УДК 519.25

КОРРЕЛЯЦИОННЫЙ И ПРОСТОЙ ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОЙ СРЕДЫ R

© 2018 г. [1]**В. Л. Егошин,** [2]**С. В. Иванов,** [3]**Н. В. Саввина,** [3]**А. Р. Ермолаев,** [4]**С. А. Мамырбекова,** [5]**Л. М. Жамалиева,** [3-6]**А. М. Гржибовский**

[1]Павлодарский филиал Государственного медицинского университета г. Семей, г. Павлодар, Казахстан; [2]Первый Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова, г. Санкт-Петербург; [3]Северо-Восточный федеральный университет им. М. К. Аммосова, г. Якутск; [4]Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан; [5]Западно-Казахстанский государственный медицинский университет им. Марата Оспанова, г. Актобе, Казахстан; [6]Северный государственный медицинский университет, г. Архангельск

В статье рассмотрены основные алгоритмы работы в программной среде R, используемые для проведения корреляционного и однофакторного линейного регрессионного анализа. Представлены базисные подходы к интерпретации результатов анализа и оценке статистических регрессионных моделей.

Ключевые слова: корреляционный анализ, однофакторный линейный регрессионный анализ, R

CORRELATION AND SIMPLE REGRESSION ANALYSIS USING R

[1]**V. L. Egoshin,** [2]**S. V. Ivanov,** [3]**N. V. Savvina,** [3]**A. R. Ermolaev,** [4]**S. A. Mamyrbekova,** [5]**L. M. Zhamaliyeva,** [3-6]**A. M. Grjibovski**

[1]Semey State Medical University, Pavlodar Campus, Pavlodar, Kazakhstan; [2]Pavlov First St. Petersburg State Medical University, St. Petersburg, Russia; [3]North-Eastern Federal University, Yakutsk, Russia; [4]Al-Farabi Kazakh National University, Almaty, Kazakhstan; [5]West Kazakhstan Marat Ospanov State Medical University, Aktobe, Kazakstan; [6]Northern State Medical University, Arkhangelsk, Russia

The article presents basic algorithms of R software using for correlation analysis and simple linear regression. Basic approaches to the interpretation of the results of analysis and evaluation of statistical regression models are presented.

Key words: correlation analysis, simple linear regression, R

Библиографическая ссылка:

Егошин В. Л., Иванов С. В., Саввина Н. В., Ермолаев А. Р., Мамырбекова С. А., Жамалиева Л. М., Гржибовский А. М. Корреляционный и простой линейный регрессионный анализ с использованием программной среды R // Экология человека. 2018. № 12. С. 55–64.

Egoshin V. L., Ivanov S. V., Savvina N. V., Ermolaev A. R., Mamyrbekova S. A., Zhamaliyeva L. M., Grjibovski A. M. Correlation and simple regression analysis using R. *Ekologiya cheloveka* [Human Ecology]. 2018, 12, pp. 55-64.

В процессе анализа результатов научных исследований для изучения связей между переменными достаточно часто требуется создание статистических моделей. Под моделью может пониматься абстрактное представление реальности в какой-либо форме (математической, физической, символической, графической), предназначенное для представления определенных аспектов этой реальности и позволяющее получить ответы на изучаемые вопросы [3]. Как правило, статистические модели описывают связи между случайными переменными, при этом связи между двумя количественными переменными изучаются методами корреляционного и простого линейного регрессионного анализа. [7, 10].

Для примера проведения анализа в статье были использованы модифицированные данные Архангельского областного регистра родов [4]. Подготовка данных к анализу представлена на рис. 1 (листинг 1). В ходе подготовки данных выбираются переменные и удаляются пропущенные значения. Обработка выбросов в данном случае предполагает их преобразование в категорию «NA» с последующим удалением.

Листинг 1

```
# импорт из файла
df <- foreign::read.spss(«Simulated_sample.sav", to.data.
frame = TRUE)

# выбор переменных в таблице данных
df <- df %>%
filter ((Maternal_age > 20 & Maternal_age <= 25) &
Gestational_age > 36) %>%
select (Maternal_height, Maternal_weight,
BIrthweight, Birthlength) %>% drop_na ()
# функция преобразования выбросов в NA
Outl_NA <- function(x) {
x [which(x %in% boxplot.stats(x)$out)] <- NA; x
}
# boxplot.stats boxplot(x, plot = FALSE)$out
# функция удаления записей с выбросами из таблицы
данных
df_out <- function(dat) {
na.omit(data.frame(apply(dat, 2, function(x) Outl_NA(x))))
}
# таблица данных после удаления выбросов
```

```
dfs <- df_out(df)
rownames (dfs) <- 1:nrow (dfs)
```

Рис. 1. Подготовка данных для анализа

В результате проведения подготовки данных количество отобранных записей – 431: женщины в возрасте от 20 до 25 лет включительно, родившие в срок более 36 недель.

Анализ выполнен в программной среде R 3.5.0. Использовались функции базового пакета и пакетов Hmisc и car, для вывода результатов тестов использовались функции пакетов knitr и pander.

Корреляционный анализ

Корреляционный анализ – метод, позволяющий оценить силу и направление связи между переменными. Вопросы применения метода приведены в пособиях по статистике [6, 11]. При проведении корреляционного анализа следует помнить, что корреляция однозначно не подразумевает причинно-следственных связей [5, 15].

При статистической обработке результатов исследования используются следующие методы корреляционного анализа:

– Метод Пирсона (Pearson correlation – r), измеряющий линейную зависимость между двумя переменными (x, y). Это тест параметрической корреляции, поскольку он подразумевает нормальное распределение данных.

– Методы Спирмена (Spearman ρ) и Кендалла (Kendall τ) – непараметрические методы, использующие ранговую корреляцию и конкордантные / дискордантные пары, соответственно.

Наиболее часто используется метод Пирсона, но его следует применять только при соблюдении следующих условий [1]:

– обе переменные являются количественными и непрерывными;

– как минимум один из признаков имеет нормальное распределение;

– зависимость между переменными носит линейный характер;

– гомоскедастичность (вариабельность одной переменной не зависит от значений другой переменной);

– независимость переменных;

– парность наблюдений (признак x и признак y относятся к одним и тем же случаям);

– достаточный объем выборки, включающий как минимум 25 наблюдений;

– для адекватной проекции расчетов на генеральную совокупность выборка должна быть репрезентативной.

