ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ «САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ УНИВЕРСИТЕТ имени академика С.П. КОРОЛЕВА»

КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ В СРЕДЕ СТАТИСТИЧЕСКОГО ПАКЕТА R

САМАРА 2010

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ имени академика С.П. КОРОЛЕВА»

КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ В СРЕДЕ СТАТИСТИЧЕСКОГО ПАКЕТА R

Утверждено Редакционно-издательским советом университета
в качестве методических указаний к практическим занятиям

САМАРА
Издательство СГАУ
2010

УДК 612.85(075)

Составитель М.В. Комарова

Рецензент доц. В.Н. Конюхов

Корреляционный и регрессионный анализ медико-биологических данных в среде статистического пакета R: метод. указания к практ. занятиям / сост. *М.В. Комарова*. — Самара: Изд-во Самар. гос. аэрокосм. ун-та, 2010. — 24 с.

В методических указаниях представлены краткие сведения о корреляционном анализе, линейной и логистической регрессионных моделях. Описаны особенности их выполнения в среде статистического пакета R. Приведены примеры реальных биомедицинских данных и требования к отчёту.

Предназначены для студентов, обучающихся по специальности 200401 «Биотехнические и медицинские аппараты и системы» при изучении курса «Системный анализ и принятие решений».

Подготовлены на кафедре радиотехники и медицинских диагностических систем.

Цель работы: ознакомиться с проведением корреляционного и регрессионного анализов в среде статистического пакета R; построить модели линейной и логистической регрессии по массивам реальных биомедицинских данных, проанализировать полученные модели.

РАБОТА 1. ИССЛЕДОВАНИЕ ТЕСНОТЫ ВЗАИМОСВЯЗЕЙ БИОМЕДИЦИСКИХ ДАННЫХ В СРЕДЕ R

Краткие теоретические сведения

Корреляция моментов Пирсона. Корреляция отражает тесноту и направленность взаимосвязи двух переменных. Если форма распределения анализируемых признаков не очень сильно отличается от нормальной и отсутствуют выбросы, рассчитывают коэффициент корреляции Пирсона (часто называемый просто коэффициентом корреляции):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (1.1)$$

где x_i *и* y_i — анализируемые показатели, n — число наблюдений.

Коэффициент корреляции изменяется в диапазоне от –1 до +1. Знак его отражает направленность взаимосвязи (+ прямая, – обратная), абсолютное значение — тесноту. В прикладных исследованиях часто используют следующие условные градации:

$\|r\|<0,3$	связь слабая
$0,3\le\|r\|<0,6$	связь умеренная
$0,6\le\|r\|<0,8$	связь сильная, тесная
$0,8\le\|r\|<0,1$	связь очень сильная

Следует помнить, что корреляция между величинами x и y не обязательно отражает причинно-следственные связи между ними.

Ранговые коэффициенты корреляции. Непараметрические, или ранговые коэффициенты корреляции Спирмена или Кендалла, рассчитывают в следующих случаях:

– форма распределения отличается от нормальной, например, скошена в ту или иную сторону;

– есть значительные выбросы (которые отражают не ошибки измерений или регистрации данных, а их реальные биологические особенности);

– шкала измерений не количественная, а порядковая;

– небольшой размер выборки.

При вычислении коэффициента корреляции *Спирмена* (r_s) величины x сортируют по возрастанию и ранжируют. Равные величины получают средние значения из рангов, которые получили бы эти значения без ограничения. Аналогично присваивают ранги величинам y. Находят разности рангов x_i и y_i, обозначаемые d_i.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}. \qquad (1.2)$$

Коэффициент корреляции *Кендалла* (τ) рассчитывают по следующему алгоритму:

1. Значения x ранжируют по возрастанию.
2. Значения y располагают соответственно x и ранжируют.
3. Для каждого ранга y_i определяют число следующих за ним значений рангов, больших y_i. Суммируют и получают n_c — число согласованных пар (от англ. *concordant*), или последовательностей.
4. Для каждого значения y_i определяют число следующих за ним рангов, меньших y_i. Суммируют и получают n_d — число рассогласованных пар (от англ. *discordant*), или инверсий.
5. Рассчитывают τ по формуле:

$$\tau = \frac{2(n_c - n_d)}{n(n-1)}. \qquad (1.3)$$

Примечание: при связанных рангах формула несколько усложняется.

Статистическая значимость коэффициента корреляции. Уровень статистической значимости коэффициента корреляции (Пирсона, Спирмена или Кендалла) зависит от числа наблюдений и может быть оценен с помощью следующей статистики:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad (1.4)$$

где r — коэффициент корреляции, n — число наблюдений. Рассчитанная статистика t имеет распределение Стьюдента с (n–2) степенями свободы. При этом проверяются следующие статистические гипотезы:

H_0: Корреляция между переменными не отличается от нуля.

