УДК 614.2:004:303.4
МРНТИ 76.75.75, 12.41.55

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ 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, Архангельск, Россия
[2]Северо-Восточный федеральный университет, Якутск, Россия
[3]Казахский Национальный университет им. Аль-Фараби, Алматы, Казахстан

Citation/
библиографиялық сілтеме/
библиографическая ссылка:

Egoshin VL, Savvina NV, Grjibovski AM. Multivariable linear regression in R brief guidelines for master and doctoral students of medicine and public health specialties. West Kazakhstan Medical journal 2019;61(1):4–15.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. R бағдарламалық ортада көптік желілік регрессия – «Медицина» және «Қоғамдық денсаулық сақтау» мамандығы бойынша магистранттар мен докторанттарға арналған қысқаша ұсынымдар. West Kazakhstan Medical journal 2019;61(1):4–15.

Егошин ВЛ, Саввина НВ, Гржибовский АМ. Множественная линейная регрессия в программной среде R – краткие рекомендации для магистрантов и докторантов по специальности «Медицина» и «Общественное здравоохранение». West Kazakhstan Medical journal 2019;61(1):4–15.

Multivariable linear regression in R: brief guidelines for Master and Doctoral students of Medicine and Public health specialties

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 linear 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 medical birth registry. In addition to the syntax we present R outputs and their interpretation.

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

R бағдарламалық ортада көптік желілік регрессия – «Медицина» және «Қоғамдық денсаулық сақтау» мамандығы 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 федералды университеті, Якутск, Ресей
[3]Әл Фараби атындағы Қазақ Ұлттық уинверситеті, Алматы, Қазақстан

Бұл мақалада көптік желілік регрессиялы талдау үшін R бағдарламалық ортаны қолданудың негізгі принциптері ұсынылған. Популяциялық регистр түрлерінің бірнеше модифицирленген нақты мәліметтерін қолдана отырып, практикалық мысал ретінде регрессиялық моделдерді жасау және бағалау үшін R-де қадамдық алгоритм және синтаксис ұсынылған. Синтаксистен бөлек, R-де көрсеткендей, нәтижелер, сонымен қатар олардың түсіндірілуі ұсынылған.

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

Множественная линейная 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, 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 множественного линейного регрессионного анализа. Представлен пошаговый алгоритм и синтаксис в R для создания и оценки регрессионных моделей в виде практического примера с использованием несколько модифицированных реальных данных популяционного регистра родов. Помимо синтаксиса представлены результаты в том виде, как их выдает

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

Received/
Келіп түсті/
Поступила:
28.02.2019

Accepted/
Басылымға қабылданды/
Принята к публикации:
06.03.2019

ISSN 1814-5620 (Print)

R, а также их интерпретация.
Ключевые слова: *R, линейная регрессия, синтаксис, листинг, моделирование, валидация.*

Множественная линейная регрессия – метод статистического анализа данных, широко используемый в статистике и машинном обучении (1-3). В исследованиях в области медицины и общественного здравоохранения этот метод также применяется достаточно часто. Однако, на постсоветском пространстве, за исключением стран Балтии, данный метод применяется значительно реже, чем в дальнем зарубежье. Краткие рекомендации по применению линейного регрессионного анализа с помощью пакетов статистических программ SPSS, Stata и Statistica были ранее опубликованы в русскоязычной литературе (4-6). В последнее время, как в англоязычном научном пространстве, так и в Казахстане набирает популярность программная среда R, которая помимо профессиональных достоинств является абсолютно бесплатной, что делает ее еще более привлекательной для исследователей. 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 (7-11) мы уже касались вопросов применения программной среды R для различных видов бивариантного анализа в биомедицинских исследованиях, в том числе и для простой линейной регрессии (12). В данной работе мы усложняем задачу и представляем использование программной среды R для многомерного линейного регрессионного анализа.

Множественная линейная регрессия – метод оценки влияния независимых переменных (т.н. предикторов или “предсказывающих” переменных) на одну зависимую непрерывную переменную. Предикторами могут быть и количественные, и качественные (категориальные) переменные. Для оценки связи создается математическая модель, включающая переменные, отобранные на основании тех или иных аргументов исследователя. Построению модели должна предшествовать первичная оценка переменных, изучение их распределения и бивариантных связей.

Модель множественной линейной регрессии, связывающая зависимую переменную с независимыми переменными $x^{1,\dots,b-1}$, может быть записана как:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

где $\epsilon \sim N(0, \sigma^2)$

В данной формуле $\beta_0, \beta_1, \beta_2, \beta_{p-1}$ – коэффициенты регрессии, ϵ – ошибка, p – общее количество коэффициентов регрессии. Предполагается, что ошибка измерения ϵ имеет нормальное распределение со средним значением равным 0. Это условие выполнения множественного линейного анализа необходимо проверять, используя хорошо известные читателям графические и статистические способы (9-10).

Выражение “линейная” в понятии “множественная линейная регрессия” связано с тем, что модель линейна в параметрах $\beta_0, \beta_1, \beta_2, \beta_{p-1}$. Это означает, что каждый параметр умножается на переменную, тогда как функция регрессии является суммой параметров умноженных на -переменные. При этом даже β_0 рассматривается как результат умножения на $x = 1$. Условие линейности также необходимо проверять. Лучше посредством построения скаттерограмм (4, 9).

Оценка модели множественной линейной регрессии включает оценку параметров модели, статистической значимости предикторов, допустимость условий выполнения метода. В основе оценки параметров модели (коэффициентов регрессии) – уменьшение суммы квадратов ошибки модели. Mean square error $(MSE) = \frac{SSE}{n-p}$ соответствует σ^2, дисперсии для ошибки модели, SSE – сумма квадратов ошибок, n – размер выборки, p – количество коэффициентов регрессии. Root mean square error $(RMSE) = \sqrt{MSE}$ оценивает σ, стандартное отклонение и известна как regression standard error или the residual standard error.

Каждый β коэффициент показывает изменение значения зависимой переменной при изменении значения, связанной с ним независимой переменной на единицу при прочих равных условиях.