При выполнении корреляционного анализа в R выполняются следующие действия:

– оценка нормальности распределения переменных: формальные тесты (Шапиро – Уилка, Андерсона – Дарлинга) и графики (квантильная диаграмма, гистограмма или диаграмма плотности);

– точечная диаграмма – для оценки линейности зависимости и гомоскедастичности;

– определение коэффициента корреляции;

– определение достигнутого уровня значимости (p-value);

– создание корреляционной матрицы и коррелограммы.

Коэффициент корреляции может иметь значения от –1 до +1. При этом значение –1 означает полную отрицательную корреляцию, +1 – полную положительную корреляцию, а 0 – отсутствие корреляции. Помимо непосредственного значения коэффициента корреляции необходимо также оценивать доверительный интервал и p-value.

В базовом пакете R для определения коэффициента корреляции используются две функции: cor и cor.test.

Функция cor позволяет вычислить коэффициент корреляции с использованием методов Пирсона, Спирмена или Кендалла. По умолчанию используется метод Пирсона. Формат записи: cor (x, y = NULL, use = “everything”, method = c (“pearson”, “kendall”, “spearman”)), где x может быть числовым вектором, матрицей или таблицей данных; числовой вектор y используется в случае, если x – также числовой вектор.

Функция cor.test может быть выполнена для двух числовых векторов, при этом функция возвращает значение статистики, коэффициента корреляции, p-value, доверительный интервал. Может быть выполнена в формате cor.test (x, y) или в формате cor. test (~ x + y, data, method = c (“pearson”, “kendall”, “spearman”))

Применив функции cor и pairs к матрице или таблице данных, можно получить матрицу значений коэффициента корреляции и графическое представление (скаттерограммы) изучаемых переменных, что позволяет оценить линейность зависимости между переменными и гомоскедастичность (рис. 2 – листинг 2).

Листинг 2

```
round(cor(dfs),3)
## Maternal_height Maternal_weight BIrthweight Birthlength
## Maternal_height 1.000 0.332 0.233 0.214
## Maternal_weight 0.332 1.000 0.114 0.092
## BIrthweight 0.233 0.114 1.000 0.686
## Birthlength 0.214 0.092 0.686 1.000
pairs(dfs)
```

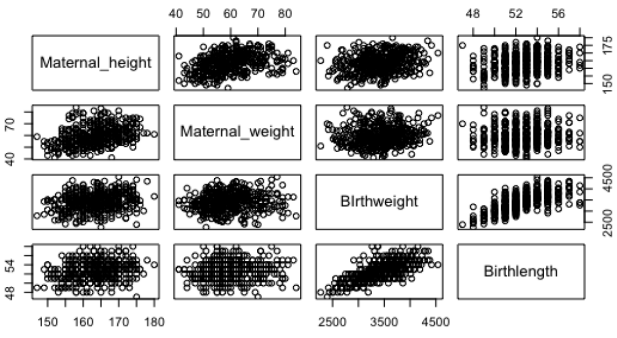


Рис. 2. Результаты применения функции pairs

Оценка нормальности распределения переменных предполагает выполнение формальных тестов и оценку диаграмм плотности и квантильных диаграмм изучаемых переменных. Алгоритм данного анализа и полученные результаты представлены на рис. 3 (листинг 3).

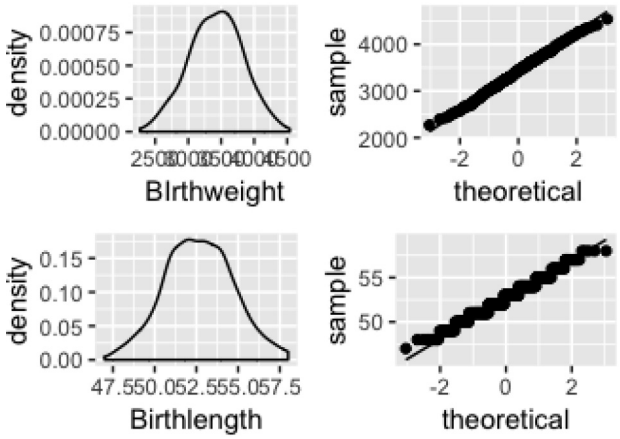
Листинг 3

```
# функция
formal_test <- function(dat) {
shw_pvalue <- apply(dat[sapply(dat, is.numeric)], 2,
function(x) shapiro.test(x)$p.value)
ad_pvalue <- apply(dat[sapply(dat, is.numeric)], 2,
function(x) nortest::ad.test(x)$p.value)
cbind(shw_pvalue, ad_pvalue)
}
knitr::kable(as.data.frame(formal_test(dfs)), digits = c( 9, 9),
caption = 'Результаты выполнения тестов', format =
'pandoc')
```

Результаты выполнения тестов

	shw_pvalue	ad_pvalue
Maternal_height	0.007969014	0.002280250
Maternal_weight	0.000049021	0.000015483
BIrthweight	0.797425177	0.851300783
Birthlength	0.000003111	0.000000000

```
# Диаграммы для переменных BIrthweight и Birthlength
g1 <- ggplot(dfs, aes(BIrthweight)) + geom_density()
g2 <- ggplot(dfs, aes(sample = BIrthweight)) + stat_qq() +
stat_qq_line()
g3 <- ggplot(dfs, aes(Birthlength)) + geom_density()
g4 <- ggplot(dfs, aes(sample = Birthlength)) + stat_qq() +
stat_qq_line()
gridExtra::grid.arrange(g1, g2, g3, g4, nrow = 2)
```




```
# Диаграммы для переменных Maternal_height и Maternal_
weight
g1 <- ggplot(dfs, aes(Maternal_height)) + geom_density()
g2 <- ggplot(dfs, aes(sample = Maternal_height)) + stat_qq()
+ stat_qq_line()
g3 <- ggplot(dfs, aes(Maternal_weight)) + geom_density()
g4 <- ggplot(dfs, aes(sample = Maternal_weight)) + stat_qq()
+ stat_qq_line()
gridExtra::grid.arrange(g1, g2, g3, g4, ncol = 2)
```

