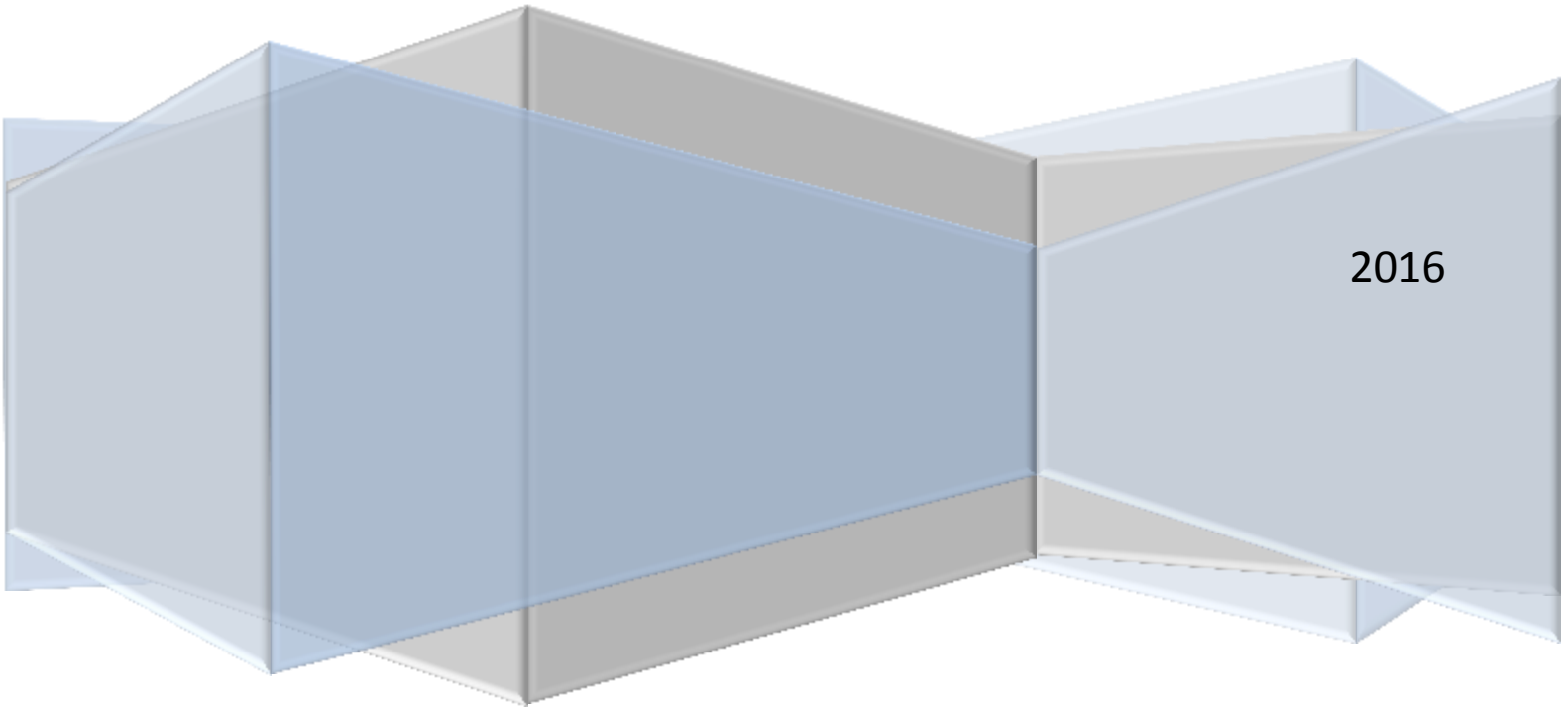
Петрозаводский государственный университет

Прикладная статистика. Решение задач с использованием языка программирования R

Учебное пособие

И. М. Шабалина, Д. П. Косицын, И. В. Пешкова



2016

Аннотация

Предлагаемое пособие содержит теоретические сведения и рекомендации по выполнению лабораторных работ по курсу «Прикладная статистика» студентами математического факультета. Описание каждого метода снабжено необходимыми теоретическими сведениями, приводится обзор библиотек и функций языка R рекомендуемых к использованию, исходные данные, фрагменты скриптов на языке R. В пособии приведены примеры работ по темам: разведочный анализ данных, проверка статистических гипотез, статистическое моделирование, корреляционный и регрессионный анализ, дисперсионный анализ, компонентный анализ.

Пособие предназначено студентам математических факультетов при изучении курсов «Прикладная статистика» и «Случайные процессы». Оно также будет полезно для экономистов и прочих специалистов, занимающихся статистическим анализом данных.

УДК 519.2
ББК 22.172
Ш 122

РЕЦЕНЗЕНТЫ

Е. А. Питухин — д.т.н., профессор кафедры прикладной математики и кибернетики;

А. В. Седов — к.т.н., доцент кафедры теории вероятностей и анализа данных

Введение

В настоящее время статистические методы используются во многих отраслях деятельности человека. Статистический анализ данных, основанный на методах теории вероятностей, математической статистики, случайных процессов, исследовании операций и прочих математических теориях, с успехом применяется для решения практических задач. Кроме того, известно значительное количество программных систем, в которых реализованы библиотеки функций для выполнения процедур анализа данных, например, специализированные пакеты Statistica, SPSS, SAS, STADIA и др. Это мощные дорогостоящие программные продукты, используемые для решения задач, связанных со сбором данных, их исследованием, визуализацией и формированием рекомендаций. Параллельно с развитием указанных пакетов в научной среде и в среде разработчиков широко распространено применение бесплатного программного обеспечения с открытым кодом (open-source решения). Использование такого программного обеспечения не требует материальных затрат на приобретение и поддержку, open-source проекты обычно реализуются заинтересованным сообществом исследователей и разработчиков. При этом значительная часть таких продуктов не имеет удобного интерфейса (командная строка), требует навыков программирования и настройки ПО. Наиболее распространенным open-source продуктом для анализа данных является платформа R, включающая язык написания скриптов, библиотеки программных модулей для решения различных задач, инструменты для разработчиков и аналитиков.

В предлагаемом учебном пособии приводится краткое описание базовых методов анализа данных, перечень библиотек и функций языка R рекомендуемых к использованию при выполнении базовых процедур статистического анализа, исходные данные, фрагменты скриптов на языке R. В пособии приведены примеры работ по темам: разведочный анализ данных, проверка статистических гипотез, статистическое моделирование, корреляционный и регрессионный анализ, дисперсионный анализ, компонентный анализ.

ГЛАВА 1. Базовые сведения о языке R

1.1. Инсталляция системы, использование пакетов программных модулей, инструментов разработчика

Для установки системы используем ресурс https://cran.rstudio.com/. В зависимости от используемой операционной системы скачиваем соответствующий дистрибутив системы R, выполняем установку. На этапе «Выбор компонентов» выбираем «32 bit Пользовательская установка», остальные настройки — по умолчанию.

Наиболее удобным, по мнению авторов, инструментом для написания и выполнения скриптов является среда RStudio, дистрибутив которой можно получить со страницы https://www.rstudio.com/products/rstudio/download/ (настройки — по умолчанию). При работе с системой R все используемые функции предоставляются в составе пакетов. Базовые и рекомендуемые пакеты обычно включаются в инсталляционный файл R. Дополнительные пакеты можно скачивать и устанавливать, например, с русского "зеркала" http://cran.gis-lab.info с помощью функции «Install Packages» в среде RStudio. Также есть возможность скачивания пакетов в виде zip-файлов с сайта http://cran.gis-lab.info/web/packages и последующей установки. Список созданных на сегодняшний день пакетов для R обширен. В [1] приводится список наиболее популярных пакетов. Наименования пакетов требующихся для выполнения лабораторных работ будут указаны в соответствующих частях пособия.

Для обеспечения возможности установки и компиляции пакетов из исходных кодов потребуется дистрибутив утилиты Rtools со страницы https://cran.rstudio.com/bin/windows/Rtools/. При установке утилиты на этапе «Выбор компонентов» выбираем «Full installation» (настройки — по умолчанию).

1.2. Типы данных, импорт-экспорт данных.

В данном разделе будут приведены типы данных и функции, используемые при написании скриптов для выполнения лабораторных работ. Авторы не ста-

вят цель привести подробное описание языка R, поскольку для этого существует большое количество других источников.

Основные типы данных, используемые в R:

- `numeric` — целые (integer) и действительные (double) числа;
- `logical` — логические объекты, принимают значения FALSE (F) и TRUE (T);
- `character` - символьные объекты.

В R можно определять переменные разных типов. При создании имен следует учитывать, что R чувствителен к регистру. Принадлежность переменной к типу можно проверить функциями `is.numeric(<имя>)`, `is.logical(<имя>)`, `is.character(<имя>)`. Для преобразования объекта в другой используются функции `as.numeric(<имя>)`, `as.logical(<имя>)`, `as.character(<имя>)`.

Составные типы данных, используемые в R: `vector` - вектор, без заголовка, `matrix` - это вектор, представленный в виде двумерного массива, `data.frame` — таблица с данными, с названиями столбцов и строк (имена наблюдений и переменных), `factor` — вектор категориальных или порядковых данных, `list` - это структура данных, позволяющий хранить в одной переменной объекты одного или разных типов, в том числе и другие списки. Тип объекта можно определить функцией `str(<имя объекта>)`.

Для загрузки данных можно использовать команды `read.table`, `read.csv`, `read.csv2` и др. Наиболее простой способ получения данных — сформировать таблицу объект-признак (строки — объекты, столбцы — значения признаков) в табличном редакторе (например, MS Excel) и сохранить как CSV файл с разделителями.

Для загрузки полученного файла, как объекта языка R используем команду:

```
> data1 <- read.csv2("data1.csv", sep = ";", header = TRUE,
row.names = 1),
```

где `sep` - вид разделителя, `header` – указывает, есть или нет заголовки у столбцов; `row.names` – указывает номер столбца с названиями наблюдений.

Для выгрузки результатов в таблицу можно использовать функции `write.table`, `write.csv`, `write.csv2` и др. Для уточнения параметров функций следует использовать команду `?<имя_функции>`.

ГЛАВА 2. Основные методы статистического анализа данных

2.1. Разведочный анализ данных

Разведочный анализ данных проводят для установления основных свойств распределений выборочных значений. Основными методами разведочного анализа данных являются интервальное и точечное оценивание параметров распределения, построение графиков оценок плотности распределения вероятностей и эмпирической функции распределения.

Рассмотрим выборку значений $x_1, x_2, \dots, x_N$ случайной величины X, N — объем выборки. В таблице 1 приведены основные точечные оценки вероятностных характеристик распределений и формулы для их вычисления.

Таблица 1. Основные очечные оценки вероятностных характеристик.

Параметр распределения	Статистическая оценка
Математическое ожидание	$m_X = \frac{1}{N}\sum_{i=1}^{N} x_i$
Дисперсия (несмещенная оценка)	$\sigma_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_X)^2$
Среднеквадратическое отклонение	$\sigma_X = \sqrt{\sigma_X^2}$
Начальный момент k -го порядка	$\nu_k = \frac{1}{N}\sum_{i=1}^{N} x_i^2$
Центральный момент k-го порядка	$\mu_k = \frac{1}{N}\sum_{i=1}^{N}(x_i - m_X)^k$
Коэффициент асимметрии	$\gamma = \frac{\mu_3}{\sigma^3}$

Параметр распределения	Статистическая оценка
Коэффициент эксцесса	$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$
Медиана	$m_e = \begin{cases} (x_s^* + x_{s+1}^*)/2\,, s = [N/2], & N - \text{четное} \\ x_s^*, \quad s = [N/2] + 1, & N - \text{нечетное} \end{cases}$ где $x_1^*, x_2^*, \dots, x_N^*$ — вариационный ряд
Размах	$R = x_{max} - x_{min}$

Интервальное оценивание подразумевает вычисление доверительных интервалов для характеристик распределения. Вероятность того, что доверительный интервал $I_\gamma(\Theta) = (\Theta_{\text{выб}} - \varepsilon, \Theta_{\text{выб}} + \varepsilon)$ содержит теоретическое значение параметра распределения Θ равна γ, и называется доверительной вероятностью. Доверительный интервал строится относительно выборочного значения параметра $\Theta_{\text{выб}}$, параметр ε (точность) определяется в зависимости от оцениваемого параметра и известных теоретических сведений о распределении.

Для математического ожидания, если дисперсия распределения неизвестна, доверительный интервал имеет вид $I_m = \left(m_x - t_{1-\frac{\gamma}{2}} \sqrt{\frac{\sigma_X^2}{N}};\ m_x + t_{1-\frac{\gamma}{2}} \sqrt{\frac{\sigma_X^2}{N}} \right)$, где t_α — квантиль уровня $\alpha = 1 - \frac{\gamma}{2}$ распределения Стьюдента с $N - 1$ степенями свободы, σ_X^2 — несмещенная выборочная дисперсия.

Эмпирической функцией распределения (функцией распределения для выборочных значений) называют функцию:

$$F^*(x) = \begin{cases} 0\ ;\ x \le x_{min} \\ \frac{N_i}{N};\ x_i^* < x \le x_{i+1}^*, i = 1, \dots, N-1\ , \\ 1\ ;\ x > x_{max} \end{cases} \qquad (1)$$

где $x_1^*, x_2^*, \dots, x_N^*$ — вариационный ряд, N_i — это количество выборочных элементов $x \le x_i^*$.

Гистограмма распределения — это график позволяющий оценить плотность распределения вероятностей случайной величины X. Гистограмма строится следующим образом. Интервал $[x_{min}; x_{max}]$ разбивается на r непересекающих-

ся интервалов $[x_{k-1}, x_k]$, в каждом из которых определяется количество выборочных значений N_k на интервале (частота) и относительная частота $\nu_k = N_k/N$. Высота «столбика» для каждого интервала $h_k = \nu_k/(x_k - x_{k-1})$, $k = 1, \dots, r$. Внешний вид гистограммы показан на рис.1.

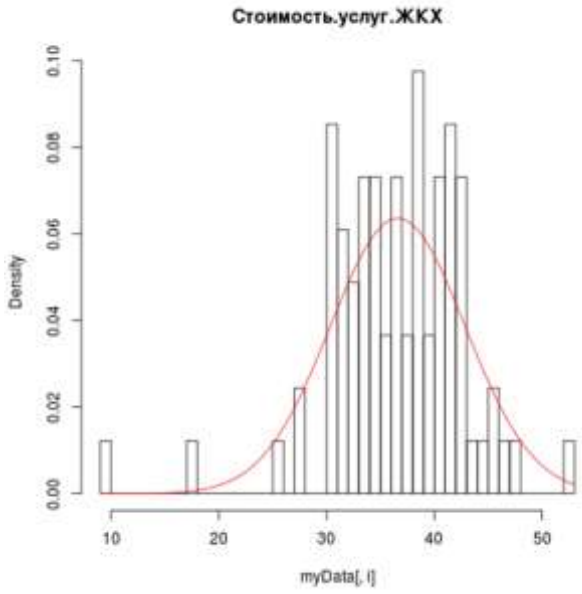


Рис. 1. Гистограмма распределения

2.2. Статистическая проверка гипотез

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известного распределения [2]. В этом случае нулевой (основной) гипотезой H_0 называют выдвинутую гипотезу, альтернативной (конкурирующей) называют гипотезу H_1 противоречащую нулевой.