При проведении регрессионного анализа оценивается нулевая гипотеза, предполагающая, что коэффициенты регрессии не отличаются от нуля. F-test оценивает нулевую гипотезу о равенстве нулю всех коэффициентов регрессии, альтернативная гипотеза заключается в том, что хотя бы один коэффициент не равен нулю. Дополнительно для каждого коэффициента вычисляется доверительный интервал с заданным уровнем доверительной вероятности (обычно 95%).

Коэффициент детерминации R^2 показывает, какая доля изменчивости зависимой переменной обусловлена моделью. Скорректированный (adjusted) R^2 показывает измененное значение доли

изменчивости зависимой переменной, с учетом количества независимых переменных. Чем выше эти показатели, тем лучше модель. Коэффициенты AIC (Akaike's information criterion) и BIC (Bayesian information criterion) служат для оценки пригодности статистической модели и используются при выборе моделей. MAPE (Mean absolute percentage error) – средняя абсолютная процентная ошибка может быть рассчитана по формуле: $MAPE = mean\left(\frac{abs(\text{actual}-\text{predict})}{\text{actual}}\right)$ где actual – реальные значения переменной, predict – предсказанные на основании созданной модели значения. Показатель MAPE используется для оценки прогностических возможностей регрессионной модели. Чем ниже эти показатели, тем лучше модель.

При применении множественного линейного регрессионного анализа необходимо убедиться в том, что соблюдаются все условия для его применения. В противном случае, результаты анализа можно считать как минимум сомнительными. Эти условия следующие:

1. Линейная связь между зависимыми и независимыми переменными;
2. Остатки независимы, подчиняются закону нормального распределения со средним арифметическим равным нулю; отсутствуют гетероскедастичность и автокорреляция остатков;
3. Отсутствуют значения, оказывающие сильное влияние на модель;
4. Отсутствует мультиколлинеарность, то есть отсутствие тесной корреляции (обычно выше 0,9) между независимыми переменными.

В программной среде R регрессионный анализ основывается на построении модели, включающей формулу:

lm(formula = зависимая переменная ~ независимые переменные, data)

В формулах R используются следующие символы (13):

Символ	Значение
~	отделяет зависимую переменную (слева от знака) от независимых
+	разделяет независимые переменные
:	обозначает взаимодействие между переменными
*	обозначает все возможные сочетания переменных

Далее происходит оценка модели с применением функций различных пакетов R. В предлагаемой работе будет продемонстрировано, как в программной среде R можно создать модель множественной линейной регрессии, оценить эту модель, в том числе с использованием визуальных методов, и улучшить её. В работе используются базовые пакеты R 3.5.2, а также пакеты dplyr, ggplot2 и другие, входящие в пакет tidyverse, а также пакеты broom, car, DescTools, GGally, gridExtra. Работа выполнена в IDE RStudio ver. 1.1.463.

Данные для практического примера анализа данных с использованием среды R представляют собой несколько измененную случайную выборку материалов Архангельского областного регистра родов (14). Подготовка данных к анализу включает отбор записей по значению, создание новых переменных, выбор переменных для включения в модель, удаление записей с пустыми значениями, удаление записей с малым количеством значений переменной, преобразование номинальных переменных, удаление выбросов.

Импорт данных и подготовка их к анализу в листинге 1.

Листинг 1

```
# установленные пакеты
library(tidyverse)
library(broom)
library(car)
library(GGally)
library(DescTools)
# импорт из файла
df <- foreign::read.spss("../data/Simulated_sample.sav",
to.data.frame = TRUE)
df1 <- df %>%
 # создание новых переменных
 mutate(BMI = Maternal_weight/(Maternal_height *
.01) ** 2,
 Smoking = factor(ifelse(Smoking_before_pregnancy
== "yes" | Smoking_during_pregnancy == "yes", "yes",
"no")),
 BMI_group = cut(BMI, breaks = c(0, 18.5, 25, 30, Inf),
 labels = c("<18.5", "18.5-25", "25-30", "30+")),
 age_group = cut(Maternal_age, breaks = c(10, 18, 25,
30, 35, Inf),
 right = FALSE, labels = c("<18", "18-25", "25-30", "30-
35", "35+"))) %>%
 # выбор переменных для включения в модель
 select(BIrthweight, Birthlength, Gestational_age,
Marital_status,
 Education, BMI_group, Maternal_height, Maternal_age,
age_group,
 Vitamins_before_pregnancy, Vitamis_during_pregnancy,
 Folic_acid_before_pregnancy, Folic_acid_during_
pregnancy,
 Infant_sex, Smoking) %>%
 na.omit() %>%
# удаление строк с пустыми значениями
 # удаление записей с малым количеством значений
переменной
 filter(Marital_status != "Other",
 !Education %in% c("None", "Unknown")) %>%
```

```
# преобразование номинальных переменных
 mutate(Marital_status = fct_relevel(Marital_status,
“Married”),
 Infant_sex = factor(Infant_sex, levels = c(“Male”,
“Female”)),
 age_group = fct_relevel(age_group, “25-30”),
 BMI_group = fct_relevel(BMI_group, “18.5-25”),
 Education = fct_relevel(Education, “Technical School”))
# удаление выбросов
# функция преобразования выбросов в NA
Outl_NA <- function(x) {x[which(x %in% boxplot.
stats(x)$out)] <- NA; x}
# таблица данных после удаления выбросов
df1 <- df1 %>%
 mutate_if(is.numeric, Outl_NA) %>%
 na.omit()
```

Для включения отобрано записей n=1641.

Предполагается получить ответы на вопросы, как влияет на массу тела новорожденного срок беременности, возраст, ИМТ, рост матери, курение, прием витаминов, фолиевой кислоты, пол новорожденного; а также сделать предположение о величине этого влияния.

Предварительная оценка данных (листинг 2) предполагает оценку каждой переменной в соответствии с её типом. В R существует много способов выполнить эту оценку. В данном случае использовалась функция skim_to_wide из пакета skimr. Для оценки распределения количественных данных и их взаимосвязи, в том числе корреляционной, будет использована функция ggpairs из пакета GGally.