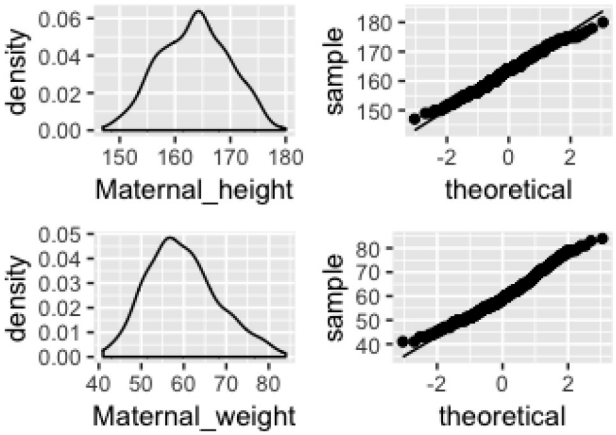


Рис. 3. Результаты анализа нормальности распределения для переменных BIrthweight и Birthlength, Maternal_height и Maternal_weight

Таким образом, результаты тестов на нормальность распределения и соответствующие графики допускают возможность использования метода Пирсона при изучении корреляция между переменными BIrthweight и Birthlength.

Нулевая гипотеза при изучении корреляции между переменными предполагает, что коэффициент корреляции равен нулю. Для оценки корреляции между переменными BIrthweight и Birthlength применим метод Пирсона, а для оценки корреляции между переменными Maternal_height и Maternal_weight — методы Спирмена и Кендалла (рис. 4 — листинг 4). Использование функции pander из пакета pander делает представление результатов теста более подходящим для публикации.

Листинг 4

```
options(scipen = -3)
cor.test(~ BIrthweight + Birthlength, dfs, method = 'pearson')
##
## Pearson's product-moment correlation
##
## data: BIrthweight and Birthlength
## t = 19.549, df = 429, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.6329680 0.7333003
## sample estimates:
## cor
## 0.6863866
pander::pander(cor.test(~ BIrthweight + Birthlength, dfs,
method = 'pearson'))
```

Pearson's product-moment correlation: BIrthweight and Birthlength

Test statistic	df	P value	Alternative hypothesis	cor
19.55	429	2.551e-61 * * *	two.sided	0.6864

```
options(scipen = 999)
pander::pander(cor.test(~ Maternal_height + Maternal_weight,
dfs, method = 'spearman'), digits = c(0, 9, 0, 4))
```

Spearman's rank correlation rho: Maternal_height and Maternal_weight

Test statistic	P value	Alternative hypothesis	rho
8604559	0.00000000000002935 * * *	two.sided	0.3552

pander::**pander**(**cor.test**(~ Maternal_height + Maternal_weight, dfs, method = 'kendall'), digits = c(4, 9, 0, 4))
Kendall's rank correlation tau: Maternal_height and Maternal_weight

Test statistic	P value	Alternative hypothesis	tau
7.416	0.0000000000001208 * * *	two.sided	0.2491

Рис. 4. Оценка корреляции между переменными BIrthweight и Birthlength и между переменными Maternal_height и Maternal_weight

В процессе выполнения анализа можно создать комбинированную корреляционную матрицу, которая будет включать не только значения коэффициентов корреляции, но и уровень их статистической значимости. Функция cor позволяет выводить коэффициент корреляции для нескольких переменных, функция cor.test представляет большой набор данных для двух переменных. Функция rcorr из пакета Hmisc позволяет представить данные о коэффициенте корреляции и уровне значимости для нескольких переменных в таблице данных (рис. 5 – листинг 5).

Листинг 5
```
(cor2 <- Hmisc::rcorr(as.matrix(dfs)))
## Maternal_height Maternal_weight BIrthweight Birthlength
## Maternal_height    1.00          0.33        0.23        0.21
## Maternal_weight    0.33          1.00        0.11        0.09
## BIrthweight        0.23          0.11        1.00        0.69
## Birthlength        0.21          0.09        0.69        1.00
##
## n= 431
##
##
## P
## Maternal_height Maternal_weight BIrthweight Birthlength
## Maternal_height    0.0000        0.0000      0.0000
## Maternal_weight    0.0000        0.0176      0.0556
## BIrthweight        0.0000        0.0176      0.0000
## Birthlength        0.0000        0.0556      0.0000
```

Рис. 5. Создание комбинированной корреллляционной матрицы.

В отличие от приведенного выше примера, использование функции Hmisc::rcorr позволяет представить данные в более удобном для восприятия виде (рис. 6 – листинг 6).

Листинг 6
```
# преобразование листа с элементами r (коэффициент корреляции)
rdfs <- reshape2::melt(cor2$r)
rdfs <- rename(rdfs, cor_coef = value)

# преобразование листа с элементами P (p-value)
rdfsp <- reshape2::melt(cor2$P)
rdfsp <- rename(rdfsp, pvalue = value)

# объединение и сортировка по значению коэффициента корреляции
rpf <- na.omit(left_join(rdfs, rdfsp, by = c('Var1', 'Var2'))) %>%
arrange(desc(cor_coef))

# таблице
knitr::kable(rpf, digits = c(4, 9),
caption = 'Комбинированная корреляционная матрица',
format = 'pandoc')
```

Комбинированная корреляционная матрица

Var1	Var2	cor_coef	pvalue
Birthlength	BIrthweight	0.6864	0.000000000
BIrthweight	Birthlength	0.6864	0.000000000
Maternal_weight	Maternal_height	0.3321	0.000000000
Maternal_height	Maternal_weight	0.3321	0.000000000
BIrthweight	Maternal_height	0.2333	0.000000975
Maternal_height	BIrthweight	0.2333	0.000000975
Birthlength	Maternal_height	0.2140	0.000007428
Maternal_height	Birthlength	0.2140	0.000007428
BIrthweight	Maternal_weight	0.1143	0.017572783
Maternal_weight	BIrthweight	0.1143	0.017572783
Birthlength	Maternal_weight	0.0923	0.055558618
Maternal_weight	Birthlength	0.0923	0.055558618

Рис. 6. Использование функции Hmisc::rcorr

Корреляционная матрица может быть представлена также в виде «вафельной диаграммы» (рис. 7 – листинг 7).

