Министерство образования и науки Российской Федерации
ФГАОУ ВО «Севастопольский государственный университет»

Институт информационных технологий
и управления в технических системах

Лабораторная работа №2
«Прогнозирование. Линейная однофакторная регрессионная модель. Множественная линейная регрессия.»

по дисциплине «Интеллектуальный анализ данных»
для студентов всех форм обучения направления подготовки
09.04.02 «Информационные системы и технологии», 09.04.03 «Прикладная информатика»

Севастополь
2019

Прогнозирование. Линейная однофакторная регрессионная модель. Множественная линейная регрессия. Методические указания к лабораторным занятиям по дисциплине «Анализ данных» / Сост.: И.П. Шумейко, О.А. Сырых – Севастополь: Изд-во СевГУ, 2018 – 20 с.

Методические указания предназначены для проведения лабораторных работ по дисциплине «Анализ данных». Целью методических указаний является помощь студентам в изучении возможностей системы RStudio. Излагаются практические сведения необходимые для выполнения лабораторной работы, требования к содержанию отчета.

Лабораторная работа № 2
Прогнозирование. Линейная однофакторная регрессионная модель. Множественная линейная регрессия.

Цель:

– исследовать возможности языка R для построения линейной однофакторной регрессионной модели и множественной линейной регрессий;

Время: 4 часа

Лабораторное оборудование: персональные компьютеры, выход в сеть Internet, RStudio.

Краткие теоретические сведения

Однофакторная линейная регрессия. Постановка задачи.

В самых разных областях знания возникает задача определения зависимости между случайными величинами, являющимися признаками одних и тех же объектов. Например, это может быть зависимость

– между ростом и весом человека;

– между силой сигнала на входе и выходе технического устройства;

– между затратами компании на рекламу и доходом от продаж;

– между уровнем инфляции и безработицей;

– между содержанием радиоактивного вещества в растениях-медоносах и в мёде, полученном от этих растений…

Этот список можно продолжать очень долго. Конечно, на практике на значение исследуемой величины влияет множество факторов, но для простоты мы будем считать, что основное влияние оказывает один из них, потому и анализ будем называть однофакторным. Будем считать, что оба признака, зависимость между которыми мы стараемся выявить, представимы как значения вещественных переменных. Предположим, что нам известны результаты n измерений. Каждое измерение i (i =1,…, n) даёт пару чисел (x_i, y_i) – значения двух признаков измеряемого объекта, (например, рост – вес, затраты на рекламу – доход и т.д.), т.е. сырые данные представимы как таблица:

№ наблюдения, i	Значения фактора, x_i	Значения переменной отклика, y_i
1	x_1	y_1
…		
i	x_i	y_i
…		
n	x_n	y_n

Здесь каждая строка соответствует одному объекту (наблюдению). Признак, который может быть непосредственно измерен (x), является фактором (предиктором), прогнозируемая переменная (y) – переменная отклика. Цель исследования – построить (линейную) функцию (регрессионную модель), которая позволит прогнозировать значение переменной отклика (y) по известному значению фактора (x).

Визуализация сырых данных

Построим систему координат, где по оси абсцисс будем откладывать значения фактора (x), по оси ординат – значения переменной отклика (y). Таким образом, каждому наблюдению (т.е. каждой паре) (x_i, y_i) (i =1,…, n) соответствует точка на координатной плоскости. Если зависимость между изучаемыми признаками была бы линейной и отсутствовала бы случайная компонента, то все эти точки лежали бы на одной прямой. Однако из-за наличия случайного «шума» точки оказываются разбросанными по координатной плоскости в виде так называемого «облака». Пример такого облака показан на приведённом ниже рисунке 1 (построенном в пакете R):

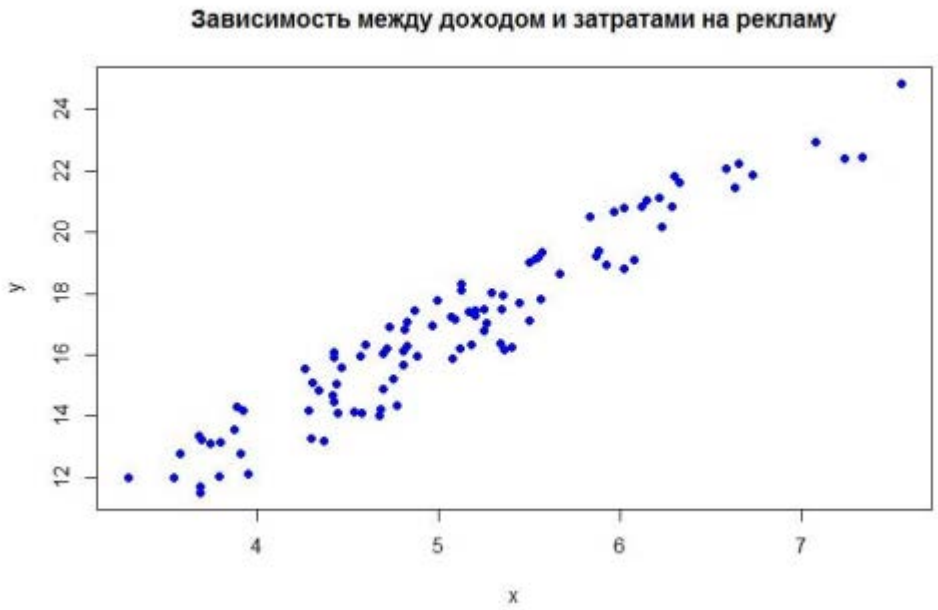


Рисунок 1.