Листинг 2

```
df1 %>%
 select_if(is.numeric) %>%
 skimr::skim_to_wide() %>%
 select(-c(1, 3 : 5)) %>%
 knitr::kable(caption = “Таб.1 Непрерывные
переменные”)
#
df1 %>%
 select_if(is.factor) %>%
 skimr::skim_to_wide() %>%
 select(-c(1, 3 : 5)) %>%
 knitr::kable(caption = “Таб.2 Категориальные
переменные”)
###
df1 %>%
 select_if(is.numeric) %>%
 ggpairs()
```

Модель 1

Используемая модель приведена в листинге 3. В качестве зависимой переменной выбрана масса тела при рождении, в качестве независимых: длина тела новорожденного, срок беременности, семейный статус, образование, группа по ИМТ, возрастная группа, пол новорожденного, курение до и/или во время беременности, прием витаминов до и во время

Таб.1 Непрерывные переменные. Описательная статистика.

variable	mean	sd	p0	p25	p50	p75	p100	hist
Birthlength	52.72	2.12	47	51	53	54	58	
BIrthweight	3448.47	447.95	2105	3140	3450	3750	4740	
Gestational_age	39.18	1.24	35	38	39	40	42	
Maternal_age	28.44	5.22	15	25	28	32	42	
Maternal_height	163.85	6.1	148	160	164	168	180	

Таб.2 Категориальные переменные. Общая информация.

variable	n_unique	top_counts	ordered
age_group	5	25-: 544, 30-: 470, 18-: 391, 35+: 227	FALSE
BMI_group	4	18.: 1069, 25-: 330, 30+: 147, <18: 95	FALSE
Education	4	Tec: 720, Hig: 571, Sec: 247, Pri: 103	FALSE
Folic_acid_before_pregnancy	2	no: 1609, yes: 32, NA: 0	FALSE
Folic_acid_during_pregnancy	2	yes: 888, no: 753, NA: 0	FALSE
Infant_sex	2	Mal: 846, Fem: 795, NA: 0	FALSE
Marital_status	3	Mar: 1155, Coh: 289, Unm: 197, Oth: 0	FALSE
Smoking	2	no: 1344, yes: 297, NA: 0	FALSE
Vitamins_before_pregnancy	2	no: 1606, yes: 35, NA: 0	FALSE
Vitamis_during_pregnancy	2	yes: 927, no: 714, NA: 0	FALSE

беременности, прием фолиевой кислоты до и во время беременности. Последние переменные включены в их мультипликативном взаимодействии.

Листинг 3

```
fit1 <- lm(BIrthweight ~ Birthlength + Gestational_age
+ Marital_status + Education + BMI_group + age_group
+ Infant_sex + Smoking + Vitamins_before_pregnancy
* Vitamis_during_pregnancy + Folic_acid_before_
pregnancy * Folic_acid_during_pregnancy, df1).
```

Для оценки модели множественной линейной регрессии будут использованы функции пакета broom: glance, tidy, augment.

Пакет broom создавался для превращения статистических объектов, таких как результат выполнения тестов, функций в "опрятные" наборы данных, что упрощает последующую работу с их представлением в табличном или графическом виде. В случае работы с объектами, получаемыми в результате регрессионного анализа, функция glance возвращает r.squared – коэффициент детерминации, R^2, adj.r.squared – скорректированный R^2, sigma - $RMSE$, p.value – p.value F – теста, AIC, BIC и некоторые другие. Функция tidy выводит данные о значениях параметров модели: величину коэффициентов (estimate), стандартную ошибку (std.error), значение Т-критерия (statistic), p.value. При включении в функцию аргумента conf.int=TRUE выводятся такие параметры как нижний (conf.low) и верхний (conf.low) доверительный интервал. Результаты выполнения этой функции информативно представлять в графическом виде. Функция augment принимает модель объекта как набор данных и добавляет информацию о каждом наблюдении, обычно это предсказанные значения в столбце .fitted, значения остатков в столбце .resid, стандартные ошибки в столбце .se.fit, Cook's distance в столбце .cooksd.

Данные о модели будут выведены в табличном и графическом виде при использовании кода из листинга 4. Также будет определен фактор инфляции дисперсии для модели (VIF), показатели фактора более 4 (по другим источникам более 8 или даже 9) указывают на мультиколлинеарность.

Листинг 4

```
# оценка мультиколлинеарности переменных в
модели
vif(fit1) %>% as.data.frame() %>% select(1) %>%
 knitr::kable(digits = 3, caption = «Таб.3 vif значения
для модели fit1»)
```

Таб.3 vif значения для модели fit1

	GVIF
Birthlength	1.243
Gestational_age	1.188
Marital_status	1.122
Education	1.356
BMI_group	1.110
age_group	1.257
Infant_sex	1.055
Smoking	1.178
Vitamins_before_pregnancy	34.487
Vitamis_during_pregnancy	1.233
Folic_acid_before_pregnancy	10.775
Folic_acid_during_pregnancy	1.248
Vitamins_before_pregnancy:Vitamis_during_pregnancy	35.582
Folic_acid_before_pregnancy:Folic_acid_during_pregnancy	10.919

```
# использование функции glance()
fit1 %>%
 glance() %>%
```

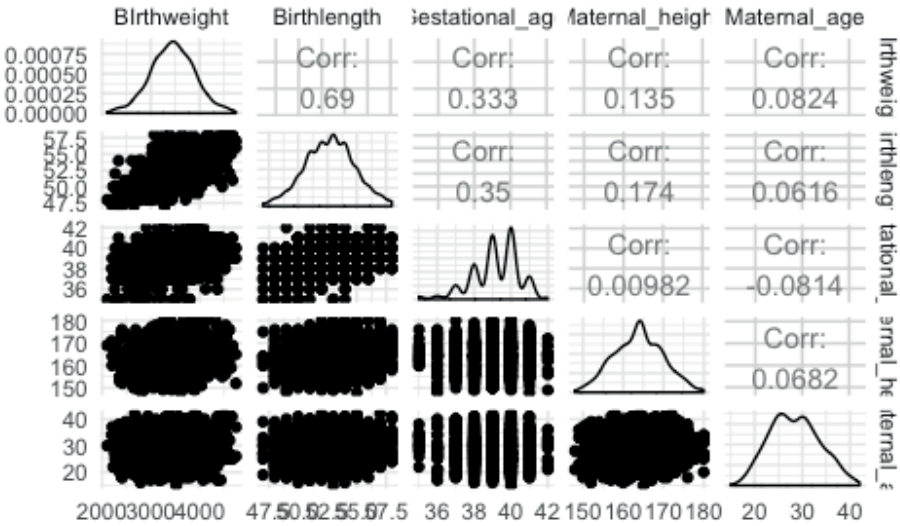

Рис.1 Графическая оценка распределения, связи между переменными, а также коэффициенты корреляции.