Листинг 7
```
g1 <- ggplot(rdfs, aes(Var1, Var2, fill = cor_coef)) +
geom_tile(color = 'grey50') +
geom_text(aes(label = round(cor_coef,3)), color = 'white', size = 3) +
scale_fill_gradient(low = 'lightgrey', high = 'black') +
theme_minimal() +
theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
labs(x = '', y = '', title = 'Correlation coefficient')

g2 <- ggplot(rdfsp, aes(Var1, Var2, fill = pvalue)) +
geom_tile(color = 'grey50') +
scale_fill_gradient(low = 'lightgrey', high = 'black') +
geom_text(aes(label = round(pvalue,3)), color = 'white', size = 3) +
theme_minimal() +
theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
labs(x = '', y = '', title = 'p-value')

gridExtra::grid.arrange(g1, g2, nrow = 1)
```

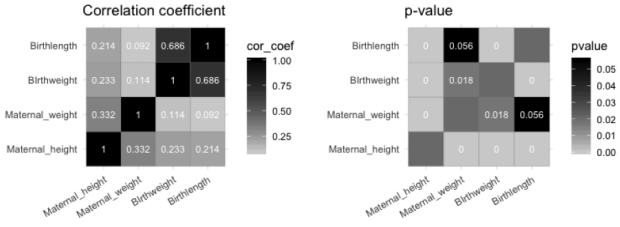


Рис. 7. Создание коррелограмм для коэффициентов корреляции (слева) и уровней значимости (справа)

Используя функции пакета corrplot, можно получить много цветных коррелограм, выполненных в различном стиле [9, 14].

Простая линейная регрессия

Регрессионный анализ широко применяется в практике статистического анализа данных [2, 8, 11, 13]. Следует отметить, что если модели корреляционного анализа позволяют изучить силу и направление связи

между переменными, то регрессионный анализ позволяет прогнозировать значения зависимой переменной (переменной отклика) по известным значениям независимой переменной (предиктора). Таким образом, в общем представлении модель простой линейной регрессии предназначена для предсказания значений количественной зависимой переменной по значениям одной количественной независимой переменной.

Цель регрессионного анализа – определить математическую формулу связи между зависимой переменной (Y) и независимой переменной (X), причем данная формула может быть использована для предсказания значения Y при известном значении X. Формула уравнения линейной регрессии следующая:

$$Y = \beta_0 + \beta_1 \cdot X_i + \varepsilon,$$

где β_0 и β_1 – коэффициенты регрессии, определяемые в ходе выполнения анализа, ε – ошибка, предположительно имеющая в случае линейной регрессии нормальное распределение со средним значением, равным нулю: $\varepsilon \sim N(0,\ \sigma^2)$.

В R при построении модели простой линейной регрессии используется следующий формат функции lm: lm(dependent_variable ~ independent_variable, data).

Функция summary возвращает данные о полученной модели линейной регрессии, а функции coef, confint возвращают значения коэффициентов регрессии и доверительные интервалы. Функция anova представляет таблицу, с помощью которой можно оценить значимость модели (рис. 8 – листинг 8).

Листинг 8

```
fit <- lm(BIrthweight ~ Birthlength, dfs)

# данные о созданной модели
summary(fit)
##
## Call:
## lm(formula = BIrthweight ~ Birthlength, data = dfs)
##
## Residuals:
## Min 1Q Median 3Q Max
## -883.24 -195.65 -15.48 184.35 937.97
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3771.833 368.244 -10.24
<0.0000000000000002 ***
## Birthlength 136.553 6.985 19.55 <0.0000000000000002 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 298 on 429 degrees of freedom
## Multiple R-squared: 0.4711, Adjusted R-squared: 0.4699
## F-statistic: 382.2 on 1 and 429 DF, p-value: <
0.00000000000000022
pander::pander(summary(fit), digits = c(0,2,3,0,0,0,0,4,4))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3772	368	-10.2	0
Birthlength	137	7	19.5	0

Fitting linear model: BIrthweight ~ Birthlength

Observations	Residual Std. Error	R^2	Adjusted R^2
431	298	0.4711	0.4699

```
# коэффициенты регрессии
coef(fit)
## (Intercept) Birthlength
## -3771.833 136.553

# доверительный интервал
confint(fit, level = .95)

## 2.5 % 97.5 %
## (Intercept) -4495.6187 -3048.0468
## Birthlength 122.8235 150.2825

pander::pander(round(cbind(coef(fit), confint(fit))), caption =
'Коэффициенты регрессии и доверительные интервалы')
```

Коэффициенты регрессии и доверительные интервалы

		2.5 %	97.5 %
(Intercept)	-3772	-4496	-3048
Birthlength	137	123	150