Здесь ось абсцисс соответствует затратам на рекламу, ось ординат – объёму продаж (зафиксированному через заданное время после проведения рекламной кампании). Поставим задачу: найти такую линейную функцию, которая наилучшим образом отражает зависимость переменной отклика y (объёма продаж) от фактора x (затрат на рекламу). Эта задача называется задачей однофакторной линейной регрессии. Приведём математическую формулировку задачи.

Математическая постановка задачи нахождения уравнения регрессии.

Как известно, линейная функция одной переменной x имеет вид:

$$y = \beta_0 + \beta x \text{ , (1)}$$

где β – тангенс угла наклона графика этой функции к оси OX,

β_0 – ордината точки пересечения этой прямой с осью OY.

Задача состоит в том, чтобы найти такие значения переменных β_0, β , при которых прямая (1) наилучшим образом проходит через облако точек (x_i, y_i), i =1,…, n. Поясним смысл задачи геометрически. Зафиксируем произвольные значения β_0 и β и построим соответствующую прямую (рис 2):

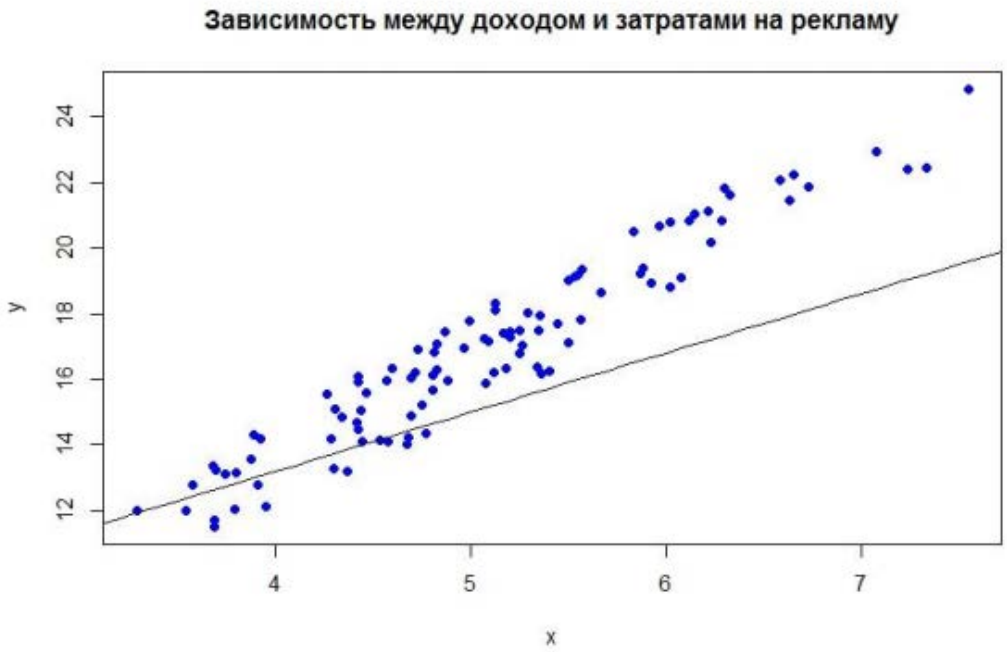


Рисунок 2.

Очевидно, построенная «наугад» прямая является не самой лучшей для данного облака точек. Формализуем понятие «качества» модели. При фиксированных β_0 и β «ожидаемое» (согласно (1)) значение y при $x = x_i$ составляет $\beta_0 + \beta x_i$, $i = 1,\ldots, n$, (т.е. точка $(x_i, \beta_0 + \beta x_i)$ лежит на построенной прямой). Но фактическое значение переменной y при $x = x_i$ составляет y_i , т.е. «ошибка» составляет $\beta_0 + \beta x_i - y$. Поставим задачу: найти значения β_0 и β, минимизирующие сумму квадратов ошибок, т.е.:

$$\sum_{i=1}^{n}\left(\beta_0 + \beta x_i - y_i\right)^2 \to \min \tag{2}$$

Метод наименьших квадратов

Принцип поиска коэффициентов регрессии путём минимизации суммы квадратов отклонений между реальными значениями признака и прогнозируемыми согласно предполагаемой форме зависимости (в нашем случае – линейной) называется методом наименьших квадратов (англ.: Least Square Method, LSM). Проиллюстрируем целевую функцию задачи (2) на рисунке 3.

Значение целевой функции задачи (2) при фиксированных значениях β_0 и β равно сумме квадратов длин построенных отрезков. Из рисунка видно, что построенная прямая – не лучшая, так как можно провести прямую, обеспечивающую меньшее значение целевой функции задачи (2).