Статистический критерий — правило (алгоритм), которое, основываясь только на выборочных данных, позволяет принять либо нулевую гипотезу, либо альтернативную [4]. Для проверки гипотез используется специально подобранная случайная величина — статистика критерия, точное или приближенное значение которой известно. На основании рассчитанной по выборочным данным статистике критерия делается вывод о справедливости гипотезы H_0 или H_1.

Статистический критерий, проверяющий гипотезу для выборки $x_1, x_2, \dots, x_N$, характеризуется множеством $W_д$ –допустимых значений, если $x_1, x_2, \dots, x_N \in$

$W_д$, то для этой выборки справедлива гипотеза H_0. Множество $W_k = R^N \setminus W_д$ – образует множество критических значений, если $x_1, x_2, \dots, x_N \in W_k$, то для этой выборки справедлива гипотеза H_1.

В силу случайности выборки при проверке гипотез статистическими критериями, неизбежно возникновение ошибок I-го и II-го рода.

Ошибка I-го рода или уровень значимости критерия — это вероятность $\alpha = p\{(x_1, x_2, \dots, x_N) \in W_к \mid H_0\}$.

Ошибка II-го рода — это вероятность $\beta' = p\{(x_1, x_2, \dots, x_N) \in W_д \mid H_1\}$. Величина $\beta = 1 - \beta'$ называется мощностью критерия.

Алгоритм проверки статистической гипотезы состоит в следующем. По выборочным значениям вычисляется статистика критерия — $S_{выб}$. Затем проверяется ее соответствие теоретическому распределению F_S, которое описывает распределение значений $S_{выб}$ в случае справедливости нулевой гипотезы. Параметры распределения F_S определяются используемым критерием. В случае односторонних критериев значение $S_{выб}$ сравнивается с квантилем, соответствующим заданному уровню значимости критерия. Если $S_{выб} < S_\alpha$, то гипотеза H_0 принимается на уровне значимости α, иначе — принимается гипотеза H_1.

В языке R процедуры проверки гипотез используют понятие p - уровня. При вычислении $S_{выб}$ определяется вероятность того, что СВ S с функцией распределения F_S примет значение большее $S_{выб}$, то есть $p = p(S > S_{выб})$. Данная вероятность является -уровнем для статистики $S_{выб}$. Другими словами — это «уровень значимости», определённый для $S_{выб}$. Если сравнить p и α, то ситуация $p > \alpha$, будет соответствовать ситуации $S_{выб} < S_\alpha$, принимается H_0, если $p < \alpha$, то $S_{выб} > S_\alpha$, гипотеза H_0 отвергается на уровне α.

В данной главе рассматриваются: критерии согласия, критерии проверки однородности и независимости двух выборок, проверки однородности и независимости нескольких выборок. Далее в таблице приведены краткие теоретические сведения о критериях в соответствии с [6],[10].

Таблица 2. Статистические критерии.

Критерий	Описание
Критерии согласия проверяют соответствие эмпирической функции распределения заданной теоретической функции $F_0(x)$; $H_0: F(x) = F_0(x)$; $H_1: F(x) \neq F_0(x)$	
Критерий Колмогорова-Смирнова	Статистика критерия $\rho = \sqrt{N}\max_{x\in(-\infty,\infty)}\lvert F_0(x) - F(x)\rvert$, где $F(x)$ — эмпирическая функция распределения, $F_0(x)$ — теоретическая функция распределения. Статистика ρ имеет распределение Колмогорова: $K(z) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2z^2}$ Для практических расчетов используется формула $\rho = \sqrt{n}\max_{1\le i\le N}\left[\left\lvert F_0(x_i^*) - \frac{2i-1}{2n}\right\rvert + \frac{1}{2n}\right]$, где x_i^*- элемент вариационного ряда, полученное значение ρ сравнивается с квантилем распределения Колмогорова, соответствующим заданному уровню α
Критерий Пирсона χ^2	Для расчета статистики критерия отрезок $[x_{min};\ x_{max}]$ делится на r частей: $\Delta_k, k = 1, \dots, r$. В каждом отрезке Δ_k, определяется n_k — частота попадания выборочных значений в отрезок, p_k- теоретическая вероятность попадания значения СВ в отрезок Δ_k из r. Статистика критерия: $\chi^2_{\text{выб}} = \sum_{k=1}^{r}\frac{(n_k - np_k)^2}{np_k}$ сравнивается с табличным значением $\chi^2_{1-\alpha}(r-l-1)$, случайной величины имеющей распределение χ^2 с $r-l-1$ степенями свободы, l — количество оцениваемых по выборке параметров распределения
Критерии проверки однородности двух выборок $(x_1, x_2, \dots, x_n)$ и $(y_1, y_2, \dots, y_m)$, с функциями распределения $F_1(x)$ и $F_2(x)$ соответственно. $H_0: F_1(x) = F_2(x)$; $H_1: F_1(x) \neq F_2(x)$	
Критерий Колмогорова-Смирнова	Статистика критерия $D = \max_x\lvert F_1^*(x) - F_2^*(x)\rvert$ Для практических расчетов используются формулы: $D^-(n,m) = \max_{1\le i\le n}\left\lvert\frac{i}{n} - F_2^*(x_i)\right\rvert = \max_{1\le j\le m}\left\lvert F_1^*(y_j) - \frac{j-1}{m}\right\rvert;$ $D^+(n,m) = \max_{1\le j\le m}\left\lvert\frac{j}{m} - F_1^*(y_i)\right\rvert = \max_{1\le i\le n}\left\lvert F_2^*(x_i) - \frac{i-1}{n}\right\rvert;$ $D = \max\left(D^-(n,m); D^+(n,m)\right)$
Критерий Вилкоксона	Строится общий вариационный ряд объединённой выборки и вычисляются ранги $r(x_i), r(y_i)$ всех элементов обеих выборок в общем вариационном ряду. Вычисляются суммарные ранги обеих выборок и статистика критерия U: $R_1 = \sum_{k=1}^{m} r(x_k);\ U_1 = nm + \frac{n(n+1)}{2} - R_1$ $R_2 = \sum_{k=1}^{m} r(y_k);\ U_2 = nm + \frac{m(m+1)}{2} - R_2$ $U = min(U_1, U_2)$ сравнивается с квантилем нормального распределения $N\left(\frac{mn}{2}; \frac{mn}{12}(m+n-1)\right)$
Критерий серий	Строится общий вариационный ряд, элементам 1-й выборки ставится «+», элементам 2-й выборки — «-». Последовательность знаков одного вида — серия. Вычисляются n_1, n_2 — количества элементов в сериях.

Критерий	Описание
	Статистика критерия: $$Z = \frac{\left\lvert n - \frac{2n_1n_2}{n_1 + n_2} - 1 \right\rvert - \frac{1}{2}}{\sqrt{\frac{2n_1n_2(2n_1n_2 - (n_1 + n_2))}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}}$$ сравнивается с квантилями нормального распределения N(0,1)
Критерии проверки однородности и независимости нескольких выборок	
Медианный критерий	Строится общий вариационный ряд, вычисляется медиана объединенной выборки. В каждой выборке определяется количество элементов меньше медианы n_i^- и больше медианы n_i^+, а также ожидаемые количества элементов больших и меньших медианы $n_{i\ \text{ожид}}$ (если количество элементов — четное, то n_i /2). Статистика критерия: $$\sum_i \frac{(n_i^-)^2}{n_{i\ \text{ожид}}} + \sum_i \frac{(n_i^+)^2}{n_{i\ \text{ожид}}} - n \quad \sim \chi^2(k-1)$$
Критерий Краскела-Уоллиса	Статистика критерия: $$K = \sum_{j=1}^{k} n_j(\frac{R_j}{n_j} - R)^2 = \sum_{j=1}^{k} \frac{R_j}{n_j}^2 - \frac{n(n+1)^2}{4}$$ $$\frac{12}{n(n+1)} K \sim \chi^2(k-1),$$ где R — средний ранг объединенной выборки, R_j — средний ранг выборки j, n_j — количество элементов в выборке j; n –количество элементов в объединенной выборке.

2.3. Корреляционный анализ данных

Две случайные величины X и Y называют независимыми, если выполнено условие: $F(xy) = F_X(x) \cdot F_Y(y)$, где $F_X(x), F_Y(y)$ функции распределения X и Y, $F(xy)$ — совместная функция распределения X и Y. Для проверки независимости признаков используются гипотезы:

$$H_0: \ F(x,y) = F_x(x)F_y(y)$$
$$H_1: \ F(x,y) \neq F_x(x)F_y(y)$$

Статистические критерии проверки независимости признаков в основном строятся с использованием коэффициентов корреляции.

Коэффициент корреляции Пирсона вычисляется по формуле:

$$r_{XY} = \frac{M[(X - M(X)(Y - M(Y)]}{\delta_X \delta_Y}, \qquad (1)$$

где δ_X, δ_Y — средние квадратичные отклонения величин X и Y. Коэффициент корреляции характеризует тесноту линейной связи между величинами, измеренными в непрерывной шкале. Свойства (1) подробно описаны в [2, 4, 5].

Статистический аналог коэффициента корреляции Пирсона вычисляется по формуле:

$$\tilde{r}_{XY} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \tilde{m}_x)(y_i - \tilde{m}_y)}{\tilde{\delta}_X \tilde{\delta}_Y}, \quad (2)$$

где $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ — выборка значений случайного вектора (X, Y), $\tilde{m}_x, \tilde{m}_y$ — выборочные средние для X и Y, n — объем выборки. Для установления наличия линейной зависимости используется статистический критерий, проверяющий гипотезу H_0: «величины независимы». В случае справедливости нулевой гипотезы на уровне значимости α, выполняется условие:

$$\tilde{r}_{XY}\sqrt{\frac{\tilde{r}_{XY} - 2}{1 - \tilde{r}_{XY}^2}} < t_{1-\frac{\alpha}{2}}(n-2), \quad (3)$$

где $t_{1-\frac{\alpha}{2}}(n-2)$ — квантиль уровня $1-\frac{\alpha}{2}$ распределения Стьюдента с $n-2$ степенями свободы.

Для установления взаимосвязи величин измеренных в номинальной или ранговой шкале используются статистические критерии, основанные на ранговых коэффициентах корреляции, и критерий χ^2 для проверки гипотезы H_0 о независимости признаков.

При вычислении коэффициента корреляции Спирмена определяются ранги элементов выборок $(x_1, x_2, \ldots x_n)$ и $(y_1, y_2, \ldots y_n)$ и вычисляются разности $d_i = rank(x_i) - rank(y_i)$ для -й пары значений. Значение коэффициента вычисляется по формуле:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}. \quad (4)$$

Проверка справедливости гипотезы H_0 выполняется аналогично (3).

При вычислении коэффициента корреляции Кендалла упорядочиваем выборку по значениям признака X. Ранжируют значения показателя Y и рассчитывают коэффициент корреляции Кендалла.

$$\tau = \frac{2S}{n(n-1)}, \qquad \text{где } S = P - Q \tag{5}$$

P — суммарное число наблюдений, следующий за текущими наблюдениями с большим значением рангов признака Y (выборка упорядочена по возрастанию X), Q - суммарное число наблюдений, следующих за текущими наблюдениями с меньшим значением рангов Y (равные ранги в расчет не берутся).

В случае справедливости H_0 на уровне значимости α, выполняется условие:

$$\tau < t_{1-\frac{\alpha}{2}}, \tag{6}$$

где $t_{1-\frac{\alpha}{2}}$ - квантиль нормального распределения с параметрами $\left(0; \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)$

Проверка независимости признаков при помощи критерия χ^2 осуществляется следующим образом. Строится таблица сопряжённости признаков следующего вида:

$Y\backslash X$	x_1	x_2	…	x_n	Сумма
y_1	n_{11}	n_{12}		n_{1n}	$n_{1\cdot}$
y_2	n_{21}	n_{22}			$n_{2\cdot}$
…	…	…	…	…	…
y_m	n_{m1}	n_{m2}		n_{mn}	$n_{m\cdot}$
Сумма	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot n}$	n

где $x_1, x_2, \dots, x_n$ — градации признака X, $y_1, y_2, \dots, y_m$ — градации признака Y, n_{ij} — количество объектов с i-ой градацией Y и -ой градацией X, $n_{\cdot j}$ — количество объектов с j-ой градацией X, $n_{i\cdot}$ — количество объектов с i-ой градацией Y.

Статистика критерия вычисляется по формуле:

$$X^2_{\text{выб}} = \sum_{ij} \frac{\left(n_{ij} - \tilde{n}_{ij}\right)^2}{\tilde{n}^2_{ij}}, \quad \tilde{n}_{ij} = \frac{n_{i.} n_{.j}}{N} \tag{5}$$

где $\tilde{n}_{ij}$ — «выравнивающая» или «ожидаемая» частота. Если выполнено условие

$X^2_{\text{выб}} < X^2_{1-\alpha}\big((n-1)(m-1)\big)$, то H_0 принимается на уровне значимости α.

2.4. Основы регрессионного анализа данных

Рассматривается случайная величина Y и случайный вектор $X = (X_1, \dots, X_n)$, при этом предполагается, что компоненты вектора X оказывают влияние на Y. В этом случае, при самых общих предположениях, можно представить зависимость результирующего признака Y и исходных признаков $X_1, \dots, X_n$ в виде следующей регрессионной модели [5]:

$$Y = f(X) + \Delta_X = Y_X + \Delta_X \tag{6}$$

где Y_X — случайная величина, индуцированная функцией регрессии $f(X)$, Δ_X — ошибка регрессии (остаток, отклонение), значение которого мало в смысле выбранной меры. Наиболее часто в практических расчетах используется мера в форме дисперсии ошибки регрессии $D[\Delta_X]$, функция регрессии доставляющая минимальное значение $D[\Delta_X]$ называется регрессией оптимальной в среднем квадратичном смысле.

Если заданы распределения $\rho_X(x)$, $\rho_Y(y)$, $\rho(x,y)$, случайные ведичины $Y, X_1, \dots X_n$ имеют конечные ненулевые дисперсии, то оптимальная среднеквадратичная регрессия имеет вид:

$$\varphi(X) = M[\,Y/X] = \int_{-\infty}^{\infty} y \frac{\rho(x,y)}{\rho_X(x)} dy, \tag{7}$$

где $M[\,Y/X]$ — условное математическое ожидание [2].

Наиболее важными для практических расчетов являются линейные регрессионные модели. Такие модели являются оптимальными в случае нормального распределения Y и X, в некоторых случаях нелинейные модели могут быть сведены к линейным [4].