```
pander::pander(anova(fit), digits = c(0, 0, 0, 0, 0, 6))
```

Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Birthlength	1	33940248	33940248	382	0
Residuals	429	38100368	88812	NA	NA

Рис. 8. Линейный регрессионный анализ

При проведении линейной регрессии оценивается нулевая гипотеза, предполагающая, что коэффициенты регрессии не отличаются от нуля (показатели Pr(>|t|) в разделе «Коэффициенты» в последней колонке). F-test оценивает нулевую гипотезу о равенстве нулю всех коэффициентов регрессии (показатель p-value в последней строке таблицы), альтернативная гипотеза – хотя бы один коэффициент не равен нулю.

Таким образом, формула уравнения линейной регрессии будет следующая: Масса тела новорожденного (г) = −3 771,8 + 136,6 × длина тела новорожденного (см). Например, новорожденный с длиной тела 52 см будет иметь массу тела 3 329 г, нижняя и верхняя границы 95 % доверительный интервал для данного значения будут равны 1 891 г и 4 767 г соответственно.

Для оценки полученной модели используется ряд показателей.

Наиболее простой показатель – коэффициент детерминации R^2 (Multiple R-squared в листинге 8). Он показывает, какая доля изменчивости зависимой переменной обусловлена моделью. Скорректированный (adjusted) R^2 (Adjusted R-squared в листинге 8) показывает значение доли изменчивости зависимой переменной, скорректированное на количество независимых переменных.

Коэффициенты AIC (Akaike's information criterion) и BIC (Bayesian information criterion) служат для оценки пригодности статистической модели и используются при выборе наиболее предпочтительной модели, если были созданы несколько моделей. Формат данных функций в R следующий: AIC(model), BIC(model).

Показатель MSE (mean squared error) является аналогом дисперсии для остатков в модели линейной регрессии, а показатель RMSE (root mean squared error) – аналогом стандартного отклонения. И тот и

другой показатель также используются при выборе моделей. Формулы их расчета:

$$MSE = \frac{1}{n\sum \mathbf{residual}^2}, \; RMSE = \sqrt{\frac{1}{n\sum \mathbf{residual}^2}}$$

Данные показатели могут быть вычислены в R с помощью следующих формул:

MSE <- mean(residuals(linear_model)**2),

RMSE <- sqrt(mean(residuals(linear_model)**2))

Показатель MAPE (Mean absolute percentage error) – средняя абсолютная процентная ошибка, которая используется для оценки предсказательных возможностей регрессионной модели и может быть рассчитана по формуле:

$$MAPE = mean\left(\frac{abs(\mathbf{actual} - \mathbf{predict})}{\mathbf{actual}}\right),$$

где actual – реальные значения переменной, predict – предсказанные на основании созданной модели значения.

Оценка модели с помощью вышеперечисленных показателей представлена в табл. 1.

Таблица 1

Использование показателей для оценки линейной регрессионной модели

Статистика	Оценка
R-Squared	Выше – лучше (> 0,70)
Adj R-Squared	Выше – лучше
F-statistic	Выше – лучше
Std. Error	Ближе к нулю – лучше
t-statistic	Больше 1,96 для p-value меньше 0,05
AIC	Ниже – лучше
BIC	Ниже – лучше
MAPE (Mean absolute percentage error)	Ниже – лучше
MSE (Mean squared error)	Ниже – лучше
RMSE (Root mean squared error)	Ниже – лучше

Для корректного выполнения простого линейного регрессионного анализа должны быть соблюдены следующие условия:

1. Линейность связи между зависимой и независимой переменными.

2. Определенные характеристики остатков:

– нормальность распределения остатков;

– нулевое среднее значение остатков;

– независимость остатков;

– отсутствие аутокорреляции остатков;

– гомоскедастичность остатков, равная дисперсия.

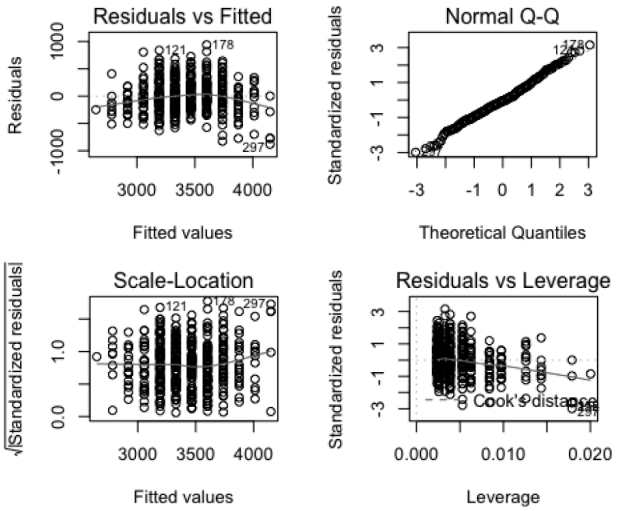
Отсутствие «влияющих» данных.

Следует отметить, что при простой линейной регрессии мультиколинеарность не оценивается.

Для оценки условий выполнения линейного регрессионного анализа используются графические методы (рис. 9 – листинг 9) и тесты.

Листинг 9

```
par(mfrow = c(2,2))
plot(fit)
```




```
par(mfrow = c(1,1))
```

Рис. 9. Результат применения функции plot(fit)

На рис. 9 объединены четыре графика: первый (вверху слева) residuals vs. fitted values – скаттерограмма между остатками и подогнанными значениями; второй (вверху справа) – квантильная диаграмма, оценивает нормальность распределения остатков; третий (внизу слева) scale-Location – скаттерограмма между стандартизованными остатками и подогнанными значениями; четвертый график (внизу справа) Cook's distance показывает точки, обладающие большим влиянием на регрессию (leverage points).

Для оценки гомоскедастичности изучаются первый и третий графики (см. рис. 9), при отсутствии гетероскедастичности отмечается полностью случайное, равное распределение точек и плоская красная линия [12].

Оценка линейности связи между переменными в модели может быть проведена путем построения точечной диаграммы (рис. 10 – листинг 10).

Листинг 10

```
ggplot(dfs, aes(Birthlength, BIrthweight)) +
geom_point() + geom_smooth(method = 'lm', se = FALSE)
```

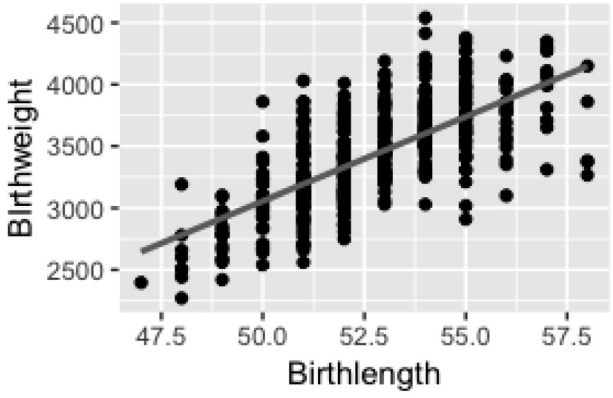


Рис. 10. Скаттерограмма, представляющая зависимость между изучаемыми переменными

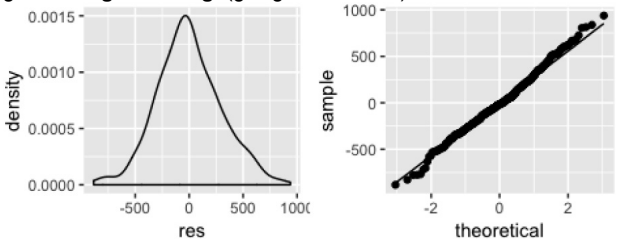
Графики и тесты, предназначенные для изучения остатков модели, представлены на рис. 11 (листинг 11).

Листинг 11