Таб.4 Модель fit1: оцениваемые показатели

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	df.residual
0.5	0.493	318.86	73.58	0	23	-11777	23602	23731	1618

```
 select(-10) %>%
 knitr::kable(caption = "Таб.4 Модель fit1:
оцениваемые показатели",
 digits = c(3, 3, 2, 2, 6, 0, 0, 0, 0, 0))
# использование функции tidy()
fit1 %>%
 tidy(conf.int = TRUE) %>%
 mutate(term = str_replace_all(term, "_", " ")) %>%
 knitr::kable(digits = c(0, 1, 1, 2, 3, 1, 1),
 caption = "Таб.5 Модель fit1: коэффициенты
регрессии и доверительные интервалы")
# использование функции augment(), удалены
некоторые столбцы из исходного набора данных
fit1 %>%
 augment() %>%
 select(c(2,15:21)) %>%
 head(3) %>%
 knitr::kable(caption = «Таб.6 Модель fit1: реальные и
добавленные данные (1-3 наблюдения)», digits = 3)
```

```
# график
fit1 %>%
 tidy(conf.int = TRUE) %>%
 filter(!str_detect(term, "Intercept")) %>%
 mutate(term = str_replace_all(term, "_", " "),
 term = str_wrap(term, 40)) %>%
 ggplot(aes(term, estimate, ymin = conf.low, ymax =
conf.high)) +
 geom_point() +
 geom_hline(yintercept = 0, linetype = "dotted", color
= "red") +
 geom_errorbar(width = .6) +
 coord_flip() +
 ggtitle("Параметры модели fit1")
```

При изучении таблицы 5 и рисунка 2 можно отметить наличие коэффициентов регрессии, доверительный интервал для которых включает 0. Эти переменные также имеют и высокое значение фактора инфляции дисперсии. Модель может быть

Таб.5 Модель fit1: коэффициенты регрессии и доверительные интервалы

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-5256.0	281.8	-18.65	0.000	-5808.7	-4703.2
Birthlength	134.4	4.1	32.52	0.000	126.3	142.5
Gestational age	41.2	6.9	5.96	0.000	27.7	54.8
Marital statusUnmarried	-66.9	25.3	-2.64	0.008	-116.6	-17.3
Marital statusCohabitant	-18.9	21.7	-0.87	0.385	-61.6	23.7
EducationPrimary (class 1-9)	66.6	35.2	1.89	0.059	-2.5	135.6
EducationSecondary (class 10-11)	16.0	24.2	0.66	0.508	-31.5	63.6
EducationHigher education	21.9	18.7	1.17	0.241	-14.8	58.6
BMI group<18.5	-39.8	34.5	-1.15	0.249	-107.4	27.8
BMI group25-30	41.3	20.5	2.01	0.044	1.0	81.5
BMI group30+	-7.1	28.7	-0.25	0.806	-63.4	49.3
age group<18	271.3	108.6	2.50	0.013	58.3	484.2
age group18-25	-48.7	22.2	-2.20	0.028	-92.1	-5.2
age group30-35	-20.5	20.3	-1.01	0.312	-60.2	19.2
age group35+	43.2	25.7	1.68	0.093	-7.2	93.7
Infant sexFemale	-31.1	16.2	-1.93	0.054	-62.9	0.6
Smokingyes	21.5	22.2	0.97	0.333	-22.0	65.0
Vitamins before pregnancyyes	109.3	319.9	0.34	0.733	-518.2	736.9
Vitamis during pregnancyyes	20.8	17.6	1.18	0.238	-13.8	55.4
Folic acid before pregnancyyes	60.9	186.9	0.33	0.744	-305.6	427.4
Folic acid during pregnancyyes	14.3	17.6	0.81	0.417	-20.3	48.9
Vitamins before pregnancyyes:Vitamis during pregnancyyes	-248.6	329.6	-0.75	0.451	-895.1	397.9
Folic acid before pregnancyyes:Folic acid during pregnancyyes	2.2	197.4	0.01	0.991	-385.0	389.4

Таб.6 Модель fit1: реальные и добавленные данные (1-3 наблюдения)

BIrthweight	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.std.resid
2920	2993.176	33.519	-73.176	0.011	318.949	0	-0.231
3190	3268.219	24.840	-78.219	0.006	318.948	0	-0.246
3850	3675.022	21.716	174.978	0.005	318.925	0	0.550