Функция линейной регрессии определяется следующим образом:

$$\varphi(X) = b_0 + \sum_{k=1}^{n} b_k(X_k - M[X_k]). \quad (8)$$

Коэффициенты модели рассчитываются по формулам:

$$b_0 = M[Y],\ b_k = \frac{\sigma_Y}{\sigma_k}\beta_k,\ \ k = 1, \dots, n,$$
$$\beta = K_X^{-1}\bar{r}_{XY},\ \bar{r}_{XY} = (r_{1Y}, \dots, r_{nY})^T, \quad (9)$$

где σ_Y — среднее квадратичное отклонение Y, σ_k — среднее квадратичное отклонение X_k, r_{kY} — коэффициенты корреляции между X_k и Y, $k = 1, \dots, n$, K_X — корреляционная матрица для вектора X, $|K_X| \neq 0$.

При практическом использовании формул (8) и (9) используются статистические оценки вероятностных характеристик.

Анализ качества построенной модели предлагается проводить в несколько этапов: вычисление коэффициента детерминации, проверка значимости функции регрессии и коэффициентов модели, анализ остатков.

Вычисление коэффициента детерминации и дисперсии ошибки.

Наиболее важным показателем качества построенной модели является коэффициент детерминации $R^2 = (K_X^{-1}r_{XY}, r_{XY})$. Данный показатель связан с дисперсией ошибки и дисперсией Y следующим соотношением:

$$D[\Delta_X] = D[Y](1 - R^2) \quad (10)$$

что позволяет заметить:

1) Если $R^2 \approx 0$, то $D[\Delta_X] \approx D[Y]$ — модель не может использоваться для прогноза значения Y;
2) Если $R^2 \approx 1$, то $D[\Delta_X] \approx 0$ — модель практически полностью определяет значение Y.

Для равнения моделей используют исправленный коэффициент детерминации (adjusted R^2), вычисляемый по формуле:

$$R^2_{adj} = R^2 - (1 - R^2)\frac{n-1}{N-n}. \quad (11)$$

Также для оценки качества модели можно вычислить оценку дисперсии ошибки $D[\Delta_X]$ и сравнить ее значение с $D[Y]$.

Проверка значимости регрессии.

Осуществляется проверка гипотезы H_0: $b_1 = b_2 = \cdots = b_n = 0$. Статистика критерия для проверки H_0:

$$\gamma = \frac{R^2}{1-R^2} \cdot \frac{N-n-1}{n}, \tag{12}$$

где N — объем выборки, n — количество исходных признаков. В случае справедливости H_0 статистика (12) имеет распределение Фишера с n и $N-n-1$ степенями свободы $F(n,\ N-n-1)$.

Проверка значимости коэффициентов регрессии.

Для каждого коэффициента b_k, $k = 1, \dots, n$, проверяется гипотеза о равенстве нулю коэффициента H_0: $b_k = 0$. Статистика критерия:

$$\gamma = \frac{b_k}{D[\Delta_X]\sqrt{a_{kk}}}, \tag{13}$$

где a_{kk} — диагональный элемент матрицы K_X^{-1}. В случае справедливости гипотезы γ имеет распределение Стьюдента с $N-n-1$ степенями свободы. Переменные модели X_k для коэффициентов которых гипотеза не была отвергнута, могут быть последовательно исключены из модели. Не рекомендуется исключать сразу все переменные, поскольку при исключении одной коэффициенты модели пересчитываются и проверка гипотез может показать другой результат.

Анализ остатков.

Линейная регрессионная модель строится в предположении, что остатки имеют нормальное распределение, центрированы и независимы от исходных признаков, результирующего признака и предсказанных значений.

При расчетах, как правило, проверка нормальности заменяется построением гистограммы остатков по которой оценивается симметричность распределения, одномодальность и близость среднего значения к нулю.

Независимость остатков также можно оценить по графикам — диаграммам рассеивания, построенным для остатков и остальных переменных модели.

При выполнении указанных выше условий, высоком значении коэффициента детерминации можно апробировать модель на данных, не участвовавших в расчетах при построении модели (контрольная выборка). Если ошибки прогноза на контрольной выборке не превышают допустимые границы (определяются в процентах от средней ошибки при построении модели) можно рекомендовать построенную модель к применению.

На основе предложенного алгоритма построения линейной регрессионной модели можно построить модели показательной, логарифмической, степенной, полиномиальной регрессии методом частичной нормализации [5].

2.5. Однофакторный дисперсионный анализ

Основная идея дисперсионного анализа (ДА) состоит в том, чтобы проверить гипотезу о несущественности влияния неколичественных факторов $A, B, C, \ldots$ на количественный результирующий признак Y [5]. В случае отвержения гипотез строится модель взаимосвязи Y и факторов $A, B, C, \ldots$.

Рассмотрим однофакторную модель дисперсионного анализа, для построения которой необходимо проверить влияние фактора A на Y. Фактор A характеризуется уровнями или значениями, которые может этот фактор принять. Поскольку фактор может принять только одно из возможных значений, то события «фактор A принял значение $a_k, k = 1,2, \ldots m$, составляют полную группу несовместных событий $H_1, H_2, \ldots H_m$. В основе расчета статистики критерия лежат формула разложения математического ожидания Y по уровням фактора A:

$$M[Y] = \sum_{k=1}^{m} M[Y/H_k]P(H_k) = \sum_{k=1}^{m} \mu_k \, P(H_k) \, , \qquad (14)$$

где $M[Y/H_k]$ — условное математическое ожидание (групповое среднее), $P(H_k)$ — вероятность -го уровня фактора, $k = 1,2, \ldots m$, и формула разложения дисперсии Y по уровням фактора A:

$$\sigma_Y^2 = \sum_{k=1}^{m}(\mu_k - \mu)^2 P(H_k) + \sum_{k=1}^{m}\sigma_k^2 P(H_k) = \sigma_{\text{факт}}^2 + \sigma_{\text{ост}}^2, \quad (15)$$

где σ_Y^2 — дисперсия Y, $\mu = M[Y]$ — математическое ожидание Y, $\sigma_k^2 = D[Y/H_k]$ — условная дисперсия, $\sigma_{\text{факт}}^2$ — факторная дисперсия, $\sigma_{\text{ост}}^2$ — остаточная дисперсия.

Проверка гипотезы о несущественности влияния фактора A на Y осуществляется для выборочных значений $y_{ik}, i = 1, \dots, n_k, \quad k = 1, \dots, m,$ сгруппированных по уровням A, n_k — количество выборочных значений с k-м уровнем A. Для такой выборки справедливы следующие выражения:

$$\sum_{k=1}^{m} n_k = N, \quad p_k = \frac{n_k}{N}, \quad \bar{y}_k = \sum_{i=1}^{n_k} y_{ik}, \quad \bar{\bar{y}} = \sum_{k=1}^{m} \bar{y}_k \, p_k, \quad (16)$$

где N — объем выборки, p_k статистическая оценка $P(H_k)$, $\bar{y}_k$ — групповое среднее, $\bar{\bar{y}}$ — общее среднее. Аналогом формулы разложения дисперсии является выражение:

$$Q_{\text{общ}}^2 = Q_{\text{факт}}^2 + Q_{\text{ост}}^2; \quad (17)$$

где $Q_{\text{факт}}^2$ - факторная сумма квадратов отклонений, статистический аналог факторной дисперсии :

$$Q_{\text{факт}}^2 = \sum_{k=1}^{m} n_k(\bar{y}_k - \bar{\bar{y}})^2, \quad (18)$$

$Q_{\text{ост}}^2$ — остаточная сумма квадратов отклонений, статистический аналог остаточной дисперсии:

$$Q_{\text{ост}}^2 = \sum_{k=1}^{m}\sum_{j=1}^{n_k}(y_{jk} - \bar{y}_k)^2. \quad (19)$$

Гипотеза о несущественности влияния фактора А на признак Yпроверяется при помощи критерия Фишера. Статистикой критерия является величина:

$$\Theta_{\text{В}} = \frac{\frac{1}{m-1} Q_{\text{факт}}^2}{\frac{1}{N-m} Q_{\text{ост}}^2}. \quad (20)$$

В случае справедливости гипотезы $\Theta_{\text{В}}$ имеет распределение Фишера с $m-1$ и $N-m$ степенями свободы. Применение критерия Фишера возможно, если Y имеет нормальное распределение.

Если гипотеза отвергается, то можно строить регрессионную модель зависимости Y от A:

$$Y = Y_A + \Delta, \tag{21}$$

где Y_A — случайная величина, принимающая значения μ_k, в зависимости от уровня фактора A. Коэффициент детерминации данной модели определяется по формуле:

$$R_A^2 = \frac{\frac{m-1}{N-m}\Theta_{\text{В}}}{1+\frac{m-1}{N-m}\Theta_{\text{В}}}. \tag{22}$$

Для модели (21) справедливы свойства линейных регрессионных моделей.

Многофакторные модели дисперсионного анализа строятся по тем же принципам, что и однофакторные. Определяют два типа моделей: с учетом взаимного влияния признаков и без учета влияния. При построении многофакторных моделей проверяются гипотезы о несущественности влияния каждого фактора на результирующий признак Y, При учете в модели взаимного влияния факторов — проверяется гипотеза для каждого сочетания факторов, размерность модели при этом существенно увеличивается.

2.6. Компонентный анализ

Основная идея метода заключается в том, что несколько переменных сильно коррелирует между собой. Это означает, что они либо взаимно определяют друг друга, либо связь обусловлена какой-то третьей величиной, которую непосредственно измерить нельзя, то есть измеряемые величины (исходные признаки $X_1, X_2, \dots, X_p$) являются лишь формой проявления величины, остающейся на заднем плане. Таким образом, задача компонентного анализа состоит в выявлении величины, которая объяснила бы наблюдение связи между переменными. Так, симптомы заболевания могут быть связаны сильными зависимостями, но

при этом исходной причиной такой взаимосвязи является наличие заболевания. Компонентный анализ (КА) подразумевает проведение анализа данных в ходе которого выявляются такие «неявные» переменные, объясняющие присутствующие в исходных данных сильные взаимосвязи. Компонентный анализ — это метод многомерного статистического анализа, исследующий внутреннюю структуру корреляционной или ковариационной матриц.

Рассмотрим выборочные значения $(x_{i1}, x_{i2}, \dots, x_{ip})$, где p — размерность пространства признаков, $i = 1, \dots, N$ — номера наблюдений, N — объем выборки. Если представить выборочные значения признаков для каждого наблюдения в виде точек в –мерном пространстве, то при наличии зависимостей между признаками и близости распределения значений признаков к нормальному (симметричное, одномодальное) точки образуют в пространстве облако в форме эллипсоида. Вдоль главной оси данного эллипсоида сосредоточен наибольший разброс исходных признаков. Вектор, вдоль которого расположена главная ось эллипсоида, может использоваться для измерения признака, являющегося линейной комбинацией исходных признаков. Такой признак является обобщающим признаком или компонентой.

С вероятностной точки зрения задача поиска компонент — это поиск линейных комбинаций исходных признаков, обладающих наибольшей дисперсией и некоррелированных между собой.

Теорема (о главных компонентах).

Если $\vec{X}$ — p-мерный случайны вектор-столбец с $M[\vec{X}] = 0$ и корреляционной матрицей $R = M[\vec{X}\vec{X}^T]$, тогда существует ортогональное преобразование B:

$$\vec{U} = B^T \vec{X} \tag{23}$$

такое, что

$$M[UU^T] = \Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix}, \tag{24}$$

где $\lambda_1, \dots, \lambda_p$ — собственные числа матрицы R.

Столбцы матрицы B определяются уравнением:

$$R\vec{\beta} = \lambda\vec{\beta}, \tag{25}$$

где $\vec{\beta}$ — собственные вектора матрицы R.

Компоненты вектора $\vec{U}$ вычисляются по формуле

$$U_{(s)} = \vec{\beta}^{(s)T}\vec{X}, \tag{26}$$

и обладают наибольшей дисперсией среди линейных комбинаций некоррелирующих с $U_{(1)}, \dots, U_{(s-1),}\ s = 1, \dots, p$.

Свойства компонент определяются свойствами собственных чисел и собственных векторов матриц. Так, количество выделенных компонент соответствуют рангу матрицы R. Основным отличием алгоритма компонентного анализа от стандартного вычисления собственных чисел и собственных векторов матриц является последовательное определение компонент в порядке убывания их дисперсий. Тем не менее, следует отметить следующие свойства:

1. Дисперсия компоненты равна соответствующему собственному числу:

$$D[U_s] = \lambda_s,\ \ s = 1, \dots, p. \tag{27}$$

2. Свойство «сохранения дисперсии»:

$$\sum_{s=1}^{p} D[U_s] = \sum_{k=1}^{p} D[X_k]. \tag{28}$$

Для интерпретации полученных результатов КА используются факторные нагрузки a_{ks} признака k в компоненте s, $k = 1, \dots, p,\ \ s = 1, \dots, p$. Матрица факторных нагрузок $A = \{\alpha_{ks}\}_{k,s=1}^{p}$ определяется по формуле:

$$A = B\Lambda^{1/2},$$

$$\Lambda^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_p} \end{pmatrix}. \tag{29}$$

Для оценки вклада компонент в суммарную дисперсию используется формула:

$$\mu_r = 100\% \cdot \frac{\lambda_r}{\sum_{s=1}^{p} \lambda_s}, \quad r = 1, \dots, p. \tag{30}$$

В результате определения компонент и их вкладов в суммарную дисперсию значительная доля дисперсии (более 90%) может прийтись только на первые m компонент. В этом случае исходные признаки имеют сильные взаимосвязи и интерпретировать (описывать смысл нового признака-компоненты) следует только первые компоненты — их называют «главные компоненты». В некоторых источниках компонентный анализ называют «метод главных компонент». График убывания собственных чисел позволяет сравнить дисперсии компонент и увидеть компоненты с наивысшими дисперсиями.

Если дисперсии компонент отличаются незначительно, то, как правило, исходные признаки имеют слабые взаимосвязи и в каждой компоненте наибольший вклад имеет только один из исходных признаков. Можно интерпретировать такую ситуацию, как отсутствие обобщающих признаков (компонент) для исходных признаков.

ГЛАВА 3. Решение задач с использованием языка R

3.1. Описание исходных данных

Исходными данными для решения задач является статистическая информация о показателях экономики в регионах РФ за 2010 год. Данные содержатся в файле «data1.csv», содержимое файла приведено в Приложении 1. Для исследований используются данные о количестве автомобилей на 100 тыс. чел. («data1.csv», столбец “avto”), цены за кв.м. жилья на первичном рынке (руб.,“house”), уровень доходов населения (руб. в мес, “expenses”), число занятых в экономике (тыс. чел.,“occupied”), стоимость товаров фиксированного набора (руб., “goods”), три группы регионов (столбец “**reg**”): центральная и северо-западная часть (“c”); юг (“s”); восточная часть (“i”).