```
mean(residuals(fit)) # среднее значение остатков
## [1] -0.00000000000000005680877
sht <- shapiro.test(residuals(fit)); pander::pander(sht)
Shapiro-Wilk normality test: residuals(fit)
```

Test statistic	P value
0.9939	0.08054

```
# визуализация рис.7
ddf <- data.frame(res = residuals(fit))
g1 <- ggplot(ddf, aes(res)) + geom_density()
g2 <- ggplot(ddf, aes(sample = res)) + stat_qq() + stat_qq_line()
gridExtra::grid.arrange(g1, g2, nrow = 1)
```




```
# оценка независимости остатков. runs test – Wald-Wolfowitz-Test
(dt_runs <- DescTools::RunsTest(residuals(fit)))
##
## Runs Test for Randomness
##
## data: residuals(fit)
## z = 0.0027976, runs = 217, m = 218, n = 213, p-value = 0.9978
## alternative hypothesis: true number ofruns is not equal the expected number
## sample estimates:
## median(x)
## -15.47766

# тест на аутокорреляцию
(dw <- durbinWatsonTest(fit))

## lag Autocorrelation D-W Statistic p-value
## 1 -0.01590408 2.031162 0.722
## Alternative hypothesis: rho != 0

# оценка гомоскедастичности – Breusch-Pagan test
(ncvt <- ncvTest(fit))
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 4.967327 Df = 1 p = 0.02583053
```

Рис. 11. Графики и тесты, предназначенные для изучения остатков модели. Диаграмма плотности остатков представлена слева, квантильная диаграмма остатков – справа

Изучение остатков показало следующие результаты:

– среднее значение остатков равно нулю;

– p-value в тесте Шапиро – Уилка равно 0,081;

– нулевая гипотеза о независимости остатков не может быть отклонена (при выполнение Wald-Wolfowitz теста p-value = 0,998);

– нулевая гипотеза об отсутствии аутокорреляции остатков (p-value = 0,722 в тесте Дарбина –Уотсона) не может быть отклонена;

– может быть отклонена нулевая гипотеза о гомоскедастичности (p-value = 0,026 в тесте Breusch-Pagan).

При оценке модели необходимо также определять наличие «влияющих» значений. Одним из способов определения таких значений является использование Cook's distance. Значимыми («влияющими») считаются значения, превышающие величину, равную 4/(n-k-1), где n – число наблюдений, k – число коэффициентов регрессии. В R можно создать таблицу данных, в которую будут включены только записи с «влияющими» значениями, и представить данные в виде графика (рис. 12 – листинг 12).

Листинг 12

```
(cutoff <- 4/((nrow(dfs)-length(fit$coefficients)-1)))
## [1] 0.009345794

# (cutoff <- 4 * mean(cooks.distance(fit)))

# создание таблицы данных с "влияющими" записями
df_cd <- dfs %>% select(Birthweight, Birthlength) %>%
mutate(n_dfs = rownames(dfs), # столбец с номерами строк таблицы dfs
cook_d = cooks.distance(fit)) %>% # столбец со значениями Cook's D
filter(cook_d > cutoff) %>% # отбор значений, превышающих cutoff
arrange(desc(cook_d)) # сортировка по значению Cook's D в порядке убывания

nrow(df_cd) # число значений, превышающих уровень cutoff
## [1] 24
pander::pander(slice(df_cd, 1:6)) # первые шесть записей в таблице данных df_cd
```

Birthweight	Birthlength	n_dfs	cook_d
3265	58	297	0.08143
3370	58	112	0.06322
3380	58	255	0.06161
3310	57	72	0.03578
3100	56	180	0.02885
3860	50	132	0.02308

```
(influence_rown <- as.numeric(df_cd$n_dfs)) # номера рядов "влияющих" значений в таблице dfs
## [1] 297 112 255 72 180 132 225 404 178 184 121 106
165 138 114 382 84
## [18] 43 367 142 224 75 209 214
# Cook's D plot
# identify D values > 4/(n-k-1)
plot(fit, which=4, cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")
```

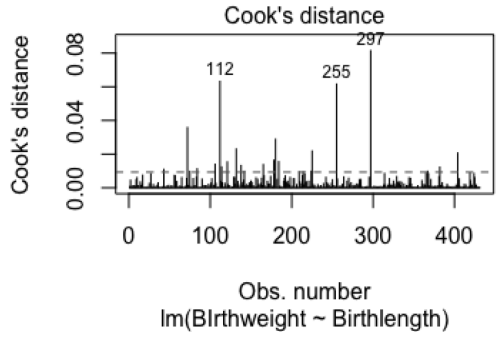


Рис. 12. Изучение «влияющих» значений

Таким образом, в результате анализа количество «влияющих» значений оказалось равным 24, номера записей в основной таблице dfs: 297, 112, 255, 72, 180, 132, 225, 404, 178, 184, 121, 106, 165, 138, 114, 382, 84, 43, 367, 142, 224, 75, 209, 214. Индивидуальное рассмотрение каждого из «влияющих» значений может понадобиться при дальнейших манипуляциях с данными.

Оценка предсказательной точности модели (рисунок 13 – листинг 13) предполагает следующие действия:

– создаются два набора данных (тренировочный и тестовый);

– 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.

Листинг 13

```
# создание тренировочного и тестового набора данных
set.seed(123)
row_index <- sample(1:nrow(dfs), .8 * nrow(dfs)) # определение номеров рядов, составляющих 80% записей
train_data <- dfs[row_index, ] # создание тренировочного набора
test_data <- dfs[-row_index, ] # создание тестового набора

# создание модели на данных тренировочного набора
model.lm <= lm(BIrthweight ~ Birthlength, train_data)
# получение предсказанных значений для тестового набора
bw_predict <- predict(model.lm, test_data)