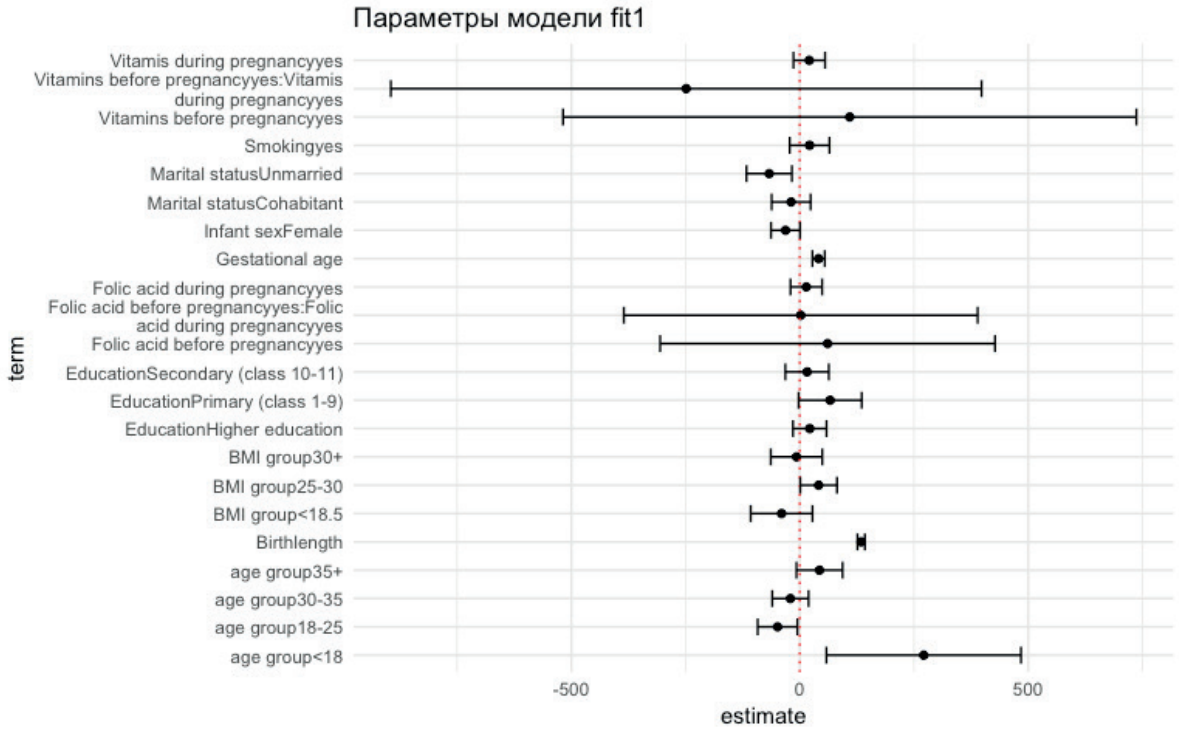

Рис.2 Графическое представление результатов оценки модели fit1

упрощена при исключении из формулы переменных с доверительным интервалом, включающим ноль, то есть незначимо связанным с исходом.

Модель 2

Создание модели 2 путем исключения переменных, включающих ноль для хотя бы одной из категорий в доверительный интервал, и результаты ее оценки в листинге 5

Листинг 5

```
fit2 <- lm(formula = BIrthweight ~ Birthlength +
Gestational_age, data = df1)
fit2 %>%
tidy(conf.int = TRUE) %>%
knitr::kable(digits = c(0, 1, 1, 2, 3, 1, 1),
caption = «Таб.6 Модель fit2: коэффициенты
регрессии и доверительные интервалы»)
fit2 %>%
tidy(conf.int = TRUE) %>%
filter(!str_detect(term, “Intercept”)) %>%
mutate(term = str_replace_all(term, “_”, “ “),
term = str_wrap(term, 40)) %>%
ggplot(aes(term, estimate, ymin = conf.low, ymax =
conf.high)) +
geom_point() +
geom_hline(yintercept = 0, linetype = “dotted”, color =
“red”) +
geom_errorbar(width = .6) +
coord_flip() +
ggtitle(“Параметры модели fit2”)
```

Модель 3

Модель можно улучшить, удалив влияющие

Таб.6 Модель fit2: коэффициенты регрессии и доверительные интервалы

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-5288.0	276.5	-19.13	0	-5830.3	-4745.7
Birthlength	137.8	4.0	34.52	0	129.9	145.6
Gestational_age	37.6	6.8	5.51	0	24.2	51.0

значения, которыми могут быть наблюдения с Cook's distance, превышающей четырехкратное среднее значение Cook's distance. В листинге 6 на графике (рис.3) продемонстрированы "влияющие" значения, пунктирная линия обозначает уровень четырехкратного среднего значения Cook's distance, которые превышают "влияющие" значения, а числа на графике указывают на номера записей. Затем создан набор данных, из которого удалены "влияющие значения", построена модель fit3.

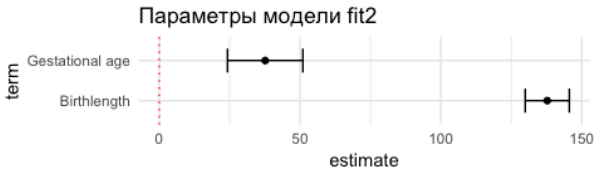

Рис.3 Графическое представление результатов оценки модели

Листинг 6

```
fit2 %>%
 augment() %>%
 mutate(n_row = row_number()) %>%
 ggplot(aes(n_row, .cooksd)) +
 geom_col() +
 geom_hline(yintercept = 4 * mean(cooks.
distance(fit2)),
 color = "red", linetype = "dotted") +
 geom_text(data = augment(fit2) %>%
 mutate(n_row = row_number()) %>%
 arrange(desc(.cooksd)) %>%
 slice(1:9),
 aes(x = n_row, y = .cooksd, label = n_row), size = 2) +
 ggtitle("Модель fit2: Cook's distance")
```

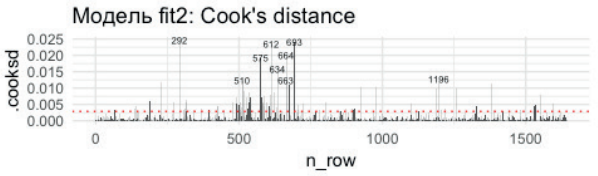

Рис.4 Cook's distance для модели fit2

```
# набор данных с удаленными "влияющими"
значениями
df_d <- fit2 %>%
 augment() %>%
 filter(.cooksd < 4 * mean(.cooksd)) %>%
 select(1:4)
# модель на основе набора данных, с удаленными
"влияющими" значениями
fit3 <- lm(formula = BIrthweight ~ Birthlength +
Gestational_age, data = df_d)
fit3_tidy <- fit3 %>%
 tidy(conf.int = TRUE)
fit3_tidy %>%
 knitr::kable(digits = c(0, 2, 2, 2, 3, 2, 2),
 caption = "Таб.7 Модель fit3: коэффициенты
регрессии и доверительные интервалы")
```

Сравнение моделей

В листинге 7 приведены результаты сравнения трех ранее полученных моделей.