3.2. Проведение разведочного анализа данных

В рамках проведения разведочного анализа данных требуется для выборочных значений параметра “avto” требуется: вычислить точечные оценки математического ожидания, дисперсии, среднеквадратичного отклонения, начальных и центральных моментов до 4го включительно, медианы, коэффициента асимметрии, коэффициента эксцесса, построить для группированной выборки гистограмму и эмпирическую функцию распределения.

Таблица 3. Пакеты и функции в R для разведочного анализа

Функция	Описание основных параметров	Назначение
Пакет **base**		
summary(x, …)	**x** —данные (vector/object)	Вычисляет основные характеристики вектора: мин, макс, медиану, среднее, 1й и 3й квартиль
Пакет **stats**		
mean(x, …)	**x** — данные (vector/object)	Вычисление оценки математического ожидания
var(x, …)	**x** — данные (vector/object/matrix)	Вычисление оценки дисперсии
sd(x, …)	**x** — данные (vector)	Вычисление оценки среднеквадратическ ого отклонения
median(x, …)	**x** — данные (vector/object)	Оценка медианы
Пакет **moments**		
all.moments(x, order.max, central, absolute, …)	**x** —данные (vector/matrix/data frame); **order.max** — максимальный порядок моментов (по умолчанию 2); **central** — логическая переменная, если установлена (TRUE), то вычисляются центральные моменты заданных порядков; **absolute** — логическая переменная, если установлена (TRUE), то вычисляются абсолютные моменты заданных порядков.	Вычисление моментов: начальных, центральных, абсолютных.
kurtosis(x, na.rm, …)	**x** — данные (тип: numeric vector/matrix/data frame)	Коэффициент асимметрии

Функция	Описание основных параметров	Назначение
skewness(x, na.rm, ...)	**x** — данные (тип: numeric vector/matrix/data frame)	Подсчет коэффициента эксцесса
ecdf(x, …)	**x** — данные (тип: numeric vector/matrix/data frame)	Вычисление значений функции распределения
qnorm(p, mean = 0, sd = 1, …)	**p** — вероятность; **mean** — среднее; **sd** — среднеквадратическое отклонение	Вычисление квантиля нормального распределения
qt(p, df, ...)	**q** — порядок квантиля; **df** — степени свободы	Вычисление квантиля распределения Стьюдента
Пакет **graphics**		
hist(x, breaks, freq, col, main, xlim, ylim, xlab, ylab, add=TRUE, ...)	**x** — данные (vector); **breaks** — количество интервалов группировки; **xlim, ylim** — пределы изменения по осям абсцисс и ординат; **freq=FALSE** — указывает, что при построении гистограммы по оси ординат откладывается не абсолютная частота, а плотность относительной частоты (freq обеспечивает нормировку гистограммы по площади); **col** — задает цвет заливки; **main, xlab, ylab** — позволяют установить подписи для графика (заголовок, ось абсцисс, ординат); **add = TRUE** — указывает, что на гистограмму будет нанесен еще один график	Построение гистограммы распределения
plot(x, xlim, ylim, main, xlab, ylab, ...)	**x** — координаты точек графика; **xlim, ylim, main, xlab, ylab** — аналогично **hist**	Построение графика
curve(func, col, add = TRUE)	**func** — функция кривой; **col** — цвет; **add = TRUE** — вывод в уже открытое графическое устройство	Вывод кривой, соответствующей функции **func**

Функция	Описание основных параметров	Назначение
par(mfrow=c(v, g), ...)	**mfrow=c(v,g)** — определяет количество объектов в одной графической области: **v** — по вертикали, **g** — по горизонтали.	Установка или запрос параметров графического изображения

Общие параметры для функций:

- **na.rm** — логическая переменная, если установлена (TRUE), то пустые значения выборки удаляются из анализа.
- **...** — другие параметры (для справки исп. команду «?<имя_ функции>»)

При работе с типом данных `dataframe`, можно обращаться к отдельным столбцам таблицы либо по названию столбца: `data1[,"признак1"]`, либо используя номер столбца: `data1[,1]` (столбец, содержащий названия наблюдений имеет номер 0).

Далее приведены команды языка R и результаты их выполнения.

```
> summary(data1)
 Min.   : 59.1          Median :215.3          3rd Qu.:239.8
 1st Qu.:193.7          Mean   :213.8          Max.   :372.3
> avto <- data1[,1]
> var(avto)
[1] 2358.068
> sd(avto)
[1] 48.55994
> median(avto)
[1] 215.3
> all.moments(avto, order.max = 4, central = TRUE)
[1]  1.000000e+00  4.856874e-15
2.328219e+03 -2.337863e+04  2.504315e+07
> all.moments(avto, order.max = 4, absolute = TRUE)
[1] 1.000000e+00 2.138025e+02
4.803974e+04 1.124320e+07 2.733151e+09
> kurtosis(avto)
[1] 4.619992
> skewness(avto)
[1] -0.2081054
> hist(avto, freq = FALSE)
> plot(ecdf(avto), ylab="Fn(x)", main = "F(x)", verticals = FALSE)
```

Примечание. Вычисление доверительных интервалов следует выполнить с использованием известных формул [2] и функций вычисления квантилей нормального распределения и распределения Стьюдента.

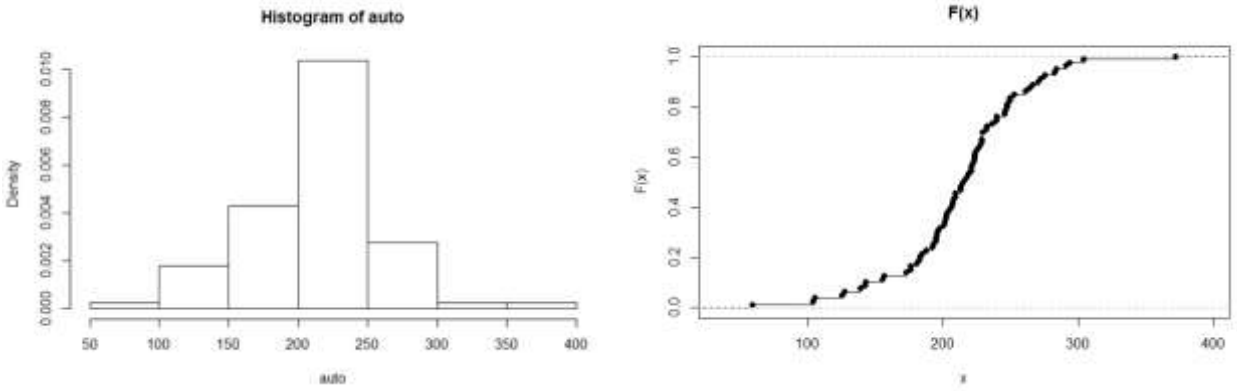


Рис. 2. Гистограмма и функция распределения

3.3. Проведение статистической проверки гипотез

В рамках проведения статистической проверки гипотез требуется:

a) Для показателя “avto” проверить гипотезу о согласованности выборочного распределения с нормальным и показательным законом распределения.
b) Проверить гипотезу об однородности значений показателя “avto” для регионов из северо-западной части и восточная части страны (“reg” принимает значения “c” и “i”).
c) Исследовать однородность показателя “avto” в трех группах (“c”, “i”, “s”)

Полученные результаты следует подтвердить графиками. Проверку гипотез осуществить на уровне значимости α=0.05.

Таблица 4. Пакеты и функции в R для проверки гипотез

Функция	Описание основных параметров	Назначение
Пакет **stats**		
`ks.test(x, y, [mean], [sd])`	**`x`** — выборка значений случайной величины *X*; **`y`** — выборка значений случайной величины *Y*, для теста на однородность двух выборок **`mean, sd`** — параметры распределения *X* для критерия согласия, **`"norm", "pexp", "punif"`** — вид распределения (нормальное, экспоненциальное, равномерное)	Критерий Колмогорова-Смирнова для одной выборки (критерий согласия), для проверки однородности двух выборок

Функция	Описание основных параметров	Назначение
chisq.test(x,p)	**x** — вектор частот на интервалах; **p** — теоретические вероятности на интервалах	Критерий Пирсона χ^2, для проверки согласованности распределения
wilcox.test(x,y) **wilcox.test(x~y)**	**x** — выборка значений случайной величины X; **y** — выборка значений случайной величины Y; **x~y -** x — выборка, y — группирующая переменная	Проверка однородности двух выборок по критерию Вилкоксона
runs.test(x, plot.it = FALSE)	**x** — выборка значений случайной величины (для 2 выборок x, y используется функция c(x, y) — создание вектора); **plot.it** — логический флаг: строить график критерия (TRUE) или нет (FALSE)	критерий серий
kruskal.test(x)	**x** — список (list) из исследуемых выборок. Для создания списка из N объектов используется функция **list(x1,…, xN)**	Критерий Краскела-Уоллиса
dnorm(x, mean, sd)	**x** — вектор квантилей; **mean, sd** — параметры нормального распределения	Плотность нормального распределения
dexp(x, rate)	**x** — вектор квантилей; **rate** — параметр показательного распределения	Плотность экспоненциального распределения
dunif(x, min, max)	**x** — вектор квантилей; **min, max** — границы отрезков	Плотность равномерного распределения
Пакет ***nortest***		
pearson.test(x, n.classes)	**x** — выборка значений; **n.classes** — количество отрезков при разбиении	Критерий χ^2 для проверки гипотезы о нормальности распределения
lillie.test(x)	**x** — выборка значений	Критерий Лиллефорса для проверки нормальности
Пакет ***MASS***		
fitdistr(x, densfun, …)	**x** — выборка значений;	Функция расчета параметров

Функция	Описание основных параметров	Назначение
	densfun — вид распределения: "exponential", “normal” (равномерное — не входит). Возвращает: **estimate**-параметры распределения	распределения
Пакет ***graphics***		
boxplot(x, ...)	**x** — выборка значений	Функция построения графика «ящик с усами»
Пакет ***ggplot2***		
ggplot(x, aes(…))+ geom_histogram(…)+ geom_density(…)+ geom_boxplot(…)	**x** — выборка значений (**dataframe**); **aes** — определяет данные и способ их отображения в графическом объекте; **geom_histogram** — гистограмма; **geom_density** — плотность распределения; **geom_boxplot** — диаграмма «размаха»	Функция инициализации графического объекта с гистограммой и оценкой плотности

Все приведенные в таблице критерии возвращают значения: **statistic** — значение статистики критерия, **p.value** — p-уровень для вычисленной статистики критерия. Прочие возвращаемые значения можно уточнить в блоке «Value» руководства по соответствующей функции.

Далее приведены команды языка R и результаты их выполнения.

а) Критерии согласия для нормального распределения, гистограмма.

```
> avto <- data1[,"avto"]
> res = lillie.test(avto)
> res
Lilliefors (Kolmogorov-Smirnov) normality test
data:  avto
D = 0.1014, p-value = 0.0432

> res = pearson.test(avto)
> res
     Pearson chi-square normality test
data:  avto
P = 10.7722, p-value = 0.2917

> hist(avto, freq = FALSE, breaks = 50, main = “avto”)
> curve(dnorm(x, mean = mean(avto), sd = sd(avto)), col = "red",
add = TRUE)
```

Проверка гипотезы по критериям показывает противоречивые результаты: - уровень для критерия Колмогорова-Смирнова равен 0.0432 ($p < \alpha$), для критерия χ^2 p-уровень 0.29 ($p > \alpha$). Однако, гистограмма распределения показывает сильные отличия от нормального графика, поэтому в данном случае следует сделать вывод об отвержении гипотезы о нормальном распределении.

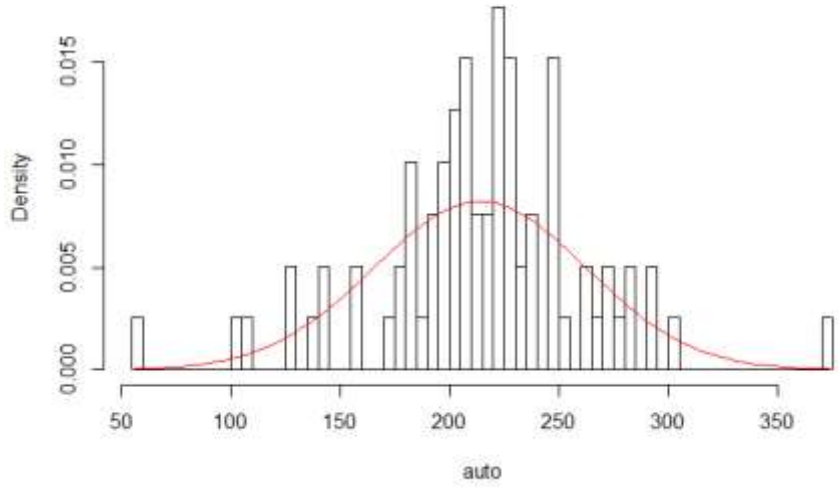


Рис. 3. Сравнение гистограммы с графиком нормального распределения

б) Критерии согласия для показательного распределения, гистограмма.

```
> par <- fitdistr(avto, "exponential")
> res = ks.test(avto, "pexp", par$estimate)
> res
     One-sample Kolmogorov-Smirnov test
data:  avto
D = 1, p-value = 2.22e-16
alternative hypothesis: two-sided

> # разбиение на интервалы для Chi-squared test
> avto.hist<-hist(avto,breaks= 10,plot=FALSE)

> # расширение интервалов
> k <- length(avto.hist$breaks)-1
> avto.hist$breaks[1] <- (-Inf)
> avto.hist$breaks[k+1] <- (+Inf)

> # теоретические частоты
> par <- fitdistr(avto, "exponential")
> avto.p.theor<- pexp(avto.hist$breaks,par$estimate)
> avto.p.theor<-(avto.p.theor[2:(k+1)]-avto.p.theor[1:k])
> chisq.test(avto.hist$counts,p=avto.p.theor)

     Chi-squared test for given probabilities
```

```
data:  avto.hist$counts
X-squared = 245.1606, df = 6, p-value < 2.2e-16

> hist(avto, freq = FALSE, breaks = 50, main = "avto");
> curve(dexp(x, rate = par$estimate), col = "red", add = TRUE)
```

В ходе проверки гипотеза о показательном распределении выборочных данных была отвергнута на уровне значимости 0.05 ($p < \alpha$).