# таблица данных с предсказанными и реальными значениями
actual_preds <- data.frame(cbind(actual = test_
data$BIrthweight,
predBW = round(bw_predict)))
# коэффициент корреляции между предсказанными и реальными значениями
(corp <- cor(actual_preds$actual, actual_preds$predBW))

## [1] 0.7576216

# средняя абсолютная процентная ошибка
(MAPE <- mean(abs(actual_preds$actual – actual_
preds$predBW) / actual_preds$actual))

## [1] 0.07151261
```

Рис. 13. Оценка предсказательной точности модели

В результате анализа коэффициент корреляции между предсказанными и реальными значениями составил 0,758, средняя абсолютная процентная ошибка – 0,072. Полученные данные позволяют сказать, что доля ошибок в данной модели составляет 7,2 %.

Дальнейшая работа с данными может улучшить качество модели (в приведенном ниже примере – удаление «влияющих» значений). Новая модель будет создана после удаления «влияющих» значений из изучаемой таблицы данных. Последующая оценка будет проводиться на измененных данных и сравниваться с показателями первой модели (рис. 14 – листинг 14).

Листинг 14

```
# удаление влиящих значений из таблицы данных
dfs_1 <- dfs[-influence_rown, ]
# построение модели на измененных данных
fit1 <- lm(BIrthweight ~ Birthlength, dfs_1)
# оценка модели
summary(fit1)
##
## Call:
## lm(formula = BIrthweight ~ Birthlength, data = dfs_1)
##
## Residuals:
## Min 1Q Median 3Q Max
## -611.27 -179.16 -19.16 175.55 714.49
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4586.635 347.186 -13.21
<0.0000000000000002 ***
## Birthlength 152.116 6.591 23.08 <0.0000000000000002 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 257.8 on 405 degrees of freedom
## Multiple R-squared: 0.5681, Adjusted R-squared: 0.567
## F-statistic: 532.6 on 1 and 405 DF, p-value: <
0.00000000000000022
pander::pander(round(cbind(coef(fit1), confint(fit1))), caption
= 'Коэффициенты регрессии и доверительные интервалы')
```

Коэффициенты регрессии и доверительные интервалы

		2.5 %	97.5 %
(Intercept)	-4587	-5269	-3904
Birthlength	152	139	165

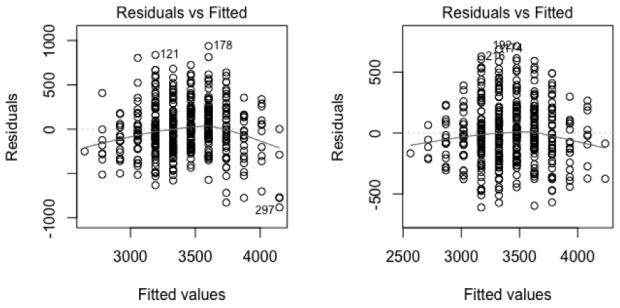
Рис. 14. Создание модели без «влияющих» значений

Далее проведем сравнение моделей, повторную оценку гомоскедастичности и прогностических возможностей модели. Сравним модели между собой, используя функцию Anova из пакета car, сравним коэффициенты AIC и BIC, показатели MSE и RMSE, повторно оценим гомоскедастичность остатков. Оценим прогностическую ценность новой модели (рис. 15 – листинг 15).

Листинг 15

```
Anova(fit, fit1, test = 'F')

## Anova Table (Type II tests)
##
## Response: BIrthweight
## Sum Sq Df F value Pr(>F)
## Birthlength 33940248 1 510.66 < 0.00000000000000022
***
## Residuals 26917957 405
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC(fit)
## [1] 6138.054

AIC(fit1)
## [1] 5678.509
```

```
BIC(fit)

## [1] 6150.252

BIC(fit1)

## [1] 5690.535

(MSE <- mean(residuals(fit)**2))

## [1] 88399.93

(MSE1 <- mean(residuals(fit1)**2))

## [1] 66137.49

(RMSE <- sqrt(mean(residuals(fit)**2)))

## [1] 297.3212

(RMSE1 <- sqrt(mean(residuals(fit1)**2)))

## [1] 257.1721

par(mfrow = c(1,2))
plot(fit, which = 1); plot(fit1, which = 1)
```




```
par(mfrow = c(1,1))

# Breusch-Pagan test
(ncvt <- ncvTest(fit1))

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.02454897 Df = 1 p = 0.8754962

# оценка прогностической ценности новой модели
set.seed(123)
row_index <- sample(1:nrow(dfs_1), .8 * nrow(dfs_1))
train_data <- dfs_1[row_index, ]
test_data <- dfs_1[-row_index, ]

model.lm <- lm(BIrthweight ~ Birthlength, train_data)
bw_predict <- predict(model.lm, test_data)

actual_preds <- data.frame(cbind(actual = test_
data$BIrthweight,
predBW = round(bw_predict)))
# коэффициент корреляции между предсказанными и
реальными значениями
(corp1 <- cor(actual_preds$actual, actual_preds$predBW))

## [1] 0.7989563

# средняя абсолютная процентная ошибка
(MAPE1 <- mean(abs(actual_preds$actual – actual_
preds$predBW) / actual_preds$actual))