H_1: Корреляция между переменными достоверно отличается от нуля.

При уровне статистической значимости $p<0,05$ мы отвергаем нулевую гипотезу и считаем, что связь между изучаемыми переменными действительно существует. Следует различать понятия статистической значимости и тесноты взаимосвязей, т.е. связи могут быть слабыми и в то же время значимыми (при большом числе наблюдений) и, наоборот, умеренными или даже сильными, но не значимыми, если наблюдений недостаточно.

Функции R для проведения корреляционного анализа

Функция `corr` ()

Для расчета коэффициентов корреляции применяют функцию:

`corr(x, use=, method= )`.

Функция возвращает матрицу корреляций между указанными переменными.

Аргументы

`x` — набор данных, между которыми вычисляются корреляции (можно указать имена переменных рабочего файла в кавычках);

`use` — обработка пропусков. Возможные значения:

`all.obs` — без пропусков (пропущенные данные вызовут ошибку);

`complete.obs` — построчная обработка пропусков;

`pairwise.complete.obs` — попарная обработка пропусков;

`method` — вид корреляции. Возможные значения:

`pearson` — корреляция Пирсона;

`spearman` — корреляция Спирмена;

`kendall` — корреляция Кендалла.

Функция `corr.test` ()

Для оценки уровня значимости коэффициентов корреляции применяют функцию:

`cor.test (x, y, alternative=, method= ).`

Аргументы

`x, y` — набор данных, должны быть одинаковой длины;

`alternative` — альтернативная гипотеза (двусторонний или односторонний тест). Возможные значения:

`two.sided`;

`greater`;

`less`;

`method` — вид тестируемой корреляции. Возможные значения:

`pearson` — корреляция Пирсона;

`spearman` — корреляция Спирмена;

`kendall` — корреляция Кендалла.

ФУНКЦИИ R ДЛЯ ГРАФИЧЕСКОЙ ИНТЕРПРЕТАЦИИ КОРРЕЛЯЦИОННОГО АНАЛИЗА

Для визуального исследования зависимости между двумя переменными используют двумерные диаграммы рассеяния, или графики разброса.

Функция `scatterplot` ()

Для создания графика разброса между двумя переменными применяют функцию:

`scatterplot(formula, data, xlab, ylab, legend.title, ellipse, reg.line, smooth).`

Аргументы

`formula` — «формула» для построения графика, применяют в форме y~x или y~x | z, где z — фактор, подразделяющий выборку на подгруппы;

`data` — массив данных, по которому строится график разброса;

`xlab` — название горизонтальной оси;

`ylab` — название вертикальной оси;

`legend.title` — заголовок легенды;

`ellipse` — при значении TRUE вместо точек на графике отображаются корреляционные эллипсы;

reg.line — отображает линию линейной регрессии при значении TRUE и не отображает её при значении FALSE;

smooth — отображает кривую нелинейной регрессии при значении TRUE и не отображает её при значении FALSE.

При необходимости построить матрицу парных графиков по нескольким переменным можно воспользоваться функцией: `scatterplot.matrix()`.

Аргументы данной функции во многом аналогичны таковым функции `scatterplot()`, иначе пишется «формула» графика: ~x1+x2+x3… Помимо указанных функций, код которых генерирует оболочка R commander, существует функция `pairs()`, которая также создаёт графики разброса.

Пример матрицы парных графиков гематологических показателей приведён на рис. 1. По диагонали представлены имена анализируемых переменных (гематологических показателей в нашем примере). Каждая точка отражает одно наблюдение, её координаты определяются значениями двух переменных. Выше и ниже диагонали с именами переменных расположены одни и те же пары переменных, но по разным осям. Например, график в первой строке и втором столбце отражает зависимость HCT1 (гематокрит, ось ординат) от HGB1 (гемоглобин, ось абсцисс), а график во второй строке, первом столбце — наоборот зависимость HGB1 от HCT1.

Если переменные тесно и линейно связаны, то множество точек данных принимает форму узкого эллипса или почти прямой. В рассматриваемом примере между HGB1 и HCT1 — очень тесная связь; между HGB1 и MCH1 — умеренная по тесноте связь. Если переменные не связаны, то точки образуют облако рассеяния (MCH1 и MCHC1).

Диаграммы рассеяния предоставляют исследователю больше информации, чем простое значение коэффициента корреляции. Они позволяют:

– выявить отсутствие однородности в выборке (например, наличие подгрупп с разным характером взаимосвязи);

– найти выбросы, или нетипичные данные, которые искусственным образом могут значительно увеличить или уменьшить коэффициент корреляции Пирсона;

– обнаружить нелинейный характер взаимосвязи.

