УДК 614.2:004.657(072)
МРНТИ 14.35.07

КВАНТИЛЬНАЯ РЕГРЕССИЯ В ПРОГРАММНОЙ СРЕДЕ R: КРАТКИЕ РЕКОМЕНДАЦИИ ДЛЯ ДОКТОРАНТОВ ПО СПЕЦИАЛЬНОСТИ «МЕДИЦИНА» И «ОБЩЕСТВЕННОЕ ЗДОРОВЬЕ»

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

[1]Северный государственный медицинский университет, Архангельск, Россия
[2]Северо-восточный федеральный университет, Якутск, Россия
[3]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

For citing/
библиографиялық сілтеме/
библиографическая ссылка:

Egoshin VL, Savvina NV, Grjibovski AM. Quantile regression in R: brief guidelines for PhD students in medicine and public health. West Kazakhstan Medical Journal 2019;61(4):194–202.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. R бағдарламалық ортадағы кванттық регрессия: «Медицина» және «Қоғамдық денсаулық» мамандықтары бойынша докоранттарға арналған қысқаша ұсынымдар. West Kazakhstan Medical Journal 2019;61(4):194–202.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. Квантильная регрессия в программной среде R: краткие рекомендации для докторантов по специальности «Медицина» и «Общественное здоровье». West Kazakhstan Medical Journal 2019;61(4):194–202.

Quantile regression in R: brief guidelines for PhD students in medicine and public health
V.L. Egoshin[1], N.V. Savvina[2], A.M. Grjibovski[1-3]
[1]Northern State Medical University, Arkhangelsk, Russia
[2]North-Eastern Federal University, Yakutsk, Russia
[3]Al-Farabi Kazakh National University, Almaty, Kazakhstan

In this paper we describe basic principles of using R package for multivariable quantile regression analysis. We present step-by-step guidelines and syntax for creation and evaluation of quantile regression models using practical example. In addition to the syntax we present R outputs and their interpretation.
Keywords: *R, quantile regression, syntax, listing, modeling.*

R бағдарламалық ортадағы кванттық регрессия: «Медицина» және «Қоғамдық денсаулық» мамандықтары бойынша докоранттарға арналған қысқаша ұсынымдар
В.Л. Егошин[1], Н.В. Саввина[2], А.М. Гржибовский[1-3]
[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 кванттық регрессиялық анализді жүзеге асыруға арналған R бағдарламалық ортасын қолданудың негізгі принциптері ұсынылған. Тәжірибелік мысал түрінде кванттық регрессиялық модельдерді құру және бағалау үшін R-де қадамдық алгоритм мен синтаксис ұсынылған. Синтаксистен басқа R және олардың интерпретациясы көрсететіндей нәтижелер ұсынылған.
Негізгі сөздер: *R, квантильді регрессия, синтаксис, листинг, моделдеу.*

Квантильная регрессия в программной среде R: краткие рекомендации для докторантов по специальности «Медицина» и «Общественное здоровье»
В.Л. Егошин[1], Н.В. Саввина[2], А.М. Гржибовский[1-3]
[1]Северный государственный медицинский университет, Архангельск, Россия
[2]Северо-восточный федеральный университет, Якутск, Россия
[3]Казахский Национальный университет им. Аль-Фараби, Алматы, Казахстан

В данной работе представлены основные принципы применения программной среды R для осуществления квантильного регрессионного анализа. Представлен пошаговый алгоритм и синтаксис в R для создания и оценки квантильных регрессионных моделей в виде практического примера. Помимо синтаксиса представлены результаты в том виде, как их выдает R, а также их интерпретация.
Ключевые слова: *R, квантильная регрессия, синтаксис, листинг, моделирование.*

Введение.

Ранее на страницах West Kazakhstan Medical Journal мы познакомили читателей с методом множественного (многомерного) линейного регрессионного анализа, с помощью которого можно оценить связь между несколькими независимыми переменными

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

Received/
Келіп түсті/
Поступила:
09.12.2019

Accepted/
Басылымға қабылданды/
Принята к публикации:
23.12.2019

ISSN 1814-5620 (Print)

Published by West Kazakhstan Marat Ospanov Medical University

(предикторами, факторными признаками) и зависимой переменной (переменной отклика, результативным признаком), которая представлена непрерывной величиной. Однако, применяя метод линейной регрессии, всегда следует помнить о его ограничениях, коих достаточно много, что ограничивает область применения линейной регрессии в биомедицинских исследованиях. Одним из основных условий, которое должно выполняться для того, чтобы результаты могли считаться валидными, это нормальное распределение остатков. Если данное условие не выполняется, а оно относительно часто не выполняется, когда результативный признак (зависимая переменная) имеет выраженное скошенное распределение, следует использовать альтернативные методы анализа, одним из которых является квантильная регрессия, которая относится к группе непараметрических методов, которые устойчивы к отклонению данных от заданного распределения. Квантильная регрессия в настоящее время находит применение в эконометрике (1-4), есть примеры ее использования в анализе дожития (5-8), однако этот метод может применяться для очень большого числа биомедицинских задач, когда мы имеем дело с сильно скошенной зависимой количественной переменной.