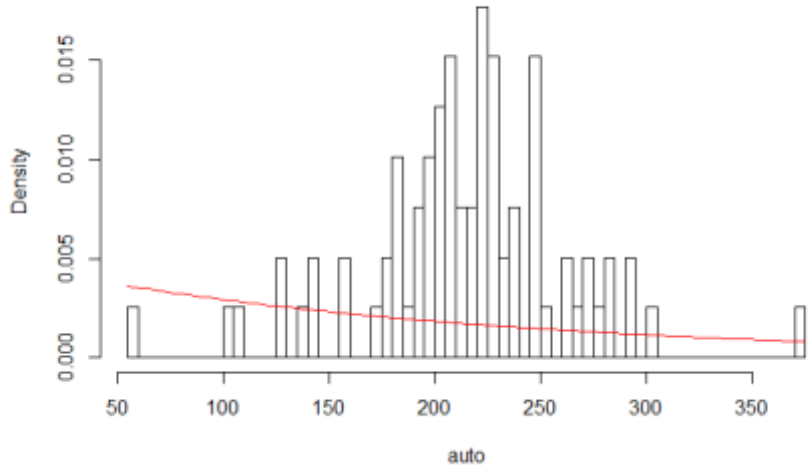


Рис. 4. Сравнение гистограммы с графиком показательного распределения

в) Критерии однородности двух выборок.

```
> x <- data1$avto[data1$reg == "c"]
> y <- data1$avto[data1$reg == "i"]
> ks.test(x,y)
     Two-sample Kolmogorov-Smirnov test
data:  x and y
D = 0.1088, p-value = 0.9756
alternative hypothesis: two-sided

> wilcox.test(x, y)
     Wilcoxon rank sum test
data:  x and y
W = 545, p-value = 0.8832
alternative hypothesis: true location shift is not equal to 0
```

Проверка показала, что гипотеза об однородности значений показателя «количество автомобилей» для двух групп регионов принята на уровне значимости 0.05 ($p > \alpha$).

г) Критерий однородности нескольких выборок.

```
> z <-data1$avto[data1$reg == "s"]
> l <- list(x,y,z)
> kruskal.test(l)

     Kruskal-Wallis rank sum test
data:  l
Kruskal-Wallis chi-squared = 2.4561, df = 2, p-value = 0.2929
> old.par <- par(mfrow=c(3,1))
> hist(x, freq = FALSE, breaks = 20, main = "c", col = "blue")
> hist(y, freq = FALSE, breaks = 20, main = "i", col = "red")
> hist(z, freq = FALSE, breaks = 20, main = "i", col = "grey")
> par(old.par)
```

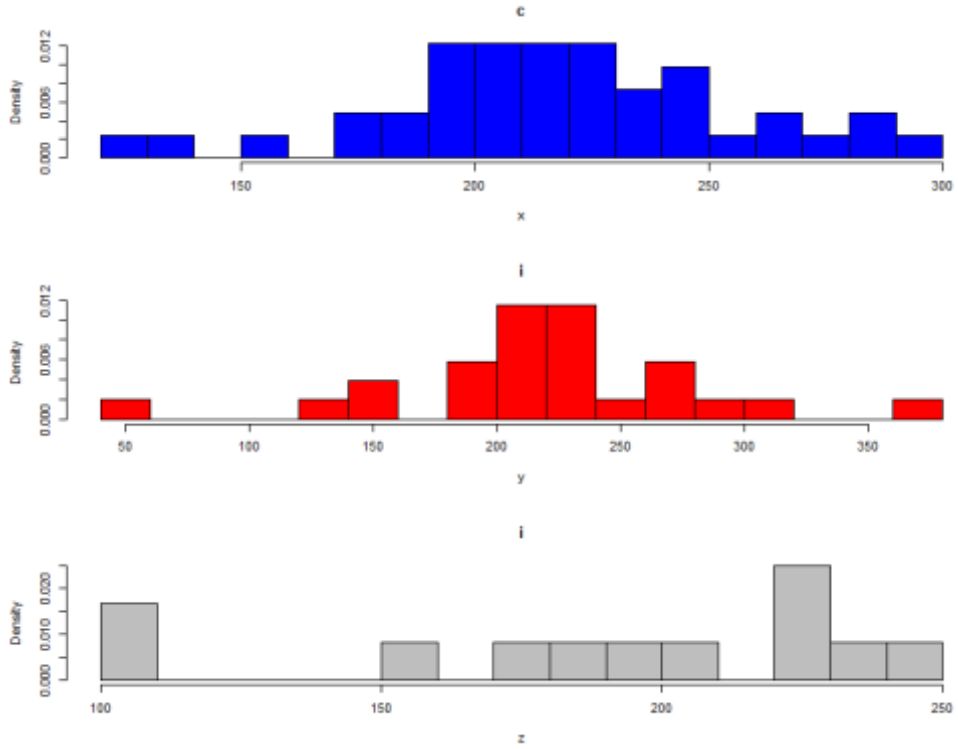


Рис .5. Сравнение гистограмм для трех групп значений.

```
> ggplot(data1, aes(avto, fill = reg)) +
+    geom_histogram(alpha = 0.5, aes(y = ..density..), position =
"dodge", binwidth = 25) +
+   geom_density(alpha = 0.2)
```

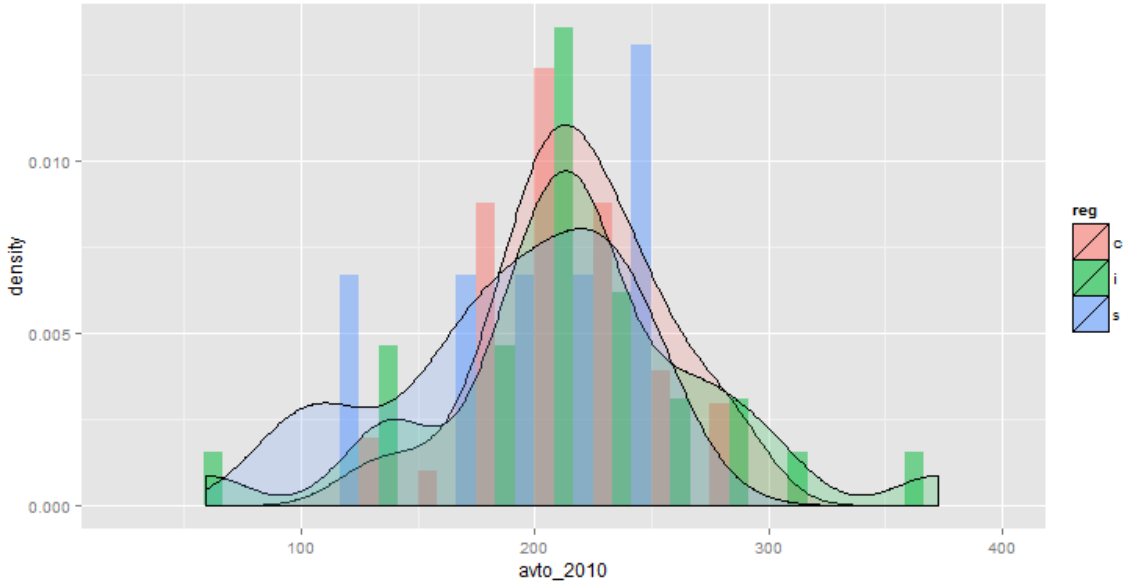


Рис. 6. Сравнение плотностей распределения для трех групп значений.

```
> boxplot(avto ~ reg,  xlab = "Регион", ylab = "Количество автомо-
билей", col = "coral", data = data1)
```

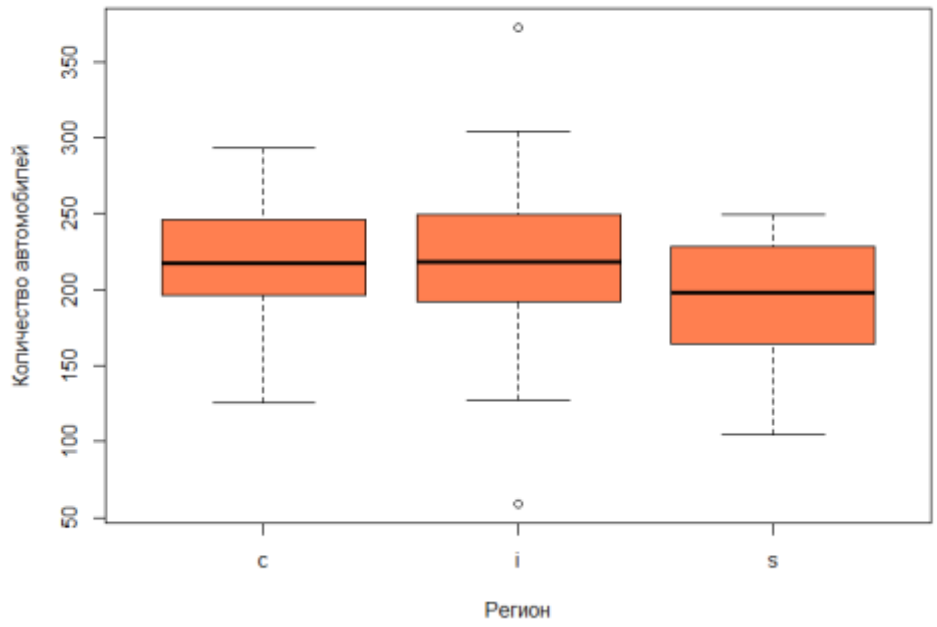


Рис. 7. Диаграммы размаха для трех групп значений.

По результатам применения критерия гипотеза об однородности значений «количество автомобилей» для трех групп регионов принимается на уровне значимости 0.05 ($p > \alpha$).

3.4. Проведение корреляционного анализа

Проводится корреляционный анализ взаимосвязи между показателями: количество автомобилей на 100 тыс. чел. (“avto”), цена за кв.м. жилья на первичном рынке (“house”), уровень доходов населения (руб. в мес, “expenses”), число занятых в экономике (тыс. чел.,“occupied”), стоимость товаров фиксированного набора (руб., “goods”). Для указанных выборочных значений:

a) построить статистическую оценку корреляционной матрицы. Для пары признаков с наибольшим коэффициентом корреляции – построить график типа «диаграмма рассеивания».
b) для пары параметров, имеющих наибольший коэффициент корреляции Пирсона, проверить гипотезу о независимости признаков с использованием коэффициентов ранговой корреляции Спирмена и Кэндела, критерия χ^2.

Таблица 5. Пакеты и функции в R для корреляционного анализа

Функция	Описание основных параметров	Назначение
Пакет **stats**		
cor(x,y, method,...)	**x** — матрица значений многомерной случайной величины или вектор значений СВ X; **y** — вектор значений СВ Y, если вычисляем корреляционную матрицу **y = NULL**; **use –** параметр, указывающий способ работы с пропущенными данными; **method** — метод вычисления коэффициента корреляции: **“pearson”, “spearman”, “kendall”**	Вычисление корреляции между двумя векторами или корреляционной матрицы для элементов data.frame
cor.test(...)	**x, y** — выборки; **alternative** — тип гипотезы, **"two.sided"** — независимость, **"greater"** — проверка на положительную взаимосвязь, **"less"** — на отрицательную взаимосвязь, **method** — метод вычисления коэффициента корреляции; **conf.level** — уровень доверия	Проверка гипотезы о независимости признаков

Функция	Описание основных параметров	Назначение
chisq.test(x, y)	**x, y** — вектора, содержащие значения признаков для объектов	Критерий Пирсона χ^2, для проверки независимости признаков

Далее приведены команды языка R и результаты их выполнения.

а) Вычисление корреляционной матрицы и проверка гипотезы о независимости с использованием коэффициента корреляции Пирсона.

```
> data1   <- read.csv2("data1.csv", sep = ";", header = TRUE,
row.names = 1)
> d1 <- data1[,1:5]
> cor(d1,y = NULL,"pearson",use = "complete.obs" )
              avto     house  expenses       goods    occupied
avto     1.0000000 0.5098676 0.4198672  0.44922077  0.27847915
house    0.5098676 1.0000000 0.6486293  0.50964737  0.45578847
expenses 0.4198672 0.6486293 1.0000000  0.78434289  0.22935814
goods    0.4492208 0.5096474 0.7843429  1.00000000 -0.08802376
occupied 0.2784792 0.4557885 0.2293581 -0.08802376  1.00000000

> all_res <- matrix(ncol = length(d1), nrow = length(d1))
> for (i in 1:length(d1)) {
+    for (j in 1:length(d1)) {
+        x<- d1[,i];   y <- d1[,j]
+        r1 <- cor.test(x, y, method = "pearson",
+     alternative = "two.sided")
+        all_res[i,j] <- r1$p.value
+    }
+}

> round(all_res,5)

[1,] 0.00000 0e+00 0.03597 0.04933 0.00280
[2,] 0.00000 0e+00 0.00000 0.00000 0.00004
[3,] 0.03597 0e+00 0.00000 0.00000 0.16851
[4,] 0.04933 0e+00 0.00000 0.00000 0.26253
[5,] 0.00280 4e-05 0.16851 0.26253 0.00000
```

Результаты проверки гипотезы о независимости признаков подтверждают наличие статистической взаимосвязи для выделенных в первой таблице пар признаков. Наибольший коэффициент корреляции наблюдаются для пары признаков **goods - expenses**. Диаграмма рассеивания представлена далее.

```
> plot(d1$expenses,d1$house, type = "p")
> abline(lm( d1$house ~ d1$expenses))
```

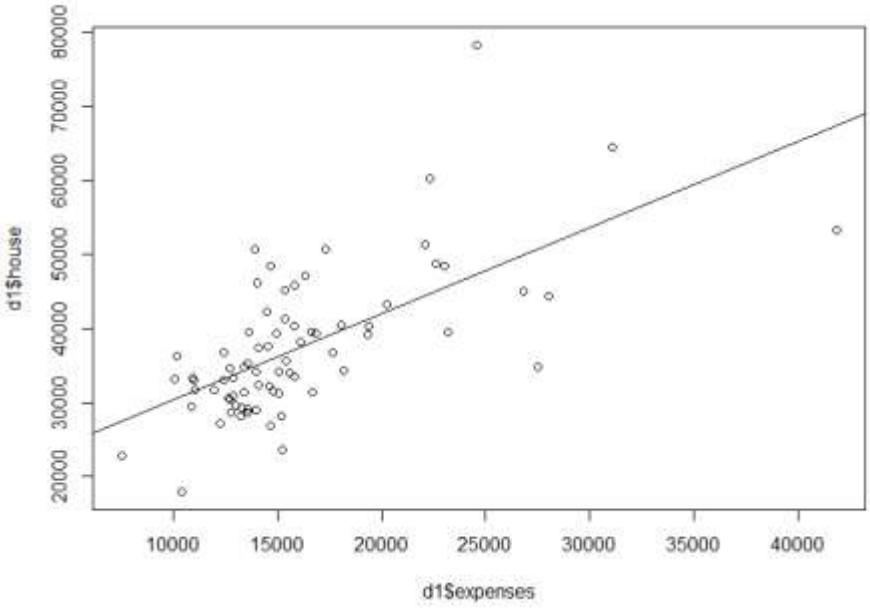


Рис. 8. Диаграмма взаимосвязи признаков.

b) Проверка гипотезы о независимости признаков с использованием коэффициентов ранговой корреляции Спирмена и Кэндела.