Таким образом, перед проведением корреляционного анализа желательно анализировать графики разброса, с помощью которых можно подобрать оптимальный срез данных для исследования (т.е. выделить определённые подгруппы или, наоборот, объединить разные подгруппы в одну, исключить выскакивающие наблюдения) и применить подходящий вид корреляции (Пирсона или его непараметрических аналогов — Спирмена или Кендалла).

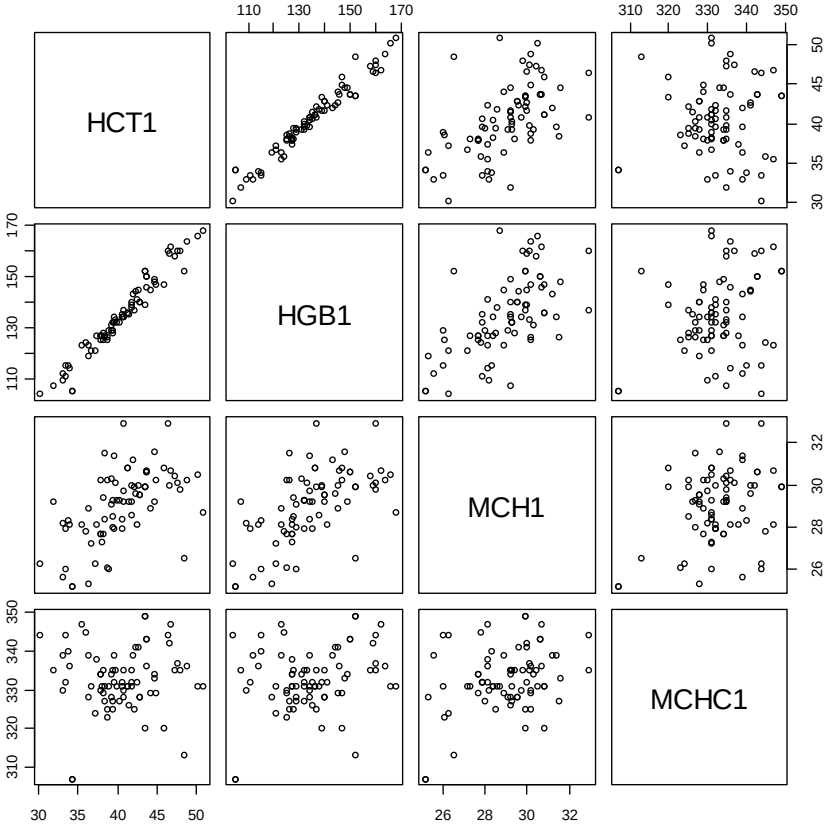

Рис. 1 — Диаграммы рассеяния биологических показателей в зависимости от тесноты взаимосвязи

Порядок работы

1. Запустите пакет R с помощью файла: R-2.7.2/bin/Rgui.exe. В окне R Console напишите: library (Rcmdr). Дальнейшая работа проходит в данной графической оболочке, которая генерирует необходимый код

через кнопочный интерфейс. Для загрузки данных в графической оболочке R commander выбрать следующие пункты меню:

`Данные`
`Импорт данных`
`Из файла SPSS`

Появится новое диалоговое окно. Со всем согласиться, нажать ОК.

Найти файл `pregn.sav`.

2. Исследуйте взаимосвязи клинико-лабораторных данных новорождённых и их матерей. Для этого воспользуйтесь корреляционным анализом:

`Статистики`
`Итоги`
`Корреляционная матрица`

В появившемся диалоговом окне укажите исследуемые переменные (*выбрать один вариант*):

Вариант 1

p_ch1 — масса ребёнка
len_ch1 — длина ребёнка
week — гестационный срок (в неделях)
tromb — тромбоциты матери
svert — свёртываемость крови матери

Вариант 2

h1 — рост матери
p_0 — масса матери
age — возраст матери
imt — индекс массы тела матери
p_len — соотношение массы и длины ребёнка

Вариант 3

Hb2 — гемоглобин матери
er2 — эритроциты матери
tp1 — общий белок матери
p_0 — масса матери 1
p_2 — масса матери 2

Выберите вначале корреляцию Пирсона, потом Спирмена. Проанализируйте полученную матрицу корреляций. Какие связи сильные, какие слабые, а какие умеренные? Для нескольких пар показателей найдите уровень значимости коэффициента корреляции.

`Статистики`
`Итоги`
`Корреляционный тест`

3. Проиллюстрируйте полученные результаты (для этих же переменных) на графиках разброса:

`Графики`
`Матрица точечных графиков`

Выбрать необходимые переменные.

Убрать галочки с «*линии наименьших квадратов*» *и* «*сгладить линии*».