Квантильная регрессия – непараметрический метод, представляющий собой процедуру оценки параметров линейной связи между независимыми переменными и заданным уровнем квантиля зависимой переменной. Квантильная регрессия позволяет получить параметры регрессии для любых квантилей распределения зависимой переменной. Это позволяет модели быть значительно менее чувствительной к выбросам (выскакивающим значениям) и к нарушениям предположений о характере распределений остатков. Метод квантильной регрессии рассматривается как качественное дополнение к классической регрессии. Это обусловлено возможностями квантильной регрессии оценить степень различия влияния факторов вдоль условных распределений зависимой переменной, выявить гетероскедастичность и асимметрию распределения ошибок.

Линейная регрессия предполагает линейную зависимость среднего арифметического значения зависимой переменной от изменения независимых переменных (предикторов). При медианной регрессии предполагается линейная зависимость условной медианы зависимой переменной от независимых переменных. При квантильной регрессии исследователь отказывается от моделирования среднего, как это было в классической линейной модели, и может моделировать любой квантиль распределения зависимой переменной.

Квантиль в математической статистике – это значение, которое заданная случайная величина не превышает с фиксированной вероятностью . Таким образом, квантиль порядка – это такое число, вероятность попасть левей которого равна .

Например, квантиль порядка 90% – это такое число, вероятность попасть левей которого составляет 90%. Поэтому квантиль возрастной нормы роста 90% – это, соответственно, такой рост, ниже которого будет 90% человек изучаемой совокупности. На рис.1 приведён график данных с нормальным распределением (среднее арифметическое – 170, стандартное отклонение – 5), последний дециль для такого распределения 90% равен 176.4. Слева от этого величины 90% значений, справа – 10%.

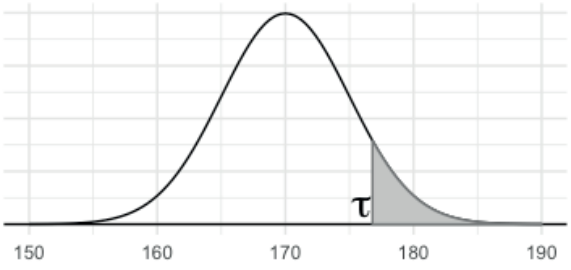

Рис.1. Нормальное распределение и графическое изображение последнего децили (закрашенная зона).

В квантильной регрессии предполагается, что квантиль порядка линейно зависит от независимых переменных.

И хотя зависимость предполагается линейной, она может быть разной для разных квантилей. Для получения оценок в квантильной регрессии минимизируется не сумма квадратов ошибок прогнозов, как в классической линейной регрессии, а взвешенная или асимметричная сумма модулей ошибок прогнозов. В данной статье мы рассмотрим возможности квантильной регрессии для расширенного понимания линейной регрессии. В работе использованы базовые пакеты R 3.6.1 (9), а также пакеты dplyr (10), ggplot2 (11), tidyr (12), purrr (13), broom (14), janitor (15), knitr (16). Работа выполнена в IDE RStudio ver. 1.2.5001. Для создания модели квантильной регрессии использовались функции пакета quantreg (17).

Используемые данные

Для демонстрации возможностей квантильной регрессии мы использовали данные с результатами соревнований по пауэрлифитнгу, проводившихся International Powerlifting Federation (IPF). Пауэрлифтинг – вид спорта, в котором спортсмены соревнуются в трёх упражнениях со штангой: приседании (squat), жиме лёжа (bench) и становой тяге (deadlift). Победителем считается достигший максимального результата (зафиксировавший максимальный вес) в упражнении за ограниченное число попыток.

Из исходного набора данных для анализа были отобраны результаты спортсменов обоего пола в возрасте 20-49 лет (5 возрастных категорий) с массой тела не более 150 кг, занявших места с 1 по 20 на турнирах, проводившихся IPF с 2000 года. Использованный в работе файл с набором данных и скрипт с кодом доступны на сайте https://github.com/valegoshin/Paper_Scripts.

Половозрастной состав участников приведён в таблице 1, созданной с использованием функций пакета janitor (листинг 1). Характер распределения результатов (зафиксированных максимальных весов штанги в упражнениях) приведён на диаграммах плотности (данные стратифицированы по упражнениям и полу) на рис.2

Листинг 1

```
# импорт данных
ipf_lifts <- read_csv(«ipf_lifts.csv»)

# преобразование данных
df <- ipf_lifts %>%
 mutate(year = lubridate::year(date)) %>%
 # выбор запис1 по условию
 filter(
 age_class %in% c(“20-23”, “24-34”, “35-39”, “40-44”, “45-49”),
 year >= 2000,
 bodyweight_kg <= 150,
 equipment == “Single-ply”,
 place %in% as.character(1:20)
 ) %>%
 # выбор переменных и переименование некоторых из них
 select(
 squat = best3squat_kg,
 bench = best3bench_kg,
 deadlift = best3deadlift_kg,
 age, age_class, bodyweight_kg, sex
 ) %>%
 # изменение формы таблицы
 pivot_longer(1:3, names_to = “exercise”, values_to = “result”)
      df %>%
 tabyl(age_class, sex) %>%
 adorn_totals(c(“row”, “col”)) %>%
 adorn_percentages(“col”) %>%
 adorn_pct_formatting(digits = 2) %>%
 adorn_ns() %>%
 adorn_title(“combined”) %>%
 knitr::kable(caption = “Таб.1 Половозрастной состав”)

      df %>%
ggplot(aes(result, fill = sex)) +
geom_density(alpha = .5) +
scale_fill_grey() +
facet_wrap(~ exercise, nrow = 1)
```

