caret. Работа выполнена в RStudio v. 1.1.463.

Для демонстрации использованы данные исследования, в котором проводилась оценка факторов риска заболевания и результаты инструментальных обследований. Данные загружены с открытого ресурса UCI Machine Learning repository и были ранее описаны Fernandes, Cardoso, and Fernandes [7]. При подготовке данных выполнено преобразование названий столбцов (функция clean_names пакета janitor). В исходном файле отсутствующие значения были помечены вопросительным знаком, произведена их замена на значения NA. Затем были отобраны переменные для включения в модель и изменен тип некоторых переменных. После этого были удалены строки, содержащие NA значения. Для показа включенных в датафрейм переменных использована функция skim_to_wide пакета skimr (листинг 1).

Листинг 1

```
raw_data <- read_csv("../data/risk_factors_cervical_
cancer.csv")
raw_data <- janitor::clean_names(raw_data)
cervical <- raw_data %>%
mutate_if(is.character, list(~ as.numeric(na_if(., "?")))) 
%>%
 select(c(1:6, 8:13, 36)) %>%
 mutate(
 smokes = factor(smokes, levels = c(0, 1)),
 hormonal_contraceptives = factor(hormonal_
contraceptives, levels = c(0, 1)),
 iud = factor(iud, levels = c(0, 1)),
 st_ds = factor(st_ds, levels = c(0, 1)),
 biopsy = factor(biopsy, levels = c(0, 1))
 ) %>%
 na.omit()
cervical %>%
 select_if(is.numeric) %>%
 skimr::skim_to_wide() %>%
 select(-c(1, 3:5)) %>%
 knitr::kable(caption = "Таб.2 Непрерывные
переменные")
cervical %>%
 select_if(is.factor) %>%
 skimr::skim_to_wide() %>%
 select(-c(1, 3:5)) %>%
knitr::kable(caption = "Таб.3 Категориальные
переменные")
```

Таб.2 Непрерывные переменные

variable	mean	sd	p0	p25	p50	p75	p100
age	27.26	8.73	13	21	26	33	84
first_sexual_intercourse	17.14	2.85	10	15	17	18	32
hormonal_contraceptives_years	2.29	3.72	0	0	0.5	3	22
iud_years	0.53	2	0	0	0	0	19
num_of_pregnancies	2.32	1.47	0	1	2	3	11
number_of_sexual_partners	2.2	1.64	1	2	2	3	28
smokes_years	1.24	4.19	0	0	0	0	37
st_ds_number	0.17	0.55	0	0	0	0	4

Для изучения отобраны 668 записей. Количество событий на переменную составляет 56, что дает достаточную статистическую мощность для планируемой процедуры моделирования.

Как предикторы будут использованы такие переменные как возраст, число половых партнеров, возраст первого полового контакта, количество беременностей, курение, использование гормональных контрацептивов и внутриматочных средств, случаи ЗППП и их количество. В качестве переменной отклика – результаты биопсии (есть или нет рак шейки матки). В листинге 2 представлено создание исходной модели и её оценка. Функция создания модели линейной регрессии для данного примера glm(biopsy ~ ., data = cervical, family = binomial). Формула biopsy ~. указывает, что переменной отклика является biopsy, точка указывает на то, что в качестве предикторов выступают все остальные переменные в наборе данных.

Таб.3 Категориальные переменные

variable	n_unique	top_counts	ordered
biopsy	2	0: 623, 1: 45, NA: 0	FALSE
hormonal_ contraceptives	2	1: 430, 0: 238, NA: 0	FALSE
iud	2	0: 593, 1: 75, NA: 0	FALSE
smokes	2	0: 572, 1: 96, NA: 0	FALSE
st_ds	2	0: 603, 1: 65, NA: 0	FALSE

Для представления данных о созданной модели использованы функции glance и tidy пакета broom. Функция tidy использована в виде tidy(conf.int = TRUE, exponentiate = TRUE), первый параметр позволяет вывести показатели доверительных интервалов для коэффициентов, второй параметр выводит потенцированные значения коэффициентов (столбец estimate) и границ доверительных интервалов. Рис.3 дает графическое представление о коэффициентах и доверительных интервалах модели.