В отчёт: по заданию 1 и 2 выберите 3 графика разброса. Укажите численное значение коэффициентов корреляции и их уровней значимости. Дайте содержательную оценку взаимосвязей: прямая или обратная, слабая или сильная. Какие значения коэффициентов корреляции больше по абсолютному значению: Пирсона или Спирмена?

4. Проведите исследование взаимосвязей в различных подгруппах обследованных. Постройте ещё раз графики разброса, нажав на кнопку «графики по группам» и введя качественный признак gestoz. Теперь на графиках разным цветом и разными маркёрами выделены наблюдения (пациенты) из различных групп по степени тяжести осложнения беременности — гестозу. Что можно увидеть, глядя на полученные графики? Различаются ли взаимосвязи в разных подгруппах?

РАБОТА 2. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Краткие теоретические сведения

Регрессионный анализ решает следующие задачи:

– восстановление зависимости между исследуемыми переменными;
– прогноз зависимой переменной (переменной отклика) по известным независимым переменным (предикторам).

Регрессия бывает:

– по числу предикторов — парная и множественная;
– по форме зависимости — линейная и криволинейная.

Исходные данные для регрессионного анализа представляют собой таблицу (матрицу), в которой строки соответствуют объектам (испытуемым), а столбцы — переменным. Все переменные при этом должны быть измерены в количественной шкале. Одна из переменных определяется исследователем как зависимая, а остальные как независимые переменные.

Парная линейная регрессия в общем случае имеет вид: $y=b_0+b_1x$. Нахождение коэффициентов регрессии основано на методе наименьших квадратов (минимизация суммы квадратов отклонений эмпирических значений признака от теоретических, полученных по уравнению регрессии).

ФУНКЦИИ R ДЛЯ ПОСТРОЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Для построения линейной регрессии в пакете R можно воспользоваться функцией `lm()`:

```
lm(formula, data, subset, weights, na.action).
```

Аргументы

`formula` — символическое описание восстанавливаемой модели. Для парной линейной регрессии имеет вид $y\sim x$, для множественной — $y\sim x_1+x_2+x_3\ldots$;

`data` — источник данных;

`subset` — подмножество данных, участвующих в построении модели, необязательный параметр;

`weights` — вектор весов, может быть или NULL, или числовым;

`na.action` — обработка пропущенных данных (NA).

Трактовка результатов

Объект, возвращаемый функцией `lm`, имеет различные поля:

`Residuals` — остатки ($y_i–bx_i$), распределение по квартилям.

`Coefficients` — коэффициенты регрессии и их статистическая значимость:

- `Estimate` — коэффициент регрессии b;
- `Std. Error` — ошибка коэффициента регрессии b;
- `t value` — статистика t для оценки уровня значимости коэффициента регрессии;
- `Pr(>|t|)` — достигнутый уровень значимости коэффициента регрессии.

`Multiple R-squared` — коэффициент детерминации модели.

`Adjusted R-squared` — скорректированный коэффициент детерминации модели.

`F-statistic` — F-статистика для модели в целом.

`p-value` — уровень значимости модели в целом.

Порядок работы

1. Постройте уравнение зависимости веса и длины новорождённых от гестационного срока родов (в неделях, переменная week). Для этого выберете следующие пункты меню в оболочке R commander:

```
Статистики
Подгонка моделей
Линейная регрессия
```

В появившемся диалоговом окне укажите: зависимая переменная — p_ch1 (или len_ch1), независимая — week. Проанализируйте график остатков (наблюдаемое минус предсказанное регрессионной моделью значение). Одинаковы ли они на всём диапазоне значений предсказывающей переменной.

```
Модели
Графики
График Компонента + остаток
```

Отразить в отчёте. По таблице коэффициентов запишите полученное уравнение регрессии. Значимы ли коэффициенты регрессии? В каком диапазоне значений данное уравнение будет хорошо работать?

2. Для двух произвольных выборок (n_1=n_2=10) восстановите парную линейную регрессию в среде Excel. Рассмотрите два варианта зависимости: y(x) и x(y).

Коэффициенты регрессии можно найти по формулам:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (2.1)$$

$$b_0 = \bar{y} - b_1\bar{x}. \qquad (2.2)$$

Рассчитайте также коэффициент корреляции Пирсона между выборками (по формуле 1.1) или его модуль по формуле:

$$y = \pm\sqrt{b_{1x(y)}b_{1y(x)}}\ . \qquad (2.3)$$

Отразите графически обе зависимости y(x) и x(y). Примерный вид графиков приведён на рис. 2.