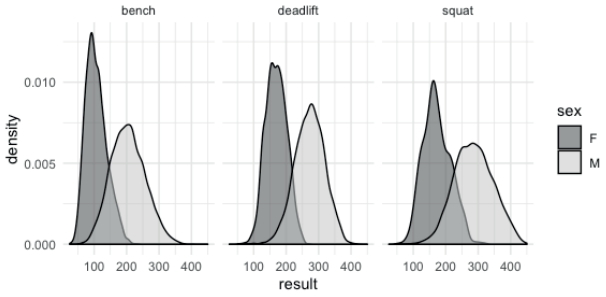

Рис.2. Характер распределения результатов соревнований в зависимости от упражнения и пола спортсмена

Линейные модели

В листинге 2 показано, как можно создать линейные модели, характеризующие зависимость результатов от веса и возраста спортсмена: 1) для всех данных, 2) стратифицированных по отдельным упражнениям и полу, и отобразить это графически.

При подготовке таблицы 3 данные были сгруппированы, объединены в группах, были созданы модели линейной регрессии в группах, созданы объекты, содержащие коэффициенты регрессии и доверительные интервалы, а затем в новую таблицу были выделены группирующие переменные и объект с коэффициентами, объект с коэффициентами был “распакован”, во вновь созданной таблице были отфильтрованы и отсортированы данные. Подобный метод использовался и в дальнейшем.

Графики на рисунках 3 и 4 созданы в пакете ggplot2, точечная диаграмма создана с использованием функции geom_jitter, график линейной регрессии – geom_smooth(method = «lm»), стратификация на подгруппы выполнена применением facet_grid.

Листинг 2

```
# линейная модель базовыми функциями R
model <- lm(result ~ bodyweight_kg + age, data = df)
summary(model)
      Call:
lm(formula = result ~ bodyweight_kg + age, data = df)
Residuals:
 Min 1Q Median 3Q Max
-246.383 -42.697 -0.551 41.484 190.050
```

Таб.1 Половозрастной состав изучаемой совокупности

age_class/sex	F	M	Total
20-23	27.83% (3939)	26.46% (6297)	26.97% (10236)
24-34	32.28% (4569)	31.78% (7563)	31.97% (12132)
35-39	10.51% (1488)	10.83% (2577)	10.71% (4065)
40-44	16.17% (2289)	18.42% (4383)	17.58% (6672)
45-49	13.20% (1869)	12.51% (2976)	12.77% (4845)
Total	100.00% (14154)	100.00% (23796)	100.00% (37950)

```
Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.86084 1.62750 40.47 <2e-16 ***
bodyweight_kg 2.11311 0.01451 145.62 <2e-16 ***
age -0.94282 0.03649 -25.84 <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
Residual standard error: 58.27 on 28240 degrees of
freedom
 (9707 observations deleted due to missingness)
Multiple R-squared: 0.4318, Adjusted R-squared: 0.4317
F-statistic: 1.073e+04 on 2 and 28240 DF, p-value: <
2.2e-16
```

```
    # коэффициенты линейной регрессии - с использованием функций пакетов dplyr и broom
df %>%
 lm(result ~ bodyweight_kg + age, data = .) %>%
 tidy(conf.int = TRUE) %>%
 knitr::kable(digits = 3, caption = “Таб.2 Коэффициенты регрессии - результат от веса спортсмена”)

    # коэффициенты линейной регрессии – стратификация по полу и упражнениям
df %>%
 group_by(exercise, sex) %>% # группировка данных
 nest() %>% # объединение данных в группах
 # создание объектов с результатами линейной регрессии
 mutate(lm_obj = map(data, ~ lm(result ~ bodyweight_kg + age, data = .))) %>%
 # создание объектов с коэффициентами и доверительными интервалами линейной регрессии
 mutate(tidy_lm = map(lm_obj, tidy, conf.int = TRUE)) %>%
 ungroup() %>% # удаление группировки
 transmute(exercise, sex, tidy_lm) %>% # создание новой таблицы
 unnest(cols = c(tidy_lm)) %>% # “распаковка” объектов с коэффициентами
 filter(term != “(Intercept)”) %>% # отбор строк
 select(term, sex, exercise, estimate, conf.low, conf.high) %>% # отбор столбцов
 arrange(desc(term), sex) %>% # сортировка
 # упорядоченный вывод данных в таблице
 knitr::kable(digits = 3, caption = “Таб.3 Коэффициенты регрессии - результат от веса и возраста спортсмена, стратифицированные по упражнениям и полу”)
    df %>%
 ggplot(aes(bodyweight_kg, result)) +
 # создание точечной диаграммы
 geom_jitter(alpha = .7, color = “grey50”, shape = “.”, size = 6) +
 # создание графика линейной регрессии
 geom_smooth(method = lm, color = “black”) +
 # стратификация данных
 facet_grid(sex ~ exercise, scales = “free_y”)
```