## [1] 0.05490008
```

Рис. 15. Сравнение модели с «влияющими» значениями с моделью без «влияющих» значений

Результаты выполнения функции Anova указывают на значимые различия между моделями. Сравнение показателей AIC, BIC, MSE, RMSE указывает на повышение качества модели (изучаемые показатели уменьшились):

AIC: было – 6138, стало – 5679;

BIC: было – 6150, стало – 5691;

MSE: было – 88400, стало – 66137;

RMSE: было – 297, стало – 257.

В измененной модели не может быть отклонена нулевая гипотеза о гомоскедастичности (p-value = 0,875 в тесте Breusch-Pagan). В измененной модели коэффициент корреляции между предсказанными и реальными значениями составил 0,799, средняя абсолютная процентная ошибка – 0,055, в первоначальной модели коэффициент корреляции между предсказанными и реальными значениями был равен 0,758, средняя абсолютная процентная ошибка – 0,072.

Полученные данные позволяют сказать, что доля ошибок при использовании измененной модели снизилась до 5,5 %.

Таким образом измененная линейная регрессионная модель имеет следующую формулу: Масса тела новорожденного (г) = −4 586,6 + 152,1 × длина тела новорожденного (см).

Соответственно новорожденный с длиной тела 52 см в среднем имеет массу тела 3 323 г, нижняя и верхняя границы 95 % доверительного интервала будут равны 1967 г и 4680 г соответственно.

В следующей работе мы рассмотрим анализ качественных данных в программной среде R.

Список литературы

1. *Гржибовский А. М.* Корреляционный анализ // Экология человека. 2008. № 9. С. 50–60.

2. *Гржибовский А. М.* Однофакторный линейный регрессионный анализ // Экология человека. 2008. № 10. С. 55–64.

3. *Когаловский М. Р. и др.* Глоссарий по информационному обществу / под ред. Ю. Е. Хохлова. М.: Институт развития информационного общества, 2009. 162 с.

4. *Усынина А. А., Одланд И. О., Пылаева Ж. А., Пастбина И. М., Гржибовский А. М.* Регистр родов Архангельской области как важный информационный ресурс для науки и практического здравоохранения // Экология человека. 2017. № 2. С. 58–64.

5. *Aldrich J.* Correlation Genuine and Spurious in Pearson and Yule // Statistical Science. 1995. Vol. 10 (4). P. 364-376.

6. *Bowers D.* Medical Statistics from Scratch. Chichester, England: John Wiley & Sons Ltd, 2008.

7. *Crawley M. J.* The R Book. 2nd ed. Wiley, 2013.

8. *Faraway J. J.* Linear Models with R. NewYork: Chapman & Hall/CRC, 2005.

9. *Hut I.* 2017. Correlation Tests, Correlation Matrix, and Corresponding Visualization Methods in R. URL: https://rstudio-pubs-static.s3.amazonaws.com/240657_5157ff98e8204c358b2118fa69162e18.html (дата обращения 10.09.2018).

10. *Kabacoff R. I.* R in Action. Data analysis and graphics with R: 2nd ed. ShelterIsland, NY: Manning Publications, 2015.

11. *Logan M.* Biostatistical Design and Analysis Using R: A Practical Guide. Wiley-Blackwell, 2010.

12. *Prabhakaran S.* 2017. How to Detect Heteroscedasticity and Rectify It? URL: https://datascienceplus.com/how-to-detect-heteroscedasticity-and-rectify-it/ (дата обращения 10.09.2018).

13. STAT501, PennState. 2018. Regression Analysis. URL: https://newonlinecourses.science.psu.edu/stat501/ (дата обращения 10.09.2018).

14. STHDA. 2016. Correlation Analyses in R. URL: http://www.sthda.com/english/wiki/correlation-analyses-in-r. (дата обращения 10.09.2018).

15. *Tufte E. R.* The Cognitive Style of Powerpoint: Pitching Out Corrupts Within. Cheshire, Connecticut: Graphics Press, 2006.

References

1. Grjibovski A. M. Correlation analysis. *Ekologiya cheloveka* [Human Ecology]. 2008, 9, pp. 50-60. [In Russian]

2. Grjibovski A. M. Simple linear regression analysis. *Ekologiya cheloveka* [Human Ecology]. 2008, 10, pp. 55-64. [In Russian]

3. Kogalovskiy M. R. i dr. *Glossariy po informatsionnomu obshchestvu* [Glossary of the Information Society]. Moscow, Institute of the Information Society, 2009, 162 p.

4. Usynina A. A., Odland Jon Øyvind, Pylaeva Zh. A., Pastbina I. M., Grjibovski A. M. Arkhangelsk County Birth Registry as an Inportant Source of Information for Research and Healthcare. *Ekologiya cheloveka* [Human Ecology]. 2017, 2, pp. 58-64. [In Russian]

5. Aldrich J. Correlation Genuine and Spurious in Pearson and Yule. *Statistical Science*. 1995, 10 (4), pp. 364-376.

6. Bowers D. *Medical Statistics from Scratch*. Chichester, England, John Wiley & Sons Ltd, 2008.

7. Crawley M. J. *The R Book*. 2nd ed. Wiley, 2013.

8. Faraway J. J. *Linear Models with R*. N Y, Chapman & Hall/CRC, 2005.

9. Hut I. 2017. *Correlation Tests, Correlation Matrix, and Corresponding Visualization Methods in R*. Available from: https://rstudio-pubs-static.s3.amazonaws.com/240657_5157ff98e8204c358b2118fa69162e18.html (accessed: 10.09.2018).

10. Kabacoff R. I. R in Action. *Data analysis and graphics with R*. 2nd ed. ShelterIsland, N Y, Manning Publications, 2015.

11. Logan M. *Biostatistical Design and Analysis Using R: A Practical Guide*. Wiley-Blackwell, 2010.

12. Prabhakaran S. 2017. *How to Detect Heteroscedasticity and Rectify It?* Available from: https://datascienceplus.com/how-to-detect-heteroscedasticity-and-rectify-it/ (accessed: 10.09.2018).

13. STAT501, PennState. 2018. *Regression Analysis*. Available from: https://newonlinecourses.science.psu.edu/stat501/ (accessed: 10.09.2018).

14. STHDA. 2016. *Correlation Analyses in R*. Available from: http://www.sthda.com/english/wiki/correlation-analyses-in-r (accessed: 10.09.2018).

15. Tufte E. R. *The Cognitive Style of Powerpoint: Pitching Out Corrupts Within*. Cheshire, Connecticut, Graphics Press, 2006.

Контактная информация:

Г*ржибовский Андрей Мечиславович* – доктор медицины, заведующий ЦНИЛ Северного государственного медицинского университета, г. Архангельск; профессор Северо-Восточного федерального университета, г. Якутск; визитинг-профессор Казахского национального университета им. Аль-Фараби, г. Алматы, Казахстан и Западно-Казахстанского государственного медицинского университета им. Марата Оспанова, г. Актобе, Казахстан; почетный доктор Международного казахско-турецкого университета, г. Туркестан (Казахстан); почетный профессор Государственного медицинского университета г. Семей (Казахстан)

Адрес: 163000 г. Архангельск, Троицкий пр., д. 51

E-mail: Andrej.Grjibovski@gmail.com