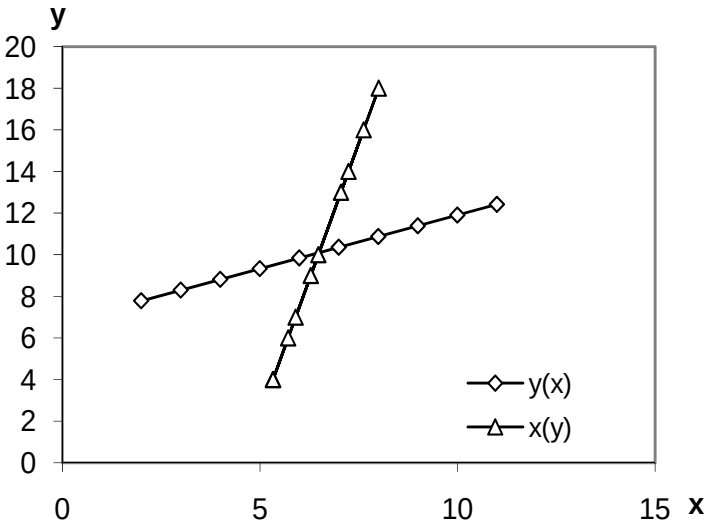

Рис. 2 — Парная линейная регрессия

После построения двух графиков изменять исходные данные так, чтобы получались различные коэффициенты корреляции. Как меняется расположение графиков и угол между прямыми? В какой точке пересекаются прямые?

РАБОТА 3. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ: ПОШАГОВОЕ ПОСТРОЕНИЕ

Краткие теоретические сведения

Множественная линейная регрессия позволяет изучить совместное воздействие нескольких независимых переменных на переменную отклика. Практическое применение двоякое: для предсказания переменной отклика и для определения интенсивности, с которой каждая независимая переменная линейно связана с зависимой. В общем случае уравнение множественной линейной регрессии имеет вид:

$$y = b_0 + b_1x_1 + b_2x_2 \ldots + b_kx_k \,. \qquad (3.1)$$

Интерпретация коэффициентов регрессии та же, что и в парной регрессии — на сколько изменится переменная отклика при увеличении предиктора на единицу.

Выполнение работы

Цель работы: изучение пошаговых методов построения множественной линейной регрессии; диагностика полученной модели; работа с частью массива данных.

Порядок работы

1. Запустите R; выполните команду: library (Rcmdr). Загрузите файл food.xls. Обратите внимание, что файл имеет расширение xls, следовательно, нужен экспорт из Excel. Выбрать первый лист.

В данном массиве приведены некоторые антропометрические данные здоровых мужчин и женщин, их возраст, а также рассчитанные на основе анкетирования съедаемые калории, жиры, углеводы, микроэлементы. Задача исследования: выявить компоненты пищи, от которых в первую очередь зависит индекс массы тела.

Для построения регрессионной модели отдельно для мужчин и женщин необходимо выбрать подмножество данных. В R консоли или окне скриптов напишите и выполните команду:

```
data_m <- subset(data1, Пол==1).
```

Обозначения

– data_m — имя нового набора данных, оно может быть произвольным;
– subset — команда R, выбирающая часть массива данных согласно определённому условию;
– data — имя исходного набора данных (если при открытии файла было присвоено по умолчанию имя Данные, то его и надо писать);
– Пол==1 — условие, по которому осуществляется выбор. В анализируемом массиве мужчины обозначены кодом 1, женщины 2. Если бы мы работали с качественным признаком, имеющим не числовой, а строковый формат, то значение нужно было бы взять в кавычки.

Теперь необходимо сделать полученный массив данных активным. Для этого выполните пункты меню R commander:

```
Данные
Активные данные
Выбрать активный набор данных
```

В появившемся диалоговом окне выберите необходимый набор данных.

Просмотрите массив данных (*из режима просмотра удобно скопировать имена переменных для построения дальнейшей модели*).

2. Для построения пошаговой множественной линейной регрессии выполните следующие команды:

Команда	**Описание**
`library(MASS)`	Вызов модуля MASS для пошагового построения регрессии
`fit1 <- lm(y~x1+x2+x3, data=mydata)`	Построение модели множественной линейной регрессии методом включения всех предикторов в общем виде:
`fit1`	имя создаваемой модели, к которому можно будет в дальнейшем обратиться для вывода различных характеристик построенной модели;
`y`	зависимая переменная (в нашем задании — ИМТ, или окружность талии или бёдер);
`x1+x2+x3`	переменные отклика (включить возраст и всё-всё, связанное с пищей: белки, различные жиры, углеводы, микроэлементы);
`mydata`	имя массива данных, на котором строится модель (например, data_m или data_f для мужчин и женщин соответственно)
`summary(fit1)`	описание полученной модели:
`fit1`	имя модели, заданное при её построении
`fit1_step <- stepAIC(fit1, direction="backward")`	построение новой модели с пошаговым исключением предикторов:
`fit1_step`	имя новой, создаваемой пошаговым методом модели;
`fit1`	имя первоначальной модели, в которой присутствовали все предикторы
`summary(fit1_step)`	описание полученной модели (с меньшим числом предикторов)
`vif(fit1_step)`	оценка мультиколлинеарности предикторов по показателю vif