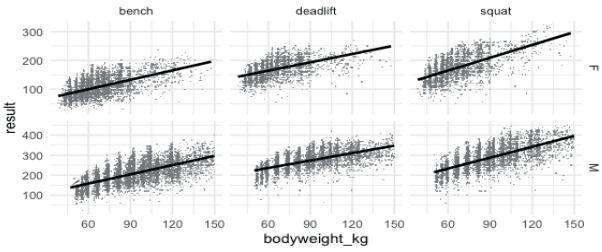

Рис.3. Графическое представление связи между массой тела спортсмена и результатом со стратификацией по упражнениям и полу спортсмена

Таб.2 Коэффициенты линейной регрессии – результат от веса и возраста спортсмена

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	65.861	1.628	40.467	0	62.671	69.051
bodyweight_kg	2.113	0.015	145.622	0	2.085	2.142
age	-0.943	0.036	-25.841	0	-1.014	-0.871

Таб.3. Коэффициенты регрессии – результат от веса и возраста спортсмена, стратифицированные по упражнениям и полу

term	ex	exercise	estimate	conf.low	conf.high
bodyweight_kg	F	squat	1.484	1.416	1.552
bodyweight_kg	F	bench	1.092	1.051	1.133
bodyweight_kg	F	deadlift	0.967	0.910	1.023
bodyweight_kg	M	squat	1.857	1.805	1.909
bodyweight_kg	M	ench	1.562	1.524	1.599
bodyweight_kg	M	deadlift	1.264	1.220	1.309
age	F	squat	-0.876	-1.000	-0.752
age	F	bench	-0.391	-0.467	-0.316
age	F	deadlift	-0.305	-0.408	-0.202
age	M	squat	-1.162	-1.286	-1.037
age	M	bench	-0.492	-0.582	-0.402
age	M	deadlift	-0.657	-0.764	-0.550

ggplot(**aes**(age, result)) +
geom_jitter(alpha = .7, color = “grey50”, shape = “.”, size = 6) +
geom_smooth(method = lm, color = “black”) +
facet_grid(sex ~ exercise, scales = “free_y”)

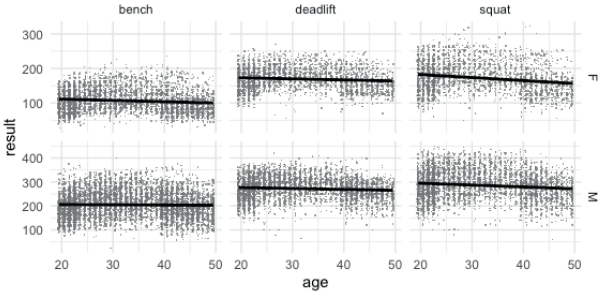

Рис.4. Графическое представление связи между возрастом спортсмена и результатом со стратификацией по упражнениям и полу спортсмена

Квантильные диаграммы

В пакете ggplot2 имеется возможность создавать квантильные диаграммы, используя функцию geom_quantile с параметром quantiles. Этот параметр является числовым вектором, в котором указываются кван-

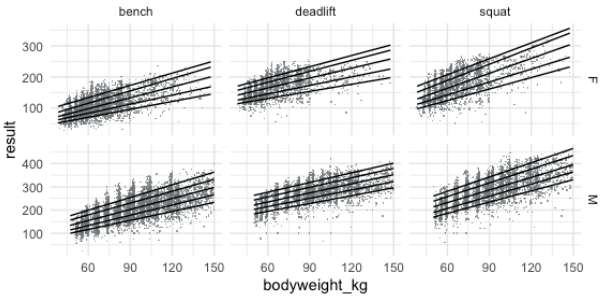

Рис.5. Квантильная диаграмма: вес тела и результат, стратифицированные по упражнениям и полу

df %>%
ggplot(**aes**(age, result)) +
geom_jitter(alpha = .7, color = “grey50”, shape = “.”, size = 6) +
geom_quantile(quantiles = **c**(.1, .25, .5, .75, .9), color = “black”) +
facet_grid(sex ~ exercise, scales = “free_y”)

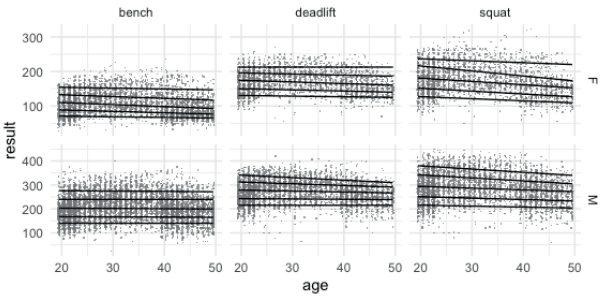

Рис.6. Квантильная диаграмма: возраст и результат, стратифицированные по упражнениям и полу

Квантильная регрессия с помощью пакета quantreg

Функцией пакета quantreg, расширяющей понимание линейной регрессии, является функция rq, выполняемая в формате rq(formula, data = data, tau = ...). Formula создается обычным способом, слева от значка ~ (тильда) зависимая переменная, справа – независимые переменные. Значения tau задаются исследователем в виде числового вектора. Функция plot(summary()) для созданной модели позволяет построить график (листинг 4).