Для демонстрации работы указанных ранговых коэффициентов, переведем значения признаков из непрерывной шкалы в ранговую (дискретизация). При разбиении на интервалы используем квантили от 0% до 100% с интервалом 20%. Такое разбиение позволяет создать 5-балльную ранговую шкалу.

```
> discret_x <- function (x)
+   intervals <- quantile(x, probs = seq(0, 1, 0.2),na.rm = TRUE)
+   for (i in 1:length(x))
+   for (k in 1:( length(intervals)-1) )
+     if(!(is.na(x[i])) && (x[i] >= intervals[k]) &&
            (x[i] <= intervals[k+1] ) )  x[i] <- k
+   return(x)
+ }
> d_expenses <- discret_x(d1$expenses)
> d_house <- discret_x(d1$house)
```

Проверка гипотезы о независимости:

```
> cor.test(d_expenses,d_house, method = "spearman", alternative =
"two.sided")

     Spearman's rank correlation rho
data:  d_expenses and d_house
S = 21938.47, p-value = 9.197e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
```

```
0.6879307

> cor.test(d_expenses,d_house, method = "kendall", alternative = "
two.sided")

     Kendall's rank correlation tau

data:  d_expenses and d_house
z = 6.0274, p-value = 1.666e-09
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.5611383
```

c) Проверка гипотезы о независимости признаков с использованием критерия χ^2. Вычисление наблюдаемых и ожидаемых частот

```
> chisq.test(d_expenses, d_house)

     Pearson's Chi-squared test

data:  d_expenses and d_house
X-squared = 49.8631, df = 16, p-value = 2.41e-05

> round(r$expected,1)
          d_house
d_expenses   1   2   3   4   5
         1 2.8 2.8 2.8 2.8 2.8
         2 3.2 3.2 3.2 3.2 3.2
         3 3.0 3.0 3.0 3.0 3.0
         4 3.2 3.2 3.2 3.2 3.2
         5 2.8 2.8 2.8 2.8 2.8

> round(r$residuals,1)
          d_house
d_expenses    1    2    3    4    5
         1  0.7  2.5  0.1 -1.7 -1.7
         2  2.7 -0.1 -0.1 -1.2 -1.2
         3  0.0  0.6  0.0  0.0 -0.6
         4 -1.8 -1.2  1.0  1.6  0.4
         5 -1.7 -1.7 -1.1  1.3  3.1
```

Результаты проверки показывают высокие значения ранговых корреляций для исследуемых признаков $r_s = 0.69$, $\tau = 0.56$. Гипотеза о независимости признаков отвергается в ходе проверки по критерию χ^2.

3.5. Построение линейной регрессионной модели

Для выборочных значений требуется построить линейную регрессию уровня доходов населения (“expenses”) на остальные параметры (“avto”, “house”, “occupied”, “goods”) и проанализировать построенную модель.

Таблица 6. Пакеты и функции в R для корреляционного анализа

Функция	Описание основных параметров	Назначение
Пакет **stats**		
lm(formula, data,…)	**formula** — объект типа формула, описывающий объясняющие и результирующие переменные; **data** — объект типа `data.frame`, выборочные данные для построения модели. Возвращает: **r.squared** — R^2; **fstatistic** — результаты проверки значимости модели; **coefficients** — коэффициенты модели; **fitted.values** — предсказанные значения Y; **residuals** — остатки.	Идентификация параметров линейной регрессионной модели, вычисления основных результатов построения модели
summary(object,…)	**object** — объект, содержащий результаты построения модели, полученный с помощью функции **lm**	Служит для получения результатов проверки значимости коэффициентов модели, исправленного R^2

Далее приведены команды языка R и результаты их выполнения.

```
> data1   <- read.csv2("data1.csv", sep = ";", header = TRUE, row.names = 1)
> d1 <- data1[,1:5]
> fit <- lm(expenses ~ avto + house + goods + occupied, d1)
> s <- summary(fit)
> print(summary(fit))

Residuals:
    Min      1Q  Median      3Q     Max 
-5618.0 -1670.3  -356.9  1710.4 13826.9 
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.079e+04  2.197e+03  -4.911 5.73e-06 ***
avto        -9.635e+00  9.341e+00  -1.031  0.30590
house        1.236e-01  5.094e-02   2.427  0.01780 *
goods        2.655e+00  3.063e-01   8.666 1.08e-12 ***
occupied     1.885e+00  6.893e-01   2.734  0.00791 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2881 on 70 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.7289, Adjusted R-squared:  0.7134
F-statistic: 47.05 on 4 and 70 DF,  p-value: < 2.2e-16
```

В приведенных результатах содержится информация о минимальном и максимальном значении ошибки регрессии, квартилях и медиане (блок `Residuals`).

Приводятся коэффициенты модели с результатами проверки значимости коэффициентов. Для переменной avto при проверке гипотезы о равенстве коэффициента модели нулю -уровень равен 0.30590 (столбец «`Pr(>|t|)`»), гипотеза принимается. Для остальных переменных $p < 0.05$, то есть гипотеза отвергается, коэффициенты значимы.

Следующий блок содержит стандартную ошибку для отклонения регрессии (`Residual standard error`) равную 2881, данные о значении $R^2 = 0.73$ и $R^2_{adj} = 0.71$, результаты проверки гипотезы о значимости функции регрессии. Указанные значения коэффициентов детерминации говорят о достаточно высокой степени адекватности модели.

```
> hist(fit$residuals, xlab='Residuals values', main='Histogram of residulas')
```

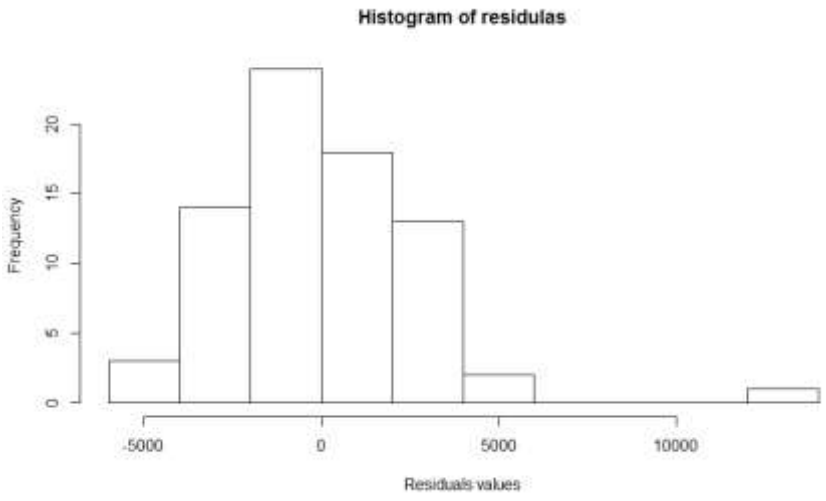


Рис. 9. Гистограмма остатков.

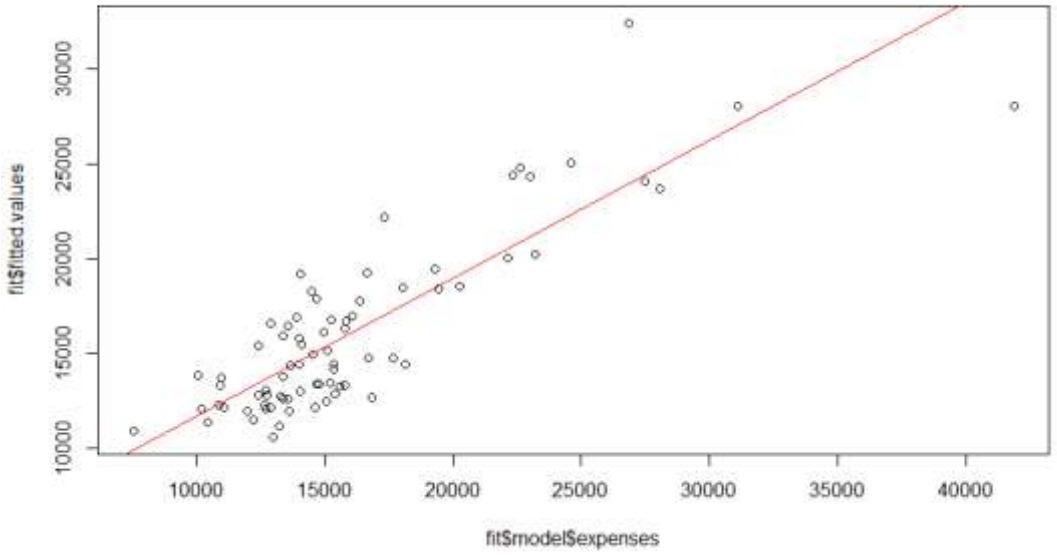


Рис. 10. График наблюдаемое значение – прогнозируемое значение.

По гистограмме (рис. 9) можно установить симметричность распределения, его одномодальность и близость среднего значения к нулю

```
> plot(fit$model$expenses, fit$fitted.values)
> abline(lm(fit$fitted.values ~ fit$model$expenses), col='red')
```

График (рис. 10) позволяет оценить степень приближения прогноза (**fit$model.expenses**) к наблюдаемым значениям (**fit$fitted.values**).

```
> plot(fit$model$expenses, fit$residuals)
> abline(lm(fit$residuals ~ fit$model$expenses), col='red')
```

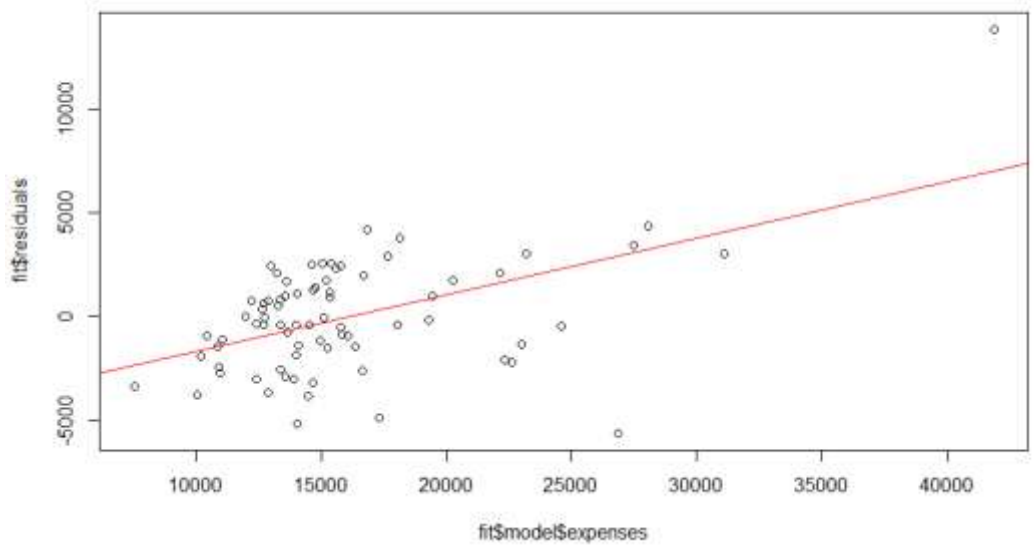


Рис. 11. График остатки – прогнозируемое значение.

График (рис. 11) позволяет оценить независимость остатков от наблюдаемых значений результирующей переменной.

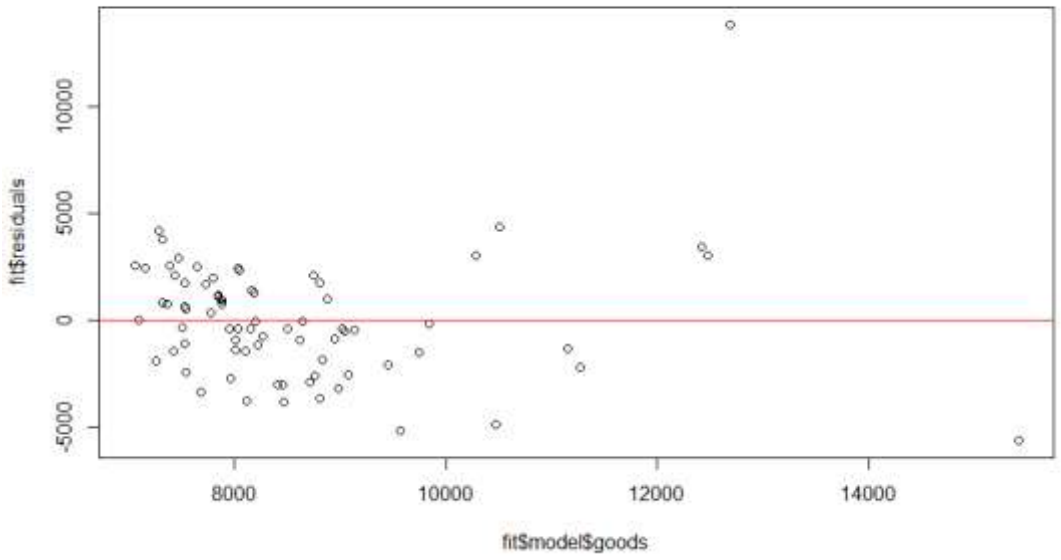


Рис. 12. График остатки – исходный признак “goods”.

Данный график рассеивания (рис. 12) позволяет оценить независимость остатков от значений исходной переменной “goods”. Аналогичные графики следует построить для остальных исходных переменных, чтобы убедиться в независимости остатков.

3.6. Проведение дисперсионного анализа данных

Исходными данными является уровень доходов населения (руб. в мес, “expenses”), число занятых в экономике (тыс. чел.,“occupied”), расположение региона (столбец “reg”). Для указанных выборочных значений:

a) Проверить гипотезу о несущественности влияния фактора A — расположение региона на результирующий фактор Y — уровень доходов населения. Построить модель, определить основные характеристики качества дисперсионной модели.
b) Проверить гипотезу о несущественности влияния двух факторов: расположение региона и число занятых в экономике (ранговое значение) на уровень доходов населения с учетом взаимного влияния признаков и без

учета взаимного влияния. Построить модель, определить основные характеристики качества дисперсионной модели.

Таблица 7. Пакеты и функции в R для дисперсионного анализа

Функция	Описание основных параметров	Назначение
Пакет **stats**		
aov(formula, data = NULL, projections = FALSE, qr = TRUE, contrasts = NULL, ...)	**formula** — формула с указанием модели; **data** — блок данных (таблица), в котором будет осуществляться поиск переменных, указанных в формуле. Возвращает: **Residuals** — внутригрупповая дисперсия; **Sum Sq** — суммы квадратов отклонений; **Mean Sq** — суммы квадратов отклонений с поправкой на степени свободы; **F value** — статистика F-критерия; **Pr(>F)** — *p*-уровень для полученной статистики	Проверка гипотезы о несущественности влияния фактора(ов)
anova(object,...)	**object** — объект, содержащий результаты, возвращенные функцией построения модели (например, lm). Возвращаемые параметры — аналогичны aov()	Дисперсионный анализ модели
model.tables(x, type = "effects", se = FALSE, cterms, …)	**x** — объект модели, полученный функцией aov(); **type** — тип выдаваемой таблицы, "effects" и "means"; **se** — булева переменная, определяющая, должно ли быть посчитано среднеквадратичное отклонение	Таблица со средними значениями элементов

Функция	Описание основных параметров	Назначение
	cterms — вектор с именами переменных. Возвращаемые значения: **tables** — таблица требуемых значений; **n** — количество элементов в группе; **se** (standard error) — средне-квадратическое отклонение	
Пакет **gplots**		
plotmeans(formula, data = NULL, xlab, ylab, main, …)	**formula** — формула с указанием модели; **data** — таблица с данными, используемыми в формуле	Графическое отображение средних с доверительными интервалами

Далее приведены команды языка R и результаты их выполнения.