VIF — аббревиатура от ***v**ariance **i**nflation **f**actor* — величина отражающая мультиколлинеарность, т.е. тесную корреляцию предикторов друг с другом.

$$VIF = \frac{1}{1 - R^2_{1.all\_pred}}, \qquad (3.2)$$

где $R_{1.all\_pred}$ — коэффициент множественной корреляции данного предиктора со всеми остальным предикторами. Значение VIF более 2 характеризует модель как мультиколлинеарную, крайне неустойчивую к небольшим изменениям предикторов и непригодную для прогноза.

Пример вывода результатов множественной линейной регрессии в пакете R

```
Coefficients:
             Estimate     Std. Err     t value Pr(>|t|)
(Intercept)  21.196479    1.229588     17.239  <2e-16 ***
Крахмал      0.003406     0.004069     0.837   0.40315
Общий.жир    0.029431     0.008472     3.474   0.00058***
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.636 on 337 degrees of freedom
Multiple R-squared: 0.036, Adjusted R-squared: 0.03028
F-statistic: 6.293 on 2 and 337 DF, p-value: 0.002073
```

Обозначения

- Первый столбец (без заголовка) — свободный член (`intercept`) и предикторы в модели;
- `Estimate` — коэффициент регрессии;
- `Std. Error` — стандартная ошибка коэффициента регрессии;
- `t value` — t-статистика, с помощью которой проверяют статистическую гипотезу отличия коэффициента регрессии от 0;
- `Pr(>|t|)` — уровень значимости предиктора, статистически значимые предикторы помечены звёздочками;
- `Residual standard error` — остаточная стандартная ошибка — показатель разброса возможных значений случайной ошибки;
- `Multiple R-squared` — коэффициент детерминации, или квадрат коэффициента множественной корреляции модели;
- `Adjusted R-squared` — скорректированный коэффициент детерминации;
- `F-statistic` — F-статистика, оценивающая значимость полученной модели множественной линейной регрессии в целом, и её уровень значимости.

Сравните коэффициенты регрессии до пошагового отбора и после него. Одинаковы они или нет?

Если в модели оказались незначимые или тесно коррелированные предикторы (с показателем vif>2), модель перестроить, исключив их из построения. Может потребоваться несколько итераций, пока не получится приемлемый результат.

Постройте диагностические графики остатков (остаток — разность между наблюдаемым и рассчитанным значением переменной отклика) регрессии. Для этого необходимо выполнить следующие команды.

Команда	**Описание**
`layout(matrix(c(1,2,3,4),2,2))`	Подготовка для построения диагностических графиков
`plot(fit)`	Построение диагностических графиков регрессии (в скобках указать имя анализируемой модели)

Примеры получаемых графиков представлены на рис. 3. График `Residuals vs Fitted` (остатки против рассчитанных значений) в хорошей модели должен быть равномерным на всем диапазоне значений переменной отклика. Аналогично трактуют график `Scale-Location`. Более точно определить закон распределения остатков можно по графику `Normal Q-Q` (график на вероятностной бумаге). Если стандартизованные остатки (откладываемые по оси ординат) нормально распределены, то все значения легли бы на прямую линию на графике; если есть выбросы или остатки распределены не по нормальному закону, то линейность графика на вероятностной бумаге нарушается. Последний график `Residuals vs Leverage` позволяет эффективно выявить выбросы, наиболее значительно влияющие на расчет коэффициентов регрессии. На всех графиках указаны номера «выскакивающих» наблюдений.

Отразите в отчёте ответы на следующие вопросы

– Все ли предикторы в модели статистически значимы?
– Есть ли в модели предикторы, для которых VIF больше 2? Если да, то сравните коэффициенты корреляции между переменной отклика и данным предиктором и коэффициентом регрессии: одинаковы ли они по знаку?
– Чему равен коэффициент детерминации модели (посмотреть на значение `Adjusted R-squared`)?
– Значима ли модель в целом (посмотреть на значение `F-statistic` и его уровень значимости — `p-value`)?

– Равномерно ли распределены остатки (residuals) по рассчитанным значениям переменной отклика?

3. Постройте аналогичные модели для женщин. Одинаковы ли пищевые факторы риска ожирения у разных полов?