Листинг 2

```
fit <- glm(biopsy ~ ., data = cervical, family = binomial)
fit %>%
 glance() %>%
 knitr::kable(caption = "Таб.4 Параметры исходной модели")
        fit %>%
 tidy(conf.int = TRUE, exponentiate = TRUE) %>%
 knitr::kable(caption = "Таб.5 Коэффициенты исходной модели", digits = 4)
        fit %>%
 tidy(conf.int = TRUE, exponentiate = TRUE) %>%
 filter(term != "(Intercept)") %>%
 mutate(term = fct_reorder(term, conf.low, min)) %>%
 ggplot(aes(term, estimate, ymin = conf.low, ymax = conf.high)) +
 geom_point() +
 geom_errorbar(width = .5) +
        geom_hline(yintercept = 1, linetype = "dotted") +
```

Таб.4 Параметры исходной модели

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
329.44	667	-157.5963	341.1927	399.7484	315.1927	655

Таб.5 Коэффициенты исходной модели

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.0215	1.1136	-3.4496	0.0006	0.0025	0.1979
age	0.9945	0.0275	-0.2010	0.8407	0.9389	1.0458
number_of_sexual_partners	0.9792	0.1176	-0.1789	0.8580	0.7549	743
first_sexual_intercourse	1.0540	0.0645	0.8153	0.4149	0.9241	1.1912
num_of_pregnancies	1.0771	0.1374	0.5405	0.5888	0.8195	1.4065
smokes1	1.0457	0.6593	0.0677	0.9460	0.2577	3.5148
smokes_years	1.0134	0.0501	0.2665	0.7898	0.9095	1.1123
hormonal_contraceptives1	0.8061	0.3871	-0.5567	0.5777	0.3777	1.7410
hormonal_contraceptives_years	1.0779	0.0414	1.8100	0.0703	0.9906	1.1675
iud1	1.7502	0.6257	0.8945	0.3710	0.4809	5.7198
iud_years	1.0054	0.1001	0.0541	0.9569	0.8048	1.2022
st_ds1	1.9648	0.9884	0.6833	0.4944	0.2551	13.0168
st_ds_number	1.2278	0.5112	0.4014	0.6881	0.4242	3.2796

expand_limits(y = 0) +
coord_flip() (Рис.3)

Полученные результаты показывают на отсутствие значимых регрессионных коэффициентов в такой модели.

Модель можно упростить и оценить снова (листинг 3). Для создания новой модели использована функция stepAIC из пакета MASS. В программе происходит создание и оценка моделей на значениях показателя AIC. Модель с наименьшим значением AIC считает наиболее подходящей. Показатели полученной модели изучены подобно тому, как это делалось с исходной моделью.

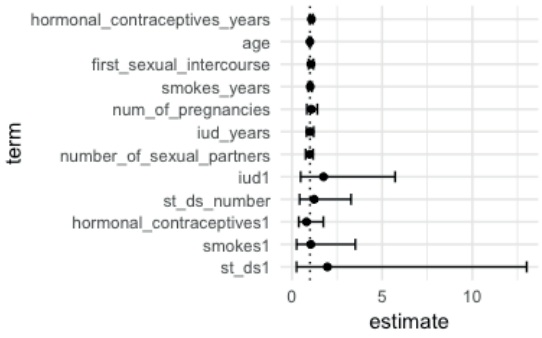


Рис.3 Коэффициенты исходной модели

Листинг 3

```
MASS::stepAIC(fit, direction = "both")
fit1 <- glm(
 formula = biopsy ~ hormonal_contraceptives_years +
st_ds,
 family = binomial, data = cervical
)
fit1 %>%
 glance() %>%
 knitr::kable(caption = "Таб.6 Параметры улучшенной
модели")
fit1 %>%
 tidy(conf.int = TRUE, exponentiate = TRUE) %>%
 knitr::kable(caption = "Таб.7 Коэффициенты
улучшенной модели", digits = 4)
```

Результаты, полученные при использовании новой модели, показывают следующее:

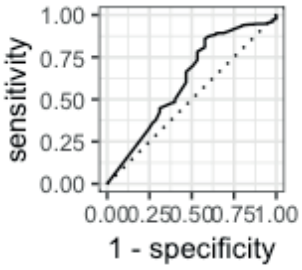
1. Наличие инфекций, передаваемых половым путем, увеличивает шансы иметь рак шейки матки в 3,01 раза при 95% доверительном интервале 1,3458-6,2430. В статьях, правда, достаточно указывать два знака после запятой.
2. Каждый год использования гормональных контрацептивов увеличивает шансы иметь заболевание в 1,07 раза при 95% доверительном интервале 1,0024 – 1,1441.