а) Однофакторный ДА.

```
> res <- aov(expenses ~ reg, data = d1)
> summary(res)
             Df    Sum Sq   Mean Sq F value   Pr(>F)
reg           2 4.654e+08 232695594    7.82 0.000816 ***
Residuals    76 2.262e+09  29757888
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Гипотеза о несущественности влияния фактора на Y отвергается ($p < 0.01$).

```
> plotmeans(expenses ~ reg, data = d1)
```

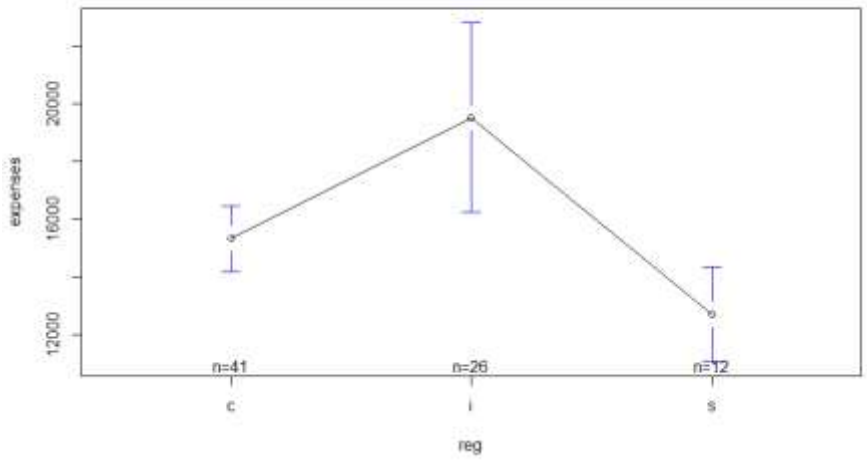


Рис. 13. Групповые средние с доверительными интервалами.

Вычисляем средние значения в группах по уровням фактора “reg”:

```
> model.tables(res, type = "means")
Tables of means
Grand mean 16294.85

 reg
        c     i     s
    15308 19515 12691
rep    41    26    12
```

Проверяем значение коэффициента детерминации для регрессионной модели:

```
> summary(lm(expenses ~ reg, data = d1))

Residuals:
   Min     1Q Median     3Q    Max
 -9465  -3221  -1383   2160  22350

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15307.9      851.9  17.968  < 2e-16 ***
regi          4206.8     1367.6   3.076  0.00291 **
regs         -2617.4     1790.4  -1.462  0.14789
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5455 on 76 degrees of freedom
Multiple R-squared:  0.1707,  Adjusted R-squared:  0.1488
F-statistic:  7.82 on 2 and 76 DF,  p-value: 0.0008162
```

Выявленное влияние фактора “reg” на уровень доходов населения “expenses” не позволяет построить регрессионную модель предсказывающую значение уровня доходов с достаточной точностью ($R^2 = 0.17$).

b) Двухфакторный ДА без учета зависимости факторов:

Используем функцию discret_x для дискретизации значения числа занятых в экономике “ occupied” (делим на 4 группы квантилями уровня $0.25, 0.5, 0.75$):

```
> d_occupied <- discret_x(d1$occupied)
> d1 <- cbind(d1, d_occupied)
```

Далее проводим двухфакторный ДА:

```
> res <- aov(expenses ~ reg + d_occupied , data = d1)
> summary(res)
```

```
             Df    Sum Sq   Mean Sq F value   Pr(>F)
reg           2 4.654e+08 232695594   7.720 0.000909 ***
d_occupied    3 6.111e+07  20370859   0.676 0.331898
Residuals    73 2.200e+09  30143657
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

Гипотеза о несущественности влияния фактора “reg” на Y отвергается ($p < 0.01$), фактора ”d_occupied” на Y — принимается ($p = 0.33$).

```
> plotmeans(expenses ~ d_occupied, data = d1)
```

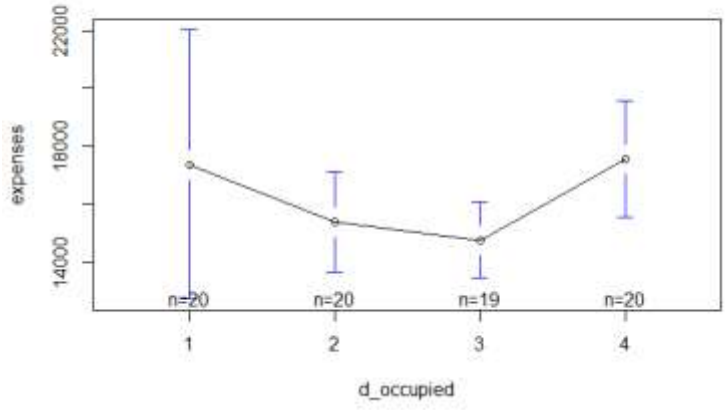


Рис. 14. Групповые средние с доверительными интервалами.

Вычисляем средние значения в группах по уровням факторов “reg” и “d_occupied”:

```
> model.tables(res, type = "means")
Tables of means
Grand mean          16294.85

 reg
        c     i     s
    15308 19515 12691
rep    41    26    12

 d_occupied
        1     2     3     4
    17249 15256 15228 17392
rep    20    20    19    20


> library(ggplot2)
> ggplot(data = d1, aes(x = d_occupied, y = expenses, colour =
reg))+ geom_boxplot()
```

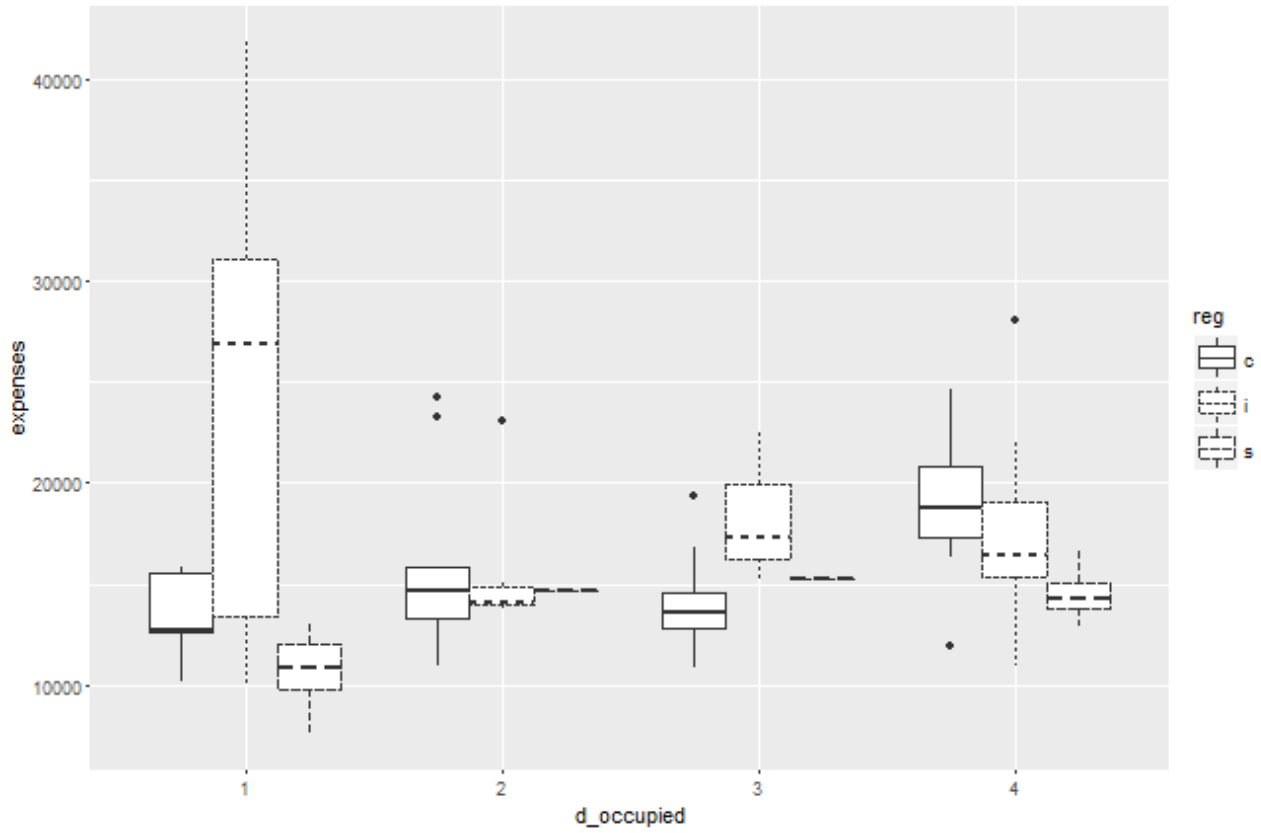


Рис. 15. Диаграммы размаха значений признака «уровень доходов» с группировкой по уровням занятости и регионам.

На диаграмме можно увидеть различия признака “expences” для регионов на каждом уровне фактора “d_occupied” при том, что различия между уровнями “d_occupied” незначительны.

c) Двухфакторный дисперсионный анализ с учетом зависимости факторов:

```
> res <- aov(expenses ~ reg + d_occupied + reg * d_occupied, data = d1)
> summary(res)
                Df    Sum Sq   Mean Sq F value   Pr(>F)
reg              2 4.654e+08 232695594   8.806 0.000402 ***
d_occupied       3 1.027e+08  34231666   1.295 0.283220
reg:d_occupied   6 3.885e+08  64746205   2.450 0.033449 *
Residuals       67 1.770e+09  26424288
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Гипотеза о несущественности влияния фактора “reg” на Y отвергается ($p < 0.01$), влияния фактора ”d_occupied” на Y – принимается ($p = 0.28$), совместного влияния факторов – отвергается на уровне 0.05 ($p = 0.0334$)

```
> model.tables(res, type = "means")
```

```
Tables of means
Grand mean          16294.85

 reg:d_occupied
     d_occupied
reg   1     2     3     4
  c   13396 15347 14023 18847
  rep     5    13    15     8
  i   24019 15642 18341 17792
  rep     9     6     3     8
  s   10708 14641 15213 14547
  rep     6     1     1     4
```

Таблица средних позволяет сравнить групповые средние в группах, сформированных уровнями двух факторов.

3.7. Проведение компонентного анализа данных

Провести компонентный анализ исходных признаков: количество автомобилей на 100 тыс. чел. (“avto”), цена за кв.м. жилья на первичном рынке (руб.,“house”), уровень доходов населения (руб. в мес, “expenses”), число занятых в экономике (тыс. чел.,“occupied”), стоимость товаров фиксированного набора (руб., “goods”). Определить число компонент, достаточное для описания исходных признаков. Вычислить факторные нагрузки признаков. Определить вклады факторов в суммарную дисперсию. Интерпретировать найденные главные компоненты с использованием факторных нагрузок и их квадратов.

Таблица 8. Пакеты и функции в R для компонентного анализа

Функция	Описание основных параметров	Назначение
Пакет **psych**		
principal(r, nfactors, rotate, n.obs = NA, covar = FALSE, scores = TRUE, method = “regression, …)	**r** — корреляционная матрица или исходная матрица; **nfactors** — количество компонент, по умолчанию 1; **rotate** — поворот, для проведения компонентного анализа "none"; **covar** — если FALSE то работаем с корреляционной матрицей, иначе — с ковариационной; **scores** — найти значения компонент для наблюдений, по умолчанию TRUE;	Анализ главных компонент с возможностью поворота осей

Функция	Описание основных параметров	Назначение
	method — метод поиска компонент, по умолчанию “regression”. Возвращает значения: **values** — собственные значения всех компонент; **n.obs** — количество наблюдений; **loadings** — матрица факторных нагрузок; **weights** — веса исходных признаков в компонентах (нормированные нагрузки); **communality** — дисперсия признаков, распределенная по компонентам; **scores** — оценки значений компонент для наблюдений;	
fa.diagram(fa.results, sort = TRUE, labels = NULL,...)	**fa.results** — результат, полученный функцией **principal**	Графическое изображение матриц нагрузок факторного анализа или анализа главных компонент
factor.plot (fa.results, cluster = NULL, cut = 0, labels = NULL, title, jiggle = FALSE, amount = .02, pch = 18, pos, . . .)	**fa.results** — результат, полученный функцией **principal**	Графическое изображение результатов анализа главных компонент
scree(rx, factors = TRUE, pc = TRUE, main = “Scree plot”, hline = NULL, add = FALSE)	**rx** — корреляционная матрица или матрица исходных данных; **factors** — показать график для факторного анализа, по умолчанию TRUE; **pc** — показать график для анализа главных компонент, по умолчанию TRUE; **main** — заголовок, по умолчанию “Scree plot”; **hline** — добавить горизонтальную	График собственных значений для факторного анализа и для анализа главных компонент

Функция	Описание основных параметров	Назначение
	линию в единице, если NULL, иначе, добавить гор.линию в указанном значении, по умолчанию NULL; **add** — добавить другой график, по умолчанию FALSE.	

Часть результатов, возвращаемых функцией **principal** относятся к процедуре факторного анализа. Этот метод позволят путем вращения векторов, соответствующих главным компонентам, находить компоненты с высокой дисперсией, при этом пренебрегая некоррелированностью компонент. Так, если в ходе проведения компонентного анализа выделяются две компоненты, берущие на себя более 90% суммарной дисперсии, то имеет смысл скорректировать линейные комбинации, определяющие компоненты так, чтобы на них распределились и оставшиеся 10% дисперсии исходных признаков. Методы факторного анализа описаны в [8], [9] и реализованы в программных системах, предназначенных для анализа данных.

Далее приведены команды языка R и результаты их выполнения.