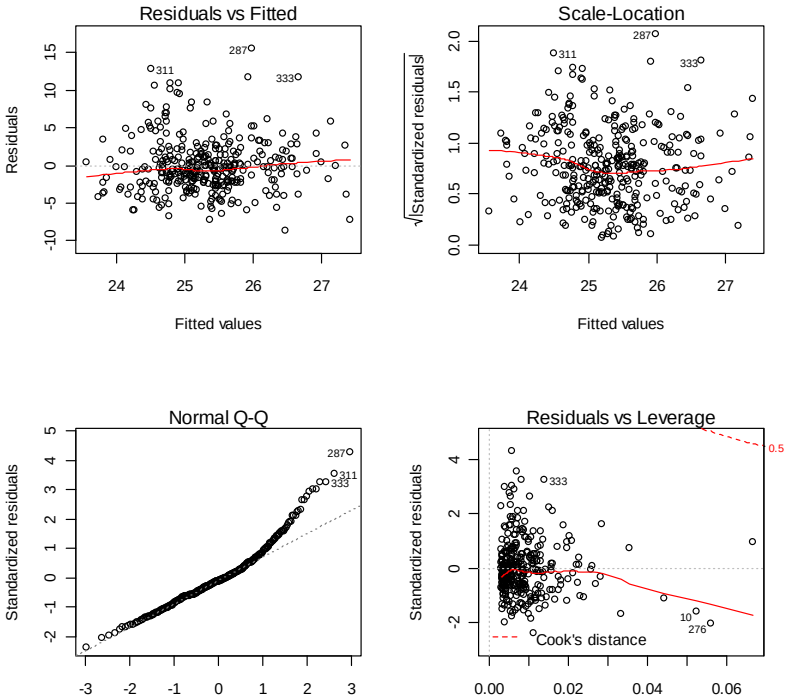

Рис. 3 — Диагностические графики в моделях множественной линейной регрессии

РАБОТА 4. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Краткие теоретические сведения

Логистическая регрессия применяется для предсказания показателя с бинарным откликом (то есть имеющим два значения, в медицинских задачах это часто наличие или отсутствие заболевания) и одной или несколькими объясняющими переменными.

Модель логистической регрессии имеет вид:

$$\ln \frac{p}{1-p} = b_0 + b_1 x_1 + \ldots + b_k x_k \,, \qquad (4.1)$$

где p — вероятность моделируемого события (исхода; заболевания); b_0, b_1… b_k — коэффициенты регрессии; x_1, x_2… x_k — объясняющие переменные.

Тогда шансы исхода (заболевания) равны:

$$шансы = \frac{p}{1-p} = \exp(b_0 + b_1 x_1 + ... + b_k x_k). \quad (4.2)$$

Вероятность развития исхода можно вычислить:

$$p = \frac{шансы}{1 + шансы}. \quad (4.3)$$

Интерпретация коэффициентов логистической регрессии отличается от интерпретации коэффициентов линейной регрессии. Практическое значение имеют экспоненциальные коэффициенты регрессии: они показывают, во сколько раз изменятся шансы предсказываемого состояния при повышении уровня предиктора на 1.

Выполнение работы

Цель работы: исследовать модель логистической регрессии для предсказания событий с двумя исходами.

Ход работы

1. Запустите R с помощью файла: R-2.7.2/bin/Rgui.exe. В окне R Console напишите: library (Rcmdr). Для загрузки данных в графической оболочке R commander выбрать следующие пункты меню:

```
Данные
Импорт данных
Из файла SPSS
```

Появится новое диалоговое окно. Со всем согласиться, нажать OK.

Найти файл `cancer.sav`.

Варианты заданий

Вариант 1

Зависимая переменная ds1 — диагноз (1 — не рак; 2 — рак)

Независимые предикторы:

- PSA — просатоспецифический антиген, лабораторный маркёр рака предстательной железы;
- CD16 — иммунологический показатель, отражающий долю лимфоцитов — так называемых натуральных киллеров, участвующих в уничтожении раковых клеток.

Вариант 2

Зависимая переменная ds1 — диагноз (1 — не рак; 2 — рак)

Независимые предикторы:

- PSA_free — свободная фракция просатоспецифического антигена, лабораторного маркёра рака предстательной железы;
- testost — тестостерон, мужской половой гормон.

Вариант 3

Зависимая переменная ds1 — диагноз (1 — не рак; 2 — рак)

Независимые предикторы:

- ro_psa — плотность просатоспецифического антигена;
- v — объём простаты (по данным УЗИ).

2. Построение модели логистической регрессии можно осуществить как с помощью оболочки R commander, так и непосредственным введением команд. При построении первым способом в окне R commander выбираем следующие пункты меню:

```
Статистики
Подгонка моделей
Обобщенная линейная модель
```

Имя модели: mylogit1 (можно и другое, можно и имеющееся по умолчанию оставить, но тогда внести коррективу в следующем задании). Первой выбираем бинарную переменную отклика; потом две независимых переменных. Семейство: `binomial`, функция связи: `logit` (они стоят по умолчанию).