Для оценки модели методами классификационного анализа с использованием пакета pROC будет построена ROC кривая, вычислена площадь под кривой (AUC), создана матрица неточностей, для которой будут вычислены показатели безошибочности, точности, полноты и F-меры (листинг 4). Первоначально создадим датафрейм, содержащий переменную с вероятностями переменной отклика. Для этого используем функцию augment пакета broom, введя параметр type.predict = «response». Затем создадим объект класса roc для выполнения дальнейших вычислений. Для этого объекта может быть вычислена площадь под кривой и вычислено наилучшее пороговое значение. Для создания матрицы неточностей создадим переменную со значениями 0 и 1, предсказанными на основании показателей вероятности признака. Параметры оценки матрицы неточностей вычислим, используя функции пакета caret.

Листинг 4

```
fit_1 <- augment(fit1, type.predict = "response")
# roc object
roc_data <- roc(biopsy ~ .fitted, data = fit_1)

yardstick::roc_curve(fit_1, biopsy, .fitted) %>% au-
toplot()
```



Таб.6 Параметры улучшенной модели

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
329.6844	667	-159.4544	324.9087	338.4216	318.9087	665

Таб.7 Коэффициенты улучшенной модели

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.0502	0.2088	-14.3278	0.0000	0.0326	0.0741
hormonal_contraceptives_years	1.0751	0.0334	2.1706	0.0300	1.0024	1.1441
st_ds1	3.0111	0.3876	2.8439	0.0045	1.3458	6.2430

```
        auc(roc_data) # roc auc
        ## Area under the curve: 0.6133
        # threshold
(thr <- coords(roc_data, "best"))
        ## threshold specificity sensitivity
## 0.0850914 0.8587480 0.4222222
        fit_1_matrix <- fit_1 %>%
 mutate(preds = as.factor(ifelse(.fitted <= thr[1], 0, 1)))
%>%
 select(preds, biopsy)
conf_matr <- with(fit_1_matrix, table(preds, biopsy))
%>%
 DescTools::Rev()
confusionMatrix(conf_matr)$overall[[1]]
        ## [1] 0.8293413
        precision(conf_matr)
        ## [1] 0.1775701
        recall(conf_matr)
        ## [1] 0.4222222
        F_meas(conf_matr)
        ## [1] 0.25
        specificity(conf_matr)
        ## [1] 0.858748
        conf_matr %>%
 knitr::kable(caption = "Таб.8 Матрица неточностей
улучшенной модели")
```

Таб.8 Матрица неточностей улучшенной модели

	1	0
1	19	88
0	26	535

```
tibble(
 parameter = c("Accuracy", "Precision", "Recall", "Spec-
ificity", "F measure"),
 value = c(
 confusionMatrix(conf_matr)$overall[[1]],
 precision(conf_matr),
 recall(conf_matr),
 specificity(conf_matr),
 F_meas(conf_matr)
 )
) %>%
 knitr::kable(
 caption = "Таб.9 Параметры оценки модели",
 align = c("l", "c"),
 digits = 3
 )
```

Таб.9 Параметры оценки модели

parameter	value
Accuracy	0.829
Precision	0.178
Recall	0.422
Specificity	0.859
F measure	0.250

Результаты убедительно показывают, что такую модель невозможно использовать в клинической практике по причине недостаточно высоких показателей модели, особенно чувствительности (0,422) и предсказательной ценности положительного результата (0,178).

Для валидационной оценки модели разделим исходные данные на тренировочный и тестовый сеты (выборки), используем улучшенную модель логистической регрессии, вычислим значения AUC для данных обоих сетов и сравним их, используя Delongi's тест (листинг 5).

Листинг 5

```
        # создание тренировочного и тестового сета
set.seed(123)
ind <- createDataPartition(cervical$biopsy, p = .7, list
= FALSE)
train_data <- cervical[ind, ]
test_data <- cervical[-ind, ]

# используемая модель
fit_train <- glm(
 formula = biopsy ~ hormonal_contraceptives_years +
st_ds,
 family = binomial, data = train_data)

# данные для создания roc объекта, основанного на
тренировочном сете
fit_train_data <- augment(fit_train, type.predict = «re-
sponse»)

# данные для создания roc объекта, основанного на
тестовом сете
fit_test_data <- augment(fit_train, newdata = test_data,
type.predict = «response»)

# roc objects
# roc объект, созданный на тренировочных данных
roc_train_data <- roc(biopsy ~ .fitted, data = fit_train_
data)

# roc объект, созданный на тестовых данных
roc_test_data <- roc(biopsy ~ .fitted, data = fit_test_data)

# тест для сравнения полученных на разных данных
объектов
(delong <- roc.test(roc_train_data, roc_test_data))
        ##
## DeLong's test for two ROC curves
##
## data: roc_train_data and roc_test_data
## D = 0.19543, df = 313.39, p-value = 0.8452
## alternative hypothesis: true difference in AUC is not
equal to 0
## sample estimates:
```