Листинг 4

model <- **rq**(result ~ bodyweight_kg + age,
data = df, tau =
c(.1, .25, .5, .75, .9)
)
summary(model)

```
Call: rq(formula = result ~ bodyweight_kg + age, tau = c(0.1, 0.25,
 0.5, 0.75, 0.9), data = df)
tau: [1] 0.1
Coefficients:
 Value Std. Error t value Pr(>|t|)
(Intercept) 5.04683 2.23447 2.25862 0.02391
bodyweight_kg 1.81107 0.02371 76.39823 0.00000
age -0.65393 0.04669 -14.00611 0.00000
Call: rq(formula = result ~ bodyweight_kg + age, tau = c(0.1, 0.25,
 0.5, 0.75, 0.9), data = df)
tau: [1] 0.25
Coefficients:
 Value Std. Error t value Pr(>|t|)
(Intercept) 25.13483 2.24021 11.21985 0.00000
bodyweight_kg 2.04458 0.02121 96.41635 0.00000
age -0.83255 0.05087 -16.36472 0.00000
Call: rq(formula = result ~ bodyweight_kg + age, tau = c(0.1, 0.25,
 0.5, 0.75, 0.9), data = df)
tau: [1] 0.5
Coefficients:
 Value Std. Error t value Pr(>|t|)
(Intercept) 59.58818 2.21504 26.90168 0.00000
bodyweight_kg 2.21046 0.02130 103.78112 0.00000
age -0.99851 0.04796 -20.81753 0.00000
Call: rq(formula = result ~ bodyweight_kg + age, tau = c(0.1, 0.25,
 0.5, 0.75, 0.9), data = df)
tau: [1] 0.75
Coefficients:
 Value Std. Error t value Pr(>|t|)
(Intercept) 90.42114 2.32754 38.84832 0.00000
bodyweight_kg 2.40112 0.02220 108.15010 0.00000
age -1.13374 0.05190 -21.84438 0.00000
Call: rq(formula = result ~ bodyweight_kg + age, tau = c(0.1, 0.25,
 0.5, 0.75, 0.9), data = df)
tau: [1] 0.9
Coefficients:
 Value Std. Error t value Pr(>|t|)
(Intercept) 126.60575 2.66769 47.45891 0.00000
bodyweight_kg 2.43031 0.02300 105.66905 0.00000
```

age -1.25114 0.05800 -21.57127 0.00000
plot(**summary**(model))

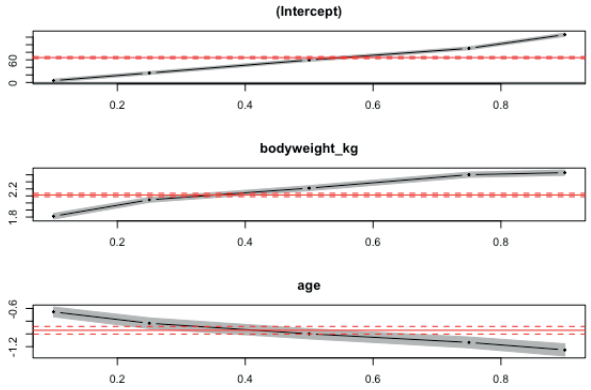

Рис.7. Диаграмма средствами пакета quantreg

Изучение квантильной регрессии с использованием функций пакетов dplyr, tidyr, purrr, broom, knitr.

Пакеты dplyr, tidyr, purrr, broom, knitr позволяют создать графики и таблицы, существенно уменьшающие количество дополнительных действий по подготовке материалов анализа данных к публикации. Объекты frq и flm созданы тем же способом, что и в листинге 2. При создании таблиц 4 и 5 из объекта frq были выделены в новую таблицу необходимые переменные, создана новая переменная, содержащая коэффициенты регрессии с доверительными интервалами, далее таблица преобразована в "широкую". Графики на рисунках 8 и 9 созданы на основе объединения объектов frq и flm.

Листинг 5

```
frq <- df %>%
 select(exercise, sex, result, bodyweight_kg) %>%
 group_by(exercise, sex) %>%
 nest() %>%
 mutate(rq_obj = map(data, ~ rq(result ~ bodyweight_
kg, data = ., tau = c(.1, .25, .5, .75, .9)))) %>%
 mutate(tidy_rq = map(rq_obj, tidy, conf.int = TRUE))
%>%
 ungroup() %>%
 transmute(exercise, sex, tidy_rq) %>%
 unnest(cols = c(tidy_rq)) %>%
 filter(term != "(Intercept)")
frq %>%
 # отбор переменных и создание новой переменной
 transmute(exercise, sex, tau, est_conf = paste0(
 as.character(round(estimate, 2)),
 " (",
 as.character(round(conf.low, 3)),
 " - ",
 as.character(round(conf.high, 3)),
 ")"
 )) %>%
 # создание "широкой" таблицы
 pivot_wider(names_from = tau, values_from = est_
conf, names_prefix = "tau_") %>%
 # упорядоченный вывод данных в таблице
 knitr::kable(caption = "Таб.4 Коэффициенты квантильной регрессии результат от массыт тела, стратифицированные по упражнениям и полу")
      flm <- df %>%
 select(exercise, sex, result, bodyweight_kg) %>%
 group_by(exercise, sex) %>%
 nest() %>%
 mutate(lm_obj = map(data, ~ lm(result ~ bodyweight_
kg, data = .))) %>%
 mutate(tidy_lm = map(lm_obj, tidy, conf.int = TRUE))
%>%
 ungroup() %>%
 transmute(exercise, sex, tidy_lm) %>%
 unnest(cols = c(tidy_lm)) %>%
 filter(term != "(Intercept)") %>%
 select(exercise, sex, term, estimate_lm = estimate, conf.
low_lm = conf.low, conf.high_lm = conf.high)
      left_join(frq, flm, by = c("exercise", "sex")) %>%
 ggplot(aes(tau, estimate)) +
 geom_ribbon(aes(ymin = conf.low, ymax = conf.high),
fill = "lightgrey", alpha = .8) +
 geom_point(alpha = .5) +
 geom_line(alpha = .5) +
 geom_hline(aes(yintercept = estimate_lm), linetype =
```