Листинг 7

```
fit_glance <-
 cbind(fit1 %>% glance() %>% select(c(1:5, 8, 9))
%>% t(),
 fit2 %>% glance() %>% select(c(1:5, 8, 9)) %>% t(),
 fit3 %>% glance() %>% select(c(1:5, 8, 9)) %>% t())

colnames(fit_glance) <- c("fit1", "fit2", "fit3")

options(scipen = 999)
fit_glance %>%
 knitr::kable(digits = 3, caption = "Таб.8 Параметры
моделей")
```

Таб.8 Параметры моделей

	fit1	fit2	fit3
r.squared	0.500	0.485	0.602
adj.r.squared	0.493	0.485	0.602
sigma	318.856	321.574	266.560
statistic	73.580	772.132	1171.349
p.value	0.000	0.000	0.000
AIC	23601.667	23609.686	21733.075
BIC	23731.341	23631.298	21754.462

Показатели увеличиваются с 0.5 в модели fit1 до 0.602 в модели fit3. Подобным же образом изменяется показатель adjusted , с 0.493 до 0.602. Показатель sigma (RMSE) уменьшается с 318.856 до 266.56. Показатели F-statistic увеличились с 73.6 до 1171.3. Показатели AIC уменьшаются с 23601.7 в модели fit1 до 21733.1 в модели fit3. Модель fit3 предпочтительнее для дальнейшей оценки и использования.

Таб.7 Модель fit3: коэффициенты регрессии и доверительные интервалы

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-6301.30	246.24	-25.59	0	-6784.31	-5818.29
Birthlength	150.51	3.58	42.06	0	143.49	157.53
Gestational_age	46.17	6.07	7.61	0	34.27	58.07

Решение проблемы мультиколлинеарности при линейном регрессионном анализе.

Мультиколлинерность – наличие линейной зависимости между независимыми переменными. На наличие мультиколлинеарности указывают 1) высокие коэффициенты корреляции между переменными; 2) высокое (допустим, более 4) значение фактора инфляции дисперсии, который может быть вычислен при использовании функции vif из пакета car. В листинге 8 получены значения vif для модели fit3,об условиях отсутствия мультиколлинеарности в нашем примере соблюдается.

Листинг 8

```
car::vif(fit3) %>% as.data.frame() %>%
setNames("vif(fit3)") %>%
knitr::kable(caption = "Таб.9 vif(fit3)", digits = 3)
```

Таб.9 vif(fit3)

	vif(fit3)
Birthlength	1.142
Gestational_age	1.142

Диагностика модели fit3

Анализ остатков

Среднее значение остатков mean(residuals(fit3)) =0. Для оценки нормальности распределения остатков создана диаграмма плотности, на которой пунктирной линией приведена диаграмма плотности для нормального распределения с параметрами существующего распределения остатков модели fit3 (рис.5 слева). Гомоскедастичность распределения остатков позволяют графически определить графики на рис.6 (листинг 9).

Листинг 9

```
g1 <- fit3 %>%
 augment() %>%
 ggplot(aes(.resid)) +
 geom_density() +
 stat_function(fun = dnorm,
 args = list(mean(residuals(fit3)), sd(residuals(fit3))),
 linetype = 3) +
 geom_vline(xintercept = 0, linetype = 2) +
 ggtitle("Диаграмма плотности")
g2 <- fit3 %>%
 augment() %>%
 ggplot(aes(sample = .resid)) + stat_qq() + stat_qq_
line() +
 ggtitle("Квантильная диаграмма")
gridExtra::grid.arrange(g1, g2, nrow = 1)
```

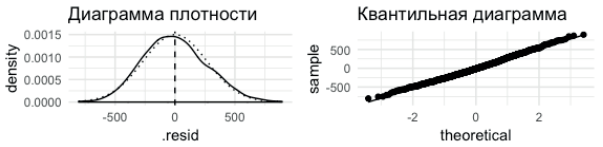

Рис.5 Диагностика модели

```
g1 <- fit3 %>%
 augment() %>%
 ggplot(aes(.fitted, .resid)) +
 geom_point(alpha = .5) +
 geom_hline(yintercept = 0, linetype = "dashed") +
 ggtitle("fitted vs. residuals")
g2 <- fit3 %>%
 augment() %>%
 ggplot(aes(.fitted, .std.resid)) + geom_point(alpha =
.5) +
 geom_hline(yintercept = 0, linetype = "dashed") +
 ggtitle("fitted vs. standardized residuals")
gridExtra::grid.arrange(g1, g2, nrow = 1)
```

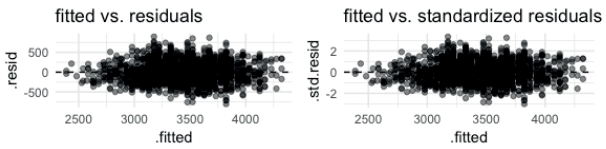

Рис.6 Диагностика модели

Проверка остатков на независимость, автокорреляцию и гомоскедастичность

Выполнение этих тестов представлено в листинге 10. Используются функции из пакетов DescTools и car.

Листинг 10

```
# оценка независимости остатков. runs test - Wald-
Wolfowitz-Test
(dt_runs <- RunsTest(residuals(fit3))) # DescTools
package
##
## Runs Test for Randomness
##
## data: residuals(fit3)
## z = -0.25398, runs = 771, m = 776, n = 775, p-value
= 0.7995
## alternative hypothesis: true number of runs is not
equal the expected number
## sample estimates:
## median(x)
## -13.5181
# тест на аутокорреляцию
(dw <- durbinWatsonTest(fit3)) # car package
## lag Autocorrelation D-W Statistic p-value
## 1 0.01432034 1.970991 0.556
## Alternative hypothesis: rho != 0
# оценка гомоскедастичности - Breusch-Pagan test
```

```
(ncvt <- ncvTest(fit3)) # car package
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.408413, Df = 1, p = 0.23532
```

Изучение остатков показывает:

- нулевая гипотеза о независимости остатков не может быть отклонена (при выполнении Wald-Wolfowitz теста p-value = 0.8);
- нулевая гипотеза об отсутствии автокорреляции остатков (p-value в тесте Дарбина-Уотсона = 0.556) не может быть отклонена;
- не может быть отклонена нулевая гипотеза о гомоскедастичности (p-value в тесте Breusch-Pagan = 0.235).