Dependent: biopsy		0	1	OR (univariable)	OR (multivariable)
age	Mean (SD)	27.1 (8.7)	29.1 (9.2)	1.02 (0.99-1.05, p=0.147)	1.01 (0.96-1.05, p=0.710)
hormonal_contraceptives_years	Mean (SD)	2.2 (3.5)	3.4 (5.6)	1.07 (1.00-1.14, p=0.034)	1.06 (0.99-1.13, p=0.101)
iud	0	57(89.4)	36(80.0)	-	-
	1	6 (10.6)	9 (20.0)	2.11 (0.92-4.40, p=0.059)	1.72 (0.71-3.83, p=0.200)
num_of_pregnancies	Mean (SD)	2.3 (1.5)	2.6 (1.3)	1.12 (0.92-1.34, p=0.229)	1.02 (0.80-1.30, p=0.874)
smokes	0	535(85.9)	37(82.2)	-	-
	1	88(14.1)	8 (17.8)	1.31 (0.55-2.78, p=0.501)	1.15 (0.47-2.51, p=0.748)
st_ds	0	568(91.2)	35(77.8)	-	-
	1	55 (8.8)	10(22.2)	2.95 (1.32-6.09, p=0.005)	2.85 (1.25-6.00, p=0.008)

```
## AUC of roc1 AUC of roc2
## 0.6175272 0.5947064
```

AUC для тренировочного сета 0,618, AUC для тестового сета 0,595. Результаты Delong's теста для двух ROC кривых: p-value = 0,845, что не позволяет отклонить нулевую гипотезу и говорит о хорошей внешней валидности исследования.

Подобный результат исследования ожидаем, так как методы, связанные с оценкой рисков, очень часто высокоспецифичны и малочувствительны. При подобных обстоятельствах возможности истинно-положительных ответов ограничены.

Функции пакета finalfit [8] позволяют получать таблицы с результатами регрессионного анализа относительно простым путем. Необходимо указать независимые (explanatory) и зависимую (dependent) переменные и применить функцию finalfit (листинг 6). Результатом выполнения функции будут таблицы, включающие нескорректированные (crude) и скорректированные (adjusted) коэффициенты с доверительными интервалами, а также достигнутые уровни значимости (p-value).

Листинг 6

```
library(finalfit)
explanatory <- c(
 "age", "hormonal_contracepti,ves_years", "iud",
 "num_of_pregnancies", "smokes", "st_ds"
)
dependent <- "biopsy"

cervical %>%
 finalfit(dependent, explanatory) %>%
 knitr::kable()
```

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 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.

Список литературы / References:

1. Moyé L. Statistical Methods for Cardiovascular Researchers. Circulation Research 2016;118(3):439–453.
2. Bertolaccini L, Pardolesi A, Solli P. The biostatistical minimum. Journal of Thoracic Disease 2017;9(10):4131–4132.
3. Hayat MJ, Powell A, Johnson T, Cadwell BL. Statistical methods used in the public health literature and implications for training of public health professionals. PLoS ONE 2017;12(6):1–10.
4. Kaplan D. Teaching Stats for Data Science." American Statistician 2018;72(1):89–96.
5. Hilbe JM. Practical Guide to Logistic Regression. CRC Press, 2015.
6. Smeden M, Reitsma JB, Eijkemans MJC, Moons KGM, Collins GS, Altman DG, de Groot JAH. Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical Methods in Medical Research 2018, 096228021878472.
7. Fernandes K, Cardoso JC, Fernandes J. Transfer learning with partial observability applied to cervical cancer screening. Iberian Conference on Pattern Recognition and Image Analysis. Springer 2017;243–50.
8. Harrison E. Tables Gallery. https://finalfit.org/articles/tables_gallery.html. Accessed 20 April 2019.