Таб.4. Коэффициенты квантильной регрессии результат от массы тела, стратифицированные по упражнениям и полу

exercise	ex	tau_0.1	tau_0.25	tau_0.5	tau_0.75	tau_0.9
squat	M	1.61 (1.514 - 1.723)	1.75 (1.695 - 1.836)	1.81 (1.749 - 1.909)	1.97 (1.905 - 2.093)	2.05 (1.925 - 2.215)
bench	M	1.29 (1.235 - 1.361)	1.47 (1.403 - 1.535)	1.58 (1.487 - 1.653)	1.7 (1.629 - 1.793)	1.82 (1.764 - 1.923)
deadlift	M	1.12 (1.072 - 1.181)	1.19 (1.137 - 1.263)	1.31 (1.244 - 1.368)	1.37 (1.293 - 1.459)	1.39 (1.289 - 1.538)
squat	F	1.24 (1.09 - 1.375)	1.39 (1.318 - 1.462)	1.6 (1.476 - 1.684)	1.74 (1.613 - 1.92)	1.71 (1.561 - 1.946)
bench	F	0.85 (0.777 - 0.888)	0.97 (0.927 - 1.049)	1.17 (1.075 - 1.259)	1.33 (1.261 - 1.418)	1.31 (1.164 - 1.421)
deadlift	F	0.76 (0.668 - 0.855)	0.92 (0.831 - 1.024)	1.06 (1.005 - 1.151)	1.18 (1.024 - 1.297)	1.21 (1.053 - 1.338)

```
2) +
 geom_hline(aes(yintercept = conf.low_lm), linetype =
3) +
 geom_hline(aes(yintercept = conf.high_lm), linetype =
3) +
 facet_grid(sex ~ exercise, scales = "free_y")
```

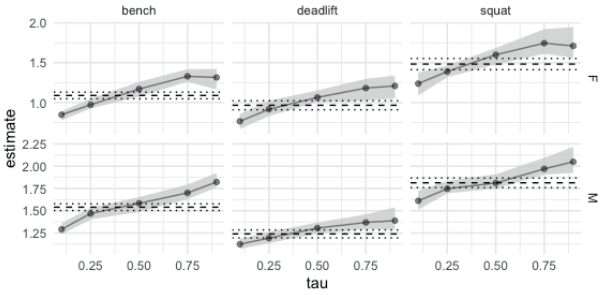

Рис.8. Квантили масса тела/результат и коэффициент линейной регрессии, стратифицированные по упражнениям и полу

```
frq <- df %>%
 select(exercise, sex, result, age) %>%
 group_by(exercise, sex) %>%
 nest() %>%
 mutate(rq_obj = map(data, ~ rq(result ~ age, data = .,
tau = c(.1, .25, .5, .75, .9)))) %>%
 mutate(tidy_rq = map(rq_obj, tidy, conf.int = TRUE))
%>%
 ungroup() %>%
 transmute(exercise, sex, tidy_rq) %>%
 unnest(cols = c(tidy_rq)) %>%
 filter(term != "(Intercept)")
frq %>%
    transmute(exercise, sex, tau, est_conf = paste0(
as.character(round(estimate, 2)), " (",
as.character(round(conf.low, 3)), " - ",
as.character(round(conf.high, 3)), ")" )) %>%
pivot_wider(names_from = tau, values_from = est_conf,
names_prefix = "tau_") %>%
    knitr::kable(caption = «Таб.5 Коэффициенты
квантильной регрессии результат от возраста, страти-
фицированные по упражнениям и полу»)
    flm <- df %>%
 select(exercise, sex, result, age) %>%
 group_by(exercise, sex) %>%
 nest() %>%
 mutate(lm_obj = map(data, ~ lm(result ~ age, data =
.))) %>%
 mutate(tidy_lm = map(lm_obj, tidy, conf.int = TRUE))
%>%
 ungroup() %>%
 transmute(exercise, sex, tidy_lm) %>%
 unnest(cols = c(tidy_lm)) %>%
 filter(term != "(Intercept)") %>%
 select(exercise, sex, term, estimate_lm = estimate, conf.
low_lm = conf.low, conf.high_lm = conf.high)
    left_join(frq, flm, by = c("exercise", "sex")) %>%
 ggplot(aes(tau, estimate)) +
 geom_ribbon(aes(ymin = conf.low, ymax = conf.high),
fill = "lightgrey", alpha = .8) +
 geom_point(alpha = .5) +
 geom_line(alpha = .5) +
 geom_hline(aes(yintercept = estimate_lm), linetype =
2) +
 geom_hline(aes(yintercept = conf.low_lm), linetype =
3) +
 geom_hline(aes(yintercept = conf.high_lm), linetype =
3) +
 facet_grid(sex ~ exercise, scales = "free_y")
```