Таким образом, условия, связанные с остатками, для нашей модели соблюдаются.

Валидация модели. Оценка предсказательной точности модели

Оценка предсказательной точности модели предполагает (листинг 11):

- создание двух наборов данных (тренировочного и тестового);
- наборы данных создаются путём разделения имеющейся таблицы данных: 80% данных составят тренировочный набор и 20% – тестовый;
- на тренировочном наборе создается модель, на основе которой предсказываются возможные значения для тестового набора данных;
- реальные данные тестового набора далее сравниваются с предсказанными по показателю MAPE (средняя абсолютная процентная ошибка).

Листинг 11

```
# validation
set.seed(123)
row_index <- sample(1:nrow(df_d), .8 * nrow(df_d))
train_data <- df_d[row_index, ] # создание тренировочного набора данных
test_data <- df_d[-row_index, ] # создание тестового набора данных
# модель, основанная на тренировочном наборе данных
lm_train <- lm(formula = BIrthweight ~ Birthlength + Gestational_age, data = train_data)
# функция для расчета MAPE с 95% доверительным интервалом
lowl <- function(x) {mean(x) - MeanSE(x) * 1.96}
highl <- function(x) {mean(x) + MeanSE(x) * 1.96}
mape_datci <- function(dat, n=3) {
 dat %>%
 set_names(«x», «y») %>%
 mutate(z = abs(x-y)/x) %>%
 summarise_at(vars(z), c(mean, lowl, highl)) %>%
 as.numeric() %>%
 round(n)
}
# MAPE для тренировочного набора данных
(mape_train <- lm_train %>%
 augment() %>%
 select(BIrthweight, .fitted) %>%
 mape_datci(4))
## [1] 0.0624 0.0598 0.0651
# MAPE для тестового набора данных
(mape_test <- lm_train %>%
 augment(newdata = test_data) %>%
 select(BIrthweight, .fitted) %>%
 mape_datci(4))
## [1] 0.0628 0.0577 0.0679
```

Полученные данные (средние значения, доверительный интервал) указывают на отсутствие различий между MAPE для тренировочного 0.0624 (0.0598 - 0.0651) и тестового набора данных 0.0628 (0.0577 - 0.0679), что говорит о достаточной валидности (достоверности) модели.

Результаты анализа

Результаты анализа указывают на возможность использования модели fit3 для оценки влияния независимых предикторов.

Таб.10 Параметры модели fit3

Переменная	Значение	95% доверительный интервал
Birthlength	150.5	143.5 - 157.5
Gestational_age	46.2	34.3 - 58.1

Масса тела новорожденного (г) = -6301.3 + 150.5 Длина тела(см) + 46.2 Срок беременности (недели)

Интерпретация показателей модели

Построенное регрессионное уравнение говорит о том, что увеличение длины тела новорожденного на 1 см увеличивает вес новорожденного на 150.5 г, 95% доверительный интервал (143.5-157.5), увеличение срока беременности на 1 неделю увеличивает вес новорожденного в среднем на 46.2 г, 95% доверительный интервал (34.3-58.1). Обращаем внимание на то, что данный пример является учебным, поэтому мы намеренно избегаем клинических гипотез и интерпретируем результаты исключительно с точки зрения аналитика, а не практикующего врача. Кроме того, следует отметить, что как масса тела, так и длина являются исходами, которые оцениваются в один и тот же момент времени, поэтому использование одной из переменных в качестве зависимой, а другой в качестве независимой переменной, мы не рекомендуем.

Таб.11 final fit (листинг 12)

	Dependent: BIrthweight		Mean (sd)	Coefficient (univariable)	Coefficient (multivariable)
1	Birthlength	[47,58]	3448.5 (447.9)	145.46 (138.06 to 152.85, p<0.001)	136.62 (128.67 to 144.58, p<0.001)
2	Gestational_age	[35,42]	3448.5 (447.9)	120.11 (103.64 to 136.59, p<0.001)	38.78 (25.32 to 52.24, p<0.001)
4	Infant_sex	Male	3500.5 (442.2)	-	-
3		Female	3393.1 (447.7)	-107.44 (-150.54 to -64.34, p<0.001)	-25.33 (-57.01 to 6.35, p=0.117)

Функции пакета final fit (15) позволяют получать таблицы с результатами регрессионного анализа относительно простым путем. Необходимо указать независимые (explanatory) и зависимую (dependent) переменные и применить функцию final fit (листинг 12). Результатом выполнения функции будет таблица 11, включающая нескорректированные (crude) и скорректированные (adjusted) коэффициенты с доверительными интервалами и p-value.

Листинг 12

```
library(finalfit)
explanatory <- c("Birthlength", "Gestational_age",
"Infant_sex")
dependent <- "BIrthweight"
df1 %>%
finalfit(dependent, explanatory) %>%
knitr::kable()
```

Результаты множественного линейного регрессионного анализа в статьях представляют чаще всего в виде таблицы с нескорректированными и скорректированными коэффициентами и доверительными интервалами для них. Этот способ является оптимальным при решении задач об оценке независимого влияния независимых переменных на переменную отклика. Для похожих для представленного примера задач данные можно представлять как, например, в (16). Когда регрессионная модель создается для прогнозирования, целесообразно представлять само уравнение регрессии, а также коэффициент детерминации, однако в медицине регрессионные модели используются относительно редко из-за большой изменчивости биологических признаков и мультифакториальной природы большинства изучаемых в медицине состояний.