То же самое можно получить, выполнив непосредственно команды в окне R console или в окне скриптов R commander:

Команда	**Описание**

`mylogit1 <-`
`glm(DS1 ~ PSA + V,`
`family=binomial(logit),`
`data=mydata)`
	Построение модели множественной линейной регрессии методом включения всех предикторов:

`mylogit1`	имя модели;
`DS1`	предсказываемое событие, должно иметь два возможных значения;
`PSA, V`	независимые предикторы;
`mydata`	имя массива данных для построения модели;
`summary(mylogit1)`	описание полученной модели;
`mylogit1`	имя модели, заданное при её построении

Пример вывода результатов логистической регрессии в пакете R

```
Coefficients:
             Estimate     Std.Error    z value Pr(>|z|)
(Intercept)  -1.273527    0.294750     -4.321  1.56e-05 ***
PSA          0.053281     0.010999     4.844   1.27e-06 ***
V            -0.006506    0.006673     -0.975  0.330
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Обозначения

- Первый столбец (без заголовка) — свободный член (`intercept`) и предикторы в модели;
- `Estimate` — коэффициент регрессии;
- `Std. Error` — стандартная ошибка коэффициента регрессии;
- `z value` — стандартизованная статистика Вальда; с её помощью проверяют статистическую гипотезу отличия коэффициента регрессии от 0;
- `Pr(>|z|)` — уровень значимости предиктора, статистически значимые предикторы помечены звёздочками.

Для нахождения экспоненциальных коэффициентов регрессии (отношение шансов) необходимо выполнить команды:

```
exp(mylogit1$coefficients)
exp(confint(mylogit1))
```

Примечание: если модель регрессии названа не `mylogit1`, а как-то иначе, то имя модели и нужно указывать!

Отразить в отчёте: цель работы, полученное уравнение регрессии; статистическую значимость предикторов; экспоненциальные коэффициенты и их доверительные интервалы. Включает ли единицу доверительный интервал, в случае если коэффициент значимый и незначимый?

3. В Excel`е изобразите графически вероятность моделируемого события от одного из значимых предикторов, а другой зафиксируйте на двух значениях (по оси абсцисс — значения предиктора, по оси ординат — вероятность события). Найти диапазон изменений предикторов можно через описательные статистики.

Совет. Возьмите 15–20 значений независимого предиктора с равным шагом от минимального до максимального (по исходным дан-

ным). Построение графика в Excel удобно разделить на два этапа: вначале рассчитать шансы, потом вероятности предсказываемого события. Затем воспользоваться функцией «точечный график». Должна получиться S-образная кривая. Полученный график зарисовать в отчет.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В каких случаях можно рассчитывать коэффициент корреляции Пирсона, а в каких Спирмена или Кендалла?
2. Для чего предназначены диаграммы рассеяния?
3. Какая функция пакета R предназначена для построения линейной регрессии? Какие характеристики модели она возвращает?
4. Какие средства диагностики мультиколлинеарности предлагает пакет R?
5. Для чего предназначены диагностические графики остатков регрессии в пакете R?
6. Напишите уравнение логистической регрессии.
7. Как отличается интерпретация коэффициентов в линейной и логистической регрессии?

РЕКОМЕНДУЕМЫЙ БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Вараксин, А.Н. Статистические модели регрессионного типа в экологии и медицине [Текст] / А.Н. Вараксин. — Екатеринбург, 2006. — 256 с.
2. Дюк, В.А. Информационные технологии в медико-биологических исследованиях [Текст] / В.А. Дюк, В.Л. Эммануэль. — СПб.: Питер, 2003. — 528 с.
3. Плис, А.И. Практикум по прикладной статистике в среде SPSS [Текст]: учеб. пособие для студентов вузов. Ч.1. Классические процедуры статистики / А.И. Плис, Н.А. Сливина. — М.: Финансы и статистика, 2004. — 287 с.
4. Славин, М.Б. Методы системного анализа в медицинских исследованиях [Текст] / М.Б. Славин. — М.: Медицина, 1989. — 304 с.

Учебное издание

КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ В СРЕДЕ СТАТИСТИЧЕСКОГО ПАКЕТА R

Методические указания к практическим занятиям

Составитель ***Комарова Марина Валериевна***

Редактор Т.С. Петренко
Доверстка Т.С. Петренко

Подписано в печать 20.05.2010. Формат 60×84 1/16.
Бумага офсетная. Печать офсетная.
Печ. л. 1,5
Тираж экз. Заказ .

Самарский государственный
аэрокосмический университет.
443086, Самара, Московское шоссе, 34.

Изд-во Самарского государственного
аэрокосмического университета.
443086, Самара, Московское шоссе, 34.