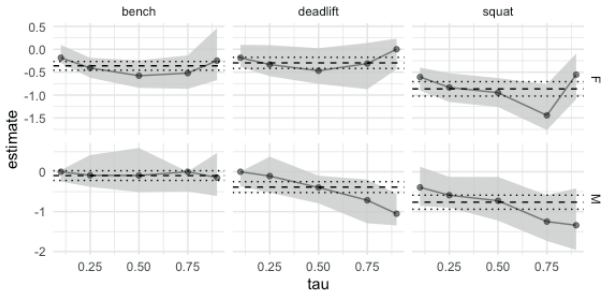

Рис.9. Квантили возраст/результат и коэффициент линейной регрессии, стратифицированные по упражнениям и полу

Анализ связи между массой тела спортсмена и результатом с учетом данных квантильной регрессии позволяет сказать, что увеличение массы спортсмена

Таб.5. Коэффициенты квантильной регрессии результат от возраста, стратифицированные по упражнениям и полу

exercise	ex	tau_0.1	tau_0.25	tau_0.5	tau_0.75	tau_0.9
squat	M	-0.39 (-0.858 - 0.125)	-0.59 (-0.9 - -0.13)	-0.73 (-1.224 - -0.131)	-1.25 (-1.734 - -0.588)	-1.34 (-1.959 - -0.416)
bench	M	0 (-0.246 - 0.041)	-0.09 (-0.378 - 0.417)	-0.09 (-0.511 - 0.595)	0 (-0.495 - 0)	-0.15 (-0.608 - 0.471)
deadlift	M	0 (-0.362 - 0)	-0.11 (-0.523 - 0.376)	-0.39 (-0.792 - -0.101)	-0.71 (-1.292 - -0.197)	-1.05 (-1.344 - -0.553)
squat	F	-0.61 (-0.899 - -0.401)	-0.83 (-1.151 - -0.523)	-0.95 (-1.259 - -0.63)	-1.44 (-1.762 - -0.765)	-0.56 (-1.089 - -0.098)
bench	F	-0.19 (-0.298 - 0.094)	-0.41 (-0.622 - -0.188)	-0.58 (-0.841 - -0.244)	-0.52 (-0.864 - -0.135)	-0.25 (-0.673 - 0.461)
deadlift	F	-0.19 (-0.385 - 0.095)	-0.33 (-0.588 - 0.089)	-0.47 (-0.755 - 0.022)	-0.31 (-0.87 - 0.137)	0 (-0.437 - 0.23)

повышает его результат в упражнениях, особенно значимо это проявляется у спортсменов, показывающих более высокие результаты. Так, коэффициент линейной регрессии для мужчин в приседании равен 1.86 (1.805-1.909) (таб. 3), коэффициенты квантильной регрессии для мужчин в приседании изменяются от 1.61 (1.514-1.723) для 10% квантиля до 2.05 (1.925-2.215) для 90% квантиля (таб. 4). Спортсмены, показывающие высокие результаты, более эффективно используют массу тела.

Изучение связи между возрастом спортсмена и результатом с учётом данных квантильной регрессии показывает, что возраст в изучаемом диапазоне не влияет на показатели спортсменов, показывающих высокие результаты. Коэффициент линейной регрессии у мужчин в жиме лёжа равен -0.49 (-0.582 – -0.402) (таб. 3), коэффициенты квантильной регрессии в жиме лёжа у мужчин для 10% квартиля 0 (-0.246 – 0.041), для 90% квартиля -0.15 (-0.608 – 0.471) (таб.5). Изучение доверительных интервалов коэффициентов линейной и квантильной регрессии на рисунке 9 позволяет говорить об отсутствии значимых изменений в коэффициентах квантильной регрессии.

Метод квантильной регрессии в дополнение к линейной регрессии позволил установить следующее:

- коэффициенты регрессии для массы тела спортсмена изменяются в зависимости от квантиля зависимой переменной (зафиксированный результат в упражнении) (рис.8),
- коэффициенты регрессии для возраста спортсмена не отличаются от коэффициента линейной регрессии при изменении квантиля зависимой переменной (рис.9).

Прогнозирование в квантильной регрессии

Выполняется с использованием функции predict, прогноз выводится для каждого квантиля, заданного исследователем в модели. Качество прогнозирования можно оценить по показателю средней абсолютной процентной ошибки , где real – реальное значение, fitted – прогнозируемое значение, n – число наблюдений. Для оценки прогнозирования создадим набор данных без пропущенных значений и разделим его на тренировочную и тестовую части. На тренировочных данных создадим модель квантильной регрессии и оценим её на тестовых данных (листинг 6).