Использованный в работе файл с набором данных и скрипт с кодом доступны на сайте https://github.com/valegoshin/Paper_Scripts.

Список литературы:

1. Gelman A, Hill J. Data Analysis Using Regression and Multilevel. Hierarchical Models. Cambridge 2007.
2. James G, Hastie T, Witten D. An Introduction to Statistical Learning with Applications in R. Springer 2013.
3. Darlington RB, Hayes AF. Regression Analysis and Linear Models. The Guilford Press 2017.
4. Гржибовский АМ. Простой линейный регрессионный анализ. Экология человека 2018;10:55–64.
5. Гржибовский АМ, Унгуряну ТН, Горбатова МА. Корреляционный и однофакторный линейный регрессионный анализ с использованием программного обеспечения SPSS и Stata. Наркология 2017;9:52–69.
6. Шарашова ЕЕ, Холматова КК, Горбатова МА, Гржибовский АМ. Применение множественного линейного регрессионного анализа в здравоохранении. Наука и здравоохранение 2017;3:5–31.
7. Егошин ВЛ, Иванов СВ, Саввина НВ, Капанова ГЖ, Гржибовский АМ. Основы работы в программной среде R при анализе исследовательских данных. Экология человека 2018;7:55–64.
8. Егошин ВЛ, Иванов СВ, Саввина НВ, Капанова ГЖ, Жамалиева ЛМ, Гржибовский АМ. Анализ категориальных данных с использованием программной среды R. Экология человека. 2019;1:51–64 .
9. Егошин ВЛ, Иванов СВ, Саввина НВ, Калмаханов СБ, Гржибовский АМ. Визуализация исследовательских данных с использованием программной среды R. Экология человека 2018;8:52–64.
10. Егошин ВЛ, Иванов СВ, Саввина НВ, Капанова ГЖ, Гржибовский АМ. Расчет показателей описательной статистики с использованием программной среды R. Экология человека 2018;9:55–64.

Spisok literatury:

1. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/ Hierarchical Models. Cambridge 2007.
2. James G, Hastie T, Witten D. An Introduction to Statistical Learning with Applications in R. Springer 2013.
3. Darlington RB, Hayes AF. Regression Analysis and Linear Models. The Guilford Press 2017.
4. Grjibovski AM. Simple linear regression analysis. Ekologiya cheloveka (Human Ecology) 2008;10:55–64.
5. Grjibovski AM, Unguryanu TN, Gorbatova MA. Correlation analysis and simple linear regression using SPSS and Stata software. Narcology 2017;9:52–69.
6. Sharashova EE, Kholmatova KK, Gorbatova MA, Grjibovski AM. Application of the multivariable linear regression analysis in healthcare using SPSS software. Science and healthcare 2017;3:5–31.
7. 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.
8. Egoshin VL, Ivanov SV, Savvina NV, Kapanova GZ, Zhamaliyeva LM, Grjibovski AM. Analysis of Categorical Variables Using R. Ekologiya cheloveka [Human Ecology]. 2019;1:51–64.
9. Egoshin VL, Ivanov SV, Savvina NV, Kalmakhanov SB, Grjibovski AM. Visualization of Biomedical Data Using R. Ekologiya cheloveka [Human Ecology] 2018;8:52–64.
10. Egoshin VL, Ivanov SV, Savvina NV, Kapanova GZh, Grjibovski AM. Descriptive statistics using R. Ekologiya cheloveka [Human Ecology] 2018;9:55–64.
11. Egoshin VL, Ivanov SV, Savvina NV, Kalmakhanov SB, Zhamaliyeva LM, Grjibovski AM. Analysis of Continuous Data Using R. Ekologiya cheloveka [Human Ecology] 2018;11:51–64.
12. Egoshin VL, Ivanov SV, Savvina NV, Ermolaev AR, Mamyrbekova

11. Егошин ВЛ, Иванов СВ, Саввина НВ, Калмаханов СБ, Жамалиева ЛМ, Гржибовский АМ. Анализ непрерывных данных с использованием программной среды R. Экология человека 2018;11:51–64.
12. Егошин ВЛ, Иванов СВ, Саввина НВ, Ермолаев АР, Мамырбекова СА, Жамалиева ЛМ, Гржибовский АМ. Корреляционный и простой линейный регрессионный анализ с использованием программной среды R. Экология человека. 2018;12:55–64.
13. Kabakoff R. Data visualization with R. https://rkabacoff.github.io/datavis/. Accessed 12 January 2019.
14. Усынина АА, Одланд ИО, Пылаева ЖА, Пастбина ИМ, Гржибовский АМ. Регистр родов Архангельской области как важный информационный ресурс для науки и практического здравоохранения ”Экология человека 2017;2:58–64.
15. Harrison E. Tables Gallery, 2019. https://finalfit.org/articles/tables_gallery.html. Accessed 12 January 2019.
16. Vikanes ÅV, Støer NC, Magnus P, Grjibovski AM. Hyperemesis gravidarum and pregnancy outcomes in the Norwegian Mother and Child Cohort - a cohort study. BMC Pregnancy and Childbirth 2013;13:169.

SA, Zhamaliyeva LM, Grjibovski AM. Correlation and simple regression analysis using R. Ekologiya cheloveka [Human Ecology] 2018;12:55–64.
13. Kabakoff R. Data visualization with R. https://rkabacoff.github.io/datavis/. Accessed 12 January 2019.
14. Usynina AA, Odland Jon Øyvind, Pylaeva ZhA, Pastbina IM, Grjibovski AM. Arkhangelsk County Birth Registry as an Inportant Source of Information for Research and Healthcare. Ekologiya cheloveka [Human Ecology] 2017;2:58–64.
15. Harrison E. Tables Gallery, 2019. https://finalfit.org/articles/tables_gallery.html. Accessed 12 January 2019.
16. Vikanes ÅV, Støer NC, Magnus P, Grjibovski AM. Hyperemesis gravidarum and pregnancy outcomes in the Norwegian Mother and Child Cohort - a cohort study. BMC Pregnancy and Childbirth 2013;13:169.