```
> install.packages("psych")
> library(psych)
> data1  <- read.csv2("data1.csv", sep = ";", dec=",", header = TR
UE, row.names = 1, encoding = "UTF-8")
> d1 <- data1[, 1:5]
> pr <- principal(d1, nfactors = 5,scores=TRUE, rotate = "none")
> pr
Principal Components Analysis
Call: principal(r = d1, nfactors = 5, rotate = "none", scores =
TRUE)
Standardized loadings (pattern matrix) based upon correlation ma-
trix
           PC1   PC2   PC3   PC4   PC5 h2       u2 com
avto      0.59  0.46 -0.66  0.11 -0.04  1 2.2e-16 2.9
house     0.88  0.20  0.08 -0.42  0.02  1 2.7e-15 1.6
expenses  0.87 -0.36  0.21  0.14 -0.23  1 1.4e-15 1.7
goods     0.77 -0.58 -0.04  0.13  0.23  1 1.4e-15 2.1
occupied  0.39  0.79  0.42  0.20  0.08  1 6.7e-16 2.3

                       PC1  PC2  PC3  PC4  PC5
SS loadings           2.63 1.33 0.66 0.26 0.11
Proportion Var        0.53 0.27 0.13 0.05 0.02
Cumulative Var        0.53 0.79 0.92 0.98 1.00
Proportion Explained  0.53 0.27 0.13 0.05 0.02
```

```
Cumulative Proportion 0.53 0.79 0.92 0.98 1.00

Mean item complexity =  2.1
Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is  0
 with the empirical chi square  0  with prob <  NA

Fit based upon off diagonal values = 1
```

Указанные результаты содержат: матрицу факторных нагрузок (Standardized loadings, h2 — суммарная дисперсия признака в компонентах, u2 — остаточная дисперсия), суммы квадратов факторных нагрузок (SS loadings), доли суммарной дисперсии, приходящиеся на компоненту (Proportion Var), накопленную долю дисперсии (Cumulative Var). Показатели Proportion Explained, Cumulative Proportion и прочие — относятся к методам факторного анализа. Значения SS loadings показывают, что на первые 2 компоненты приходится почти 80% суммарной дисперсии исходных признаков.

График значений собственных чисел также указывает на высокие значения дисперсий первых двух компонент:

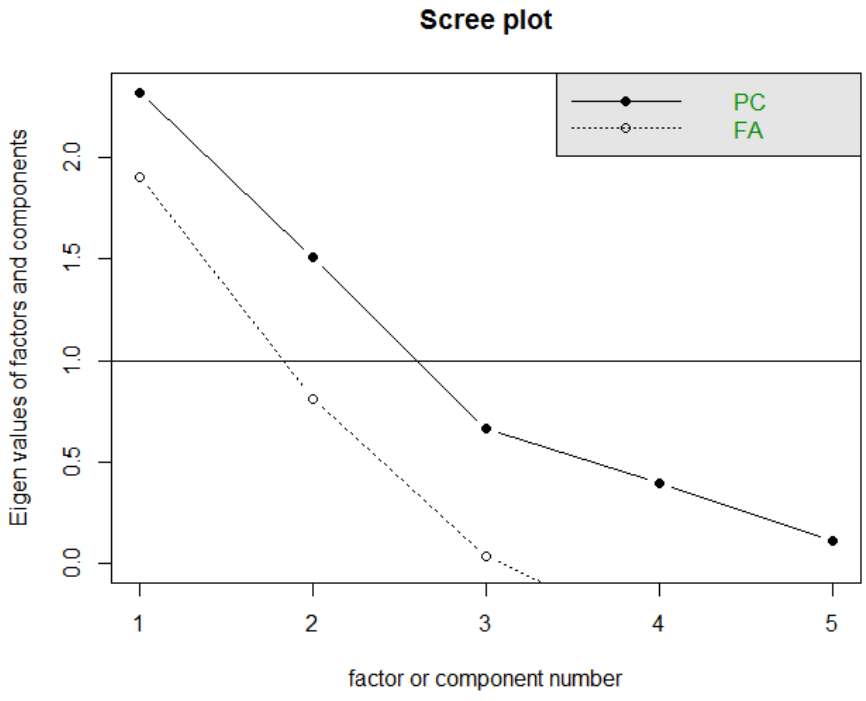


Рис. 16. График значений собственных чисел.

Для интерпретации полученных компонент можно использовать матрицу квадратов факторных нагрузок а также графическое изображение результатов анализа главных компонент.

```
> pr$loadings^2
```

```
Loadings:
         PC1   PC2   PC3   PC4   PC5
avto     0.477 0.128 0.290 0.103
house    0.559 0.185       0.246
expenses 0.613 0.268
goods    0.449 0.495
occupied 0.222 0.430 0.297
```

```
> fa.diagram(pr)
```

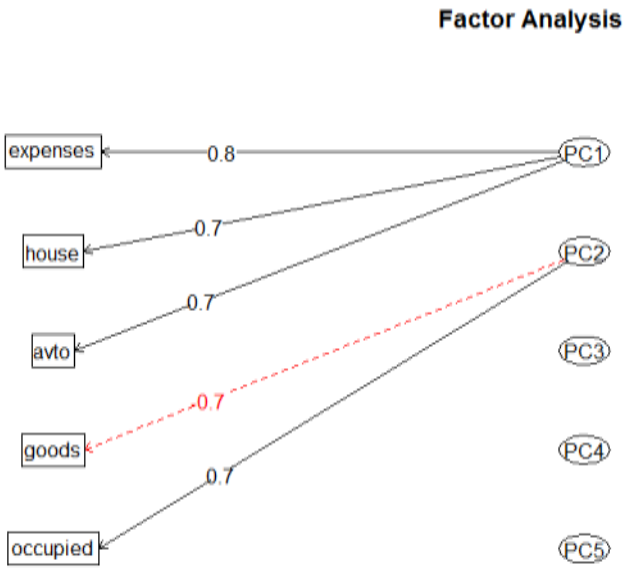


Рис. 17. Диаграмма влияния исходных признаков на компоненты.

На диаграмме видно, что наибольший вклад в первую компоненту вносят признаки: количество автомобилей на 100 тыс. чел. (“avto”), цены за кв.м. жилья на первичном рынке (“house”), уровень доходов населения (“expenses”), все переменные входят в линейную комбинацию с положительными знаками. Поскольку все три признака уровень доходов и потенциальных расходов, можно назвать компоненту «уровень благосостояния населения в регионе»

Во вторую компоненту наибольший вклад вносят признаки: число занятых в экономике (“occupied”), стоимость товаров фиксированного набора (руб.,

“goods”), с противоположными знаками. Это означает, что при высоких показателях занятости, цены на товары фиксированного набора низкие и наоборот. Данная компонента может говорить о том, что в регионе развита экономика, есть возможность привлечения дополнительных трудовых ресурсов. Если для региона при низких значениях первой компоненты наблюдается высокое значение второй, то данный регион можно охарактеризовать как регион с развитой экономикой, большим количеством предприятий, уровнем цен, обеспечивающим достаточно комфортное проживание.

Для оценки результатов можно также вывести таблицу значений компонент для регионов. Далее представлен фрагмент таблицы, содержащий регионы с крайними значениями компонент.

```
> pr$scores
                                        PC1          PC2
Московская область                2.309619291  1.896209329
Ямало-Ненецкий автономный округ   2.520177989 -2.461972789
Чукотский автономный округ        0.270673337 -5.226016770
Республика Ингушетия             -2.459797932 -1.035579348
Ростовская область                0.491286347  1.181949391
```

Полученные результаты позволяют классифицировать регионы по группам со схожими значениями компонент, например, методами кластерного анализа [11].

Заключение

В предложенном пособии изложены основные теоретические сведения, необходимые для проведения исследования данных базовыми методами статистического анализа. Также приводятся примеры расчетов, выполненные с использованием языка R. Следует обратить внимание на то, что полученные результаты могут быть получены в R разными способами, не обязательно при помощи библиотек и функций, описанных в данном пособии.

При решении практических задач статистическими методами наибольшую важность играет правильность выбора метода и последующая интерпретация полученных результатов, а выбор инструмента расчета важен только с позиции обеспечения требуемой функциональности и производительности.

Список использованных источников

1. Мастицкий, С. Э. Статистический анализ и визуализация данных с помощью R / С. Э. Мастицкий, В. К. Шитиков. — Москва : ДМК Пресс, 2014. — 496 с. : ил.
2. Вентцель, Е. С. Теория вероятностей : учебник для вузов / Е. С. Вентцель. — Москва : Высшая школа, 1999.— 576 с. : ил.
3. Гмурман, В. Е. Теория вероятностей и математическая статистика : учебное пособие для вузов / В. Е. Гмурман. — Изд. 9-е, стереотипное. — Москва : Высшая школа, 2003. — 479 с. : ил.
4. Айвазян, С. А. Прикладная статистика. Основы эконометрики: учебник для вузов / С. А. Айвазян, В. С. Мхитарян. В 2 т. — Москва : Юнити-Дана. — 2001.
5. Чернецкий, В. И. Математическое моделирование стохастических систем / В. И. Чернецкий. — Петрозаводск: Издательство ПетрГУ, 1994. — 488 с. : ил.
6. Рогов, А. А. Проверка статистических гипотез : учебно-методическое пособие / А. А. Рогов, А. В. Воронин, С. Т. Коржов. — Петрозаводск: Издательство ПетрГУ, 2005. — 38 с.
7. Питухин, Е. А. Введение в анализ временных рядов. Основные понятия и методы / Е. А. Питухин, И. М. Шабалина. — Петрозаводск : Издательство ПетрГУ. 2012. — 42 с.
8. Андерсон, Т. Введение в многомерный статистический анализ / Т. Андерсон. — Москва : Физматгиз, 1963. — 479 с.
9. Харман, Г. Современный факторный анализ / Г. Харман. — Москва : Статистика, 1972. — 489 с.
10. Большев, Л. Н. Таблицы математической статистики. / Л.Н. Большев, Н. В. Смирнов — Москва : Наука. 1965. — 416 с.
11. Мандель И.Д. Кластерный анализ / И.Д. Мандель. — Москва : Финансы и статистика, 1988. — 176 с.

Оглавление

Приложение 1. Содержимое файла data1.csv

name	avto	house	expenses	goods	occupied	reg
Белгородская область	213,21	39340	16839	7286,7	693,5	c
Брянская область	125,14	29373	13298	7542,6	571,6	c
Владимирская область	206,44	36778	12424	8404,9	703,6	c
Воронежская область	232,48	35219	13580	8711	1054,3	c
Ивановская область	175,8	33100	10980	7969	490,2	c
Калужская область	229,43	45207	15342	7877,9	480,2	c
Костромская область	205,54	30649	12656	7781,7	321,5	c
Курская область	203,57	26917	14694	8189,6	573,9	c
Липецкая область	245,6	33595	15804	8028,8	544,9	c
Московская область	293,32	60233	22324	9455,3	2901,1	c
Орловская область	213,16	29683	13017	7157,3	391,9	c
Рязанская область	272,44	39561	13663	8271,8	502,8	c
Смоленская область	246,03	31597	14770	8165,2	495,8	c
Тамбовская область	209,34	28580	13592	7870,9	503,6	c
Тверская область	218,72	50688	13925	8458,2	588,8	c
Тульская область	252,79	41216	15358	7850,6	771,1	c
Ярославская область	184,18	37551	14548	8150,7	643,9	c
Республика Карелия	263,51	45787	15851	8950,3	336,7	c
Республика Коми	214,16	39542	23220	10284,6	467,5	c
Архангельская область	192,88	39238	19310	9838,7	607,7	c
Вологодская область	228,83	34213	13999	8836,1	598,1	c
Калининградская область	283,54	40278	15808	9039,8	471,4	c
Ленинградская область	266,58	48473	14673	8979,3	741,1	c
Мурманская область	247,69		24274	11062,9	434,8	c
Новгородская область	217,69	33969	15582	8050,2	315	c
Псковская область	221,76	34651	12698	7954	325,8	c
г. Санкт- Петербург	282,89	78243	24594	9133,4	2466,3	c
Республика Адыгея	239,89	27144	12236	7883,3	152,4	s
Республика Калмыкия	180,85	22895	7540	7684,6	114	s
Краснодарский край	249,13	39592	16648	8757,4	2274,2	s
Астраханская область	229,41	32240	14641	7648,4	447,7	s
Волгоградская область	200,81	37386	14122	8011,3	1229,7	s
Ростовская область	227,34	42337	14503	8463	1895,7	s
Республика Дагестан	104,17	28159	15213	7531	949	s
Республика Ингушетия	105,3		9596	7170,7	68,3	s
Кабардино-Балкарская	156,48		11215	7220,9	309,9	s
Карачаево-Черкесская	172,76	18000	10431	8013,5	170,6	s
Республика Северная Осетия - Алания	195,02	28245	13228	7437,3	299,3	s
Ставропольский край	223,62	30931	12913	8803,2	1236,5	s
Республика Башкортостан	223,59	36811	17677	7467,6	1770,6	c
Республика Марий Эл	155,2	36348	10195	7261,8	318,1	c
Республика Мордовия	176,04	31681	11055	7527,3	385	c
Республика Татарстан	197,43	34315	18158	7320,1	1810,5	c
Удмуртская Республика	194,56	33094	12423	7504,5	759,2	c
Чувашская Республика	139,09	29438	10885	7425,3	574,6	c

name	avto	house	expenses	goods	occupied	reg
Пермский край	187,98	40312	19422	8882,3	1304,8	c
Кировская область	195,31	31386	13385	8032,8	664,2	c
Нижегородская область	207,64	47062	16358	8107,3	1710,9	c
Оренбургская область	247,27	34883	13398	7314	1070,9	c
Пензенская область	223	30357	12700	7531,5	667,3	c
Самарская область	236,38	43220	20279	8808,9	1509,4	c
Саратовская область	239,71	31679	11961	7091,4	1209	c
Ульяновская область	196,21	33330	12905	7371	602,6	c
Курганская область	226,06	29167	13601	7728,1	409	i
Свердловская область	275,44	51341	22128	8747,9	2064,1	i
Тюменская область	260,86	44306	28049	10507,5	1928,4	i
Ямало-Ненецкий автономный округ	220,77	53294	41865	12684,3	367,8	i
Челябинская область	228,55	31375	16714	7796,9	1665,7	i
Республика Алтай	192,39	34758	13399	9073,6	93,9	i
Республика Бурятия	182,68	28933	13998	8507,1	417,1	i
Республика Тыва	127,41	33175	10050	8115,7	106,1	i
Республика Хакасия	223,72	28718	12776	8196,6	243	i
Алтайский край	215,31	33386	10926	7543	1079,4	i
Забайкальский край	208,92	32425	14070	7847,2	490,1	i
Красноярский край	249,45	40491	18047	9015,3	1439,3	i
Иркутская область	202,6	39420	14965	8220,2	1140,2	i
Кемеровская область	201,69	35641	15416	7057,8	1294,7	i
Новосибирская область	232,15	38263	16090	8614,8	1286,6	i
Омская область	207,32	31271	15070	7391,8	944,6	i
Томская область	220,89	34084	15098	8641,3	491,9	i
Республика Саха (Якутия)	142,59	48481	23024	11154,3	481,1	i
Камчатский край	372,33	44949	26841	15415,1	189,1	i
Приморский край	304,22	50665	17347	10472,4	980,1	i
Хабаровский край	183,44	48786	22607	11271,1	729,4	i
Амурская область	201,79	46162	14064	9564,8	437,9	i
Магаданская область	270,01	34851	27489	12417,2	89,9	i
Сахалинская область	291,18	64500	31078	12472,8	288,7	i
Еврейская автономная область	142,46	23615	15249	9740,7	81,3	i
Чукотский автономный округ	59,13		37422	15457,4	35,9	i