Листинг 6

```
# создание набора данных для прогнозирования
df1 <- df %>%
filter(exercise == «squat») %>%
select(result, age, bodyweight_kg) %>%
drop_na()
# разделение на тренировочный и тестовый наборы
set.seed(123999)
ind <- caret::createDataPartition(df1$result, p = .7, list = FALSE)
train_data <- df1[ind, ]
test_data <- df1[-ind, ]
    # модель квантильной регрессии
rq_train <- rq(result ~ bodyweight_kg + age, train_data, tau = c(.1, .25, .5, .75, .9))
    # прогнозные значения для тестового набора
predictions <- predict(rq_train, newdata = test_data, interval = “prediction”)
    # получение оценок MAPE для тренировочного набора
train_mape <- rq_train %>%
augment() %>%
select(result, .tau, .fitted) %>%
    mutate(mape = abs(result - .fitted) / result) %>%
group_by(.tau) %>%
summarise(mape_train = mean(mape))
# получение оценок MAPE для тестового набора
test_mape <- data.frame(
 real_d = test_data$result,
 fitted_d = predictions
) %>%
    set_names(“result”, “tau10”, “tau25”, “tau50”, “tau75”, “tau90”) %>%
pivot_longer(-result, names_to = “tau”, values_to = “preds”) %>%
mutate(mape = abs(result - preds) / result) %>%
group_by(tau) %>%
summarise(mape_test = mean(mape))
# объединение результатов и вывод их в таблице
    bind_cols(train_mape, test_mape) %>%
    select(tau, mape_train, mape_test) %>%
knitr::kable(digits = 3, caption = “Таб.6 MAPE для тренировочных и предсказанных значений”)
```

Таб.6. MAPE для тренировочных и предсказанных значений

tau	mape_train	mape_test
tau10	0.263	0.264
tau25	0.193	0.196
tau50	0.181	0.184
tau75	0.247	0.248
tau90	0.346	0.347

Данные в таблице 6 показывают, что значения средней абсолютной процентной ошибки сходные для каждого квантиля.

Работа с R

Программная среда R является свободно распространяемой кросс-платформенной программной средой, используемой для статистического анализа 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/. Азы работы в программной среде R, а также примеры биомедицинских задач и методов их решения с помощью R разбирались в наших более ранних публикациях (18 и др.), которые можно найти на сайте www.elibrary.ru/.

Список литературы / References:

1. Koenker R, Hallock KF. Quantile regression. Journal of Economic Perspectives 2001;15(4):143–156.
2. Koenker R. Censored quantile regression redux. Journal of Statistical Software 2008;27(6):1–25.
3. Koenker R. Quantile Regression. New York: Cambridge university press, 2005.
4. Choi S, Kang S, Huang X. Smoothed quantile regression analysis of competing risks. Biom J. 2018;60(5):934–946.
5. Luo X, Huang CY, Wang L. Quantile regression for recurrent gap time data. Biometrics 2013;69(2):375–385.
6. Анохина ЮЕ, Мартынов БВ, Гайдар БВ. Прогностическая значимость длительности безрецидивного периода у пациентов со злокачественными глиомами головного мозга. Вестник Российской Военно-медицинской академии. 2013;2(42):44–48 *Anokhina YE, Martynov BV, Gaidar BV. Prognostic significance of the duration of exacerbation-free period among patients with malignant gliomas. Vestnik Rossiyskoy voenno-medicinskoy akademii 2013;2(42):44–48.[In Russian]*
7. Li R, Peng L. Assessing Quantile Prediction with Censored Quantile Regression Models. Biometrics. 2017;73(2)(6):517–528.
8. Hong HG, Christiani DC, Li Y. Quantile regression for survival data in modern cancer research: expanding statistical tools for precision medicine. Precision Clinical Medicine 2019;2(2):90–99.
9. R Core Team. A Language and Environment for Statistical Computing. https://www.R-project.org/ [Accessed 2 October 2019]
10. Wickham H, Romain F, Lionel H, Müller K. Dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr. [Accessed 2 October 2019]
11. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org. [Accessed 2 October 2019]
12. Wickham H, Henry L. Tidyr: Tidy Messy Data. https://CRAN.R-project.org/package=tidyr. [Accessed 2 October 2019]
13. Henry L, Wickham H. Purrr: Functional Programming Tools. https://CRAN.Rproject.org/package=purrr. [Accessed 2 October 2019]
14. Robinson D, Hayes A. Broom: Convert Statistical Analysis Objects into Tidy Tibbles. https://CRAN.R-project.org/package=broom. [Accessed 2 October 2019]
15. Firke S. Janitor: Simple Tools for Examining and Cleaning Dirty Data. https://CRAN.R-project.org/package=janitor. [Accessed 2 October 2019]
16. Xie Y. 2015. Dynamic Documents with R and Knitr. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.name/knitr/. [Accessed 2 October 2019]
17. Koenker R. Quantreg: Quantile Regression. https://CRAN.R-project.org/package=quantreg. [Accessed 2 October 2019]
18. Егошин ВЛ, Иванов СВ, Саввина НВ, Капанова ГЖ, Гржибовский АМ. Основы работы в программной среде R при анализе биомедицинских данных. Экология человека. 2018;7:55–64. *Egoshin VL, Ivanov SV, Savvina NV, Kapanova GZh. Grjibovski AM. Basic Principles of Biomedical Data Analysis in R. Ekologiya cheloveka [Human Ecology]. 2018;7:55–64.[In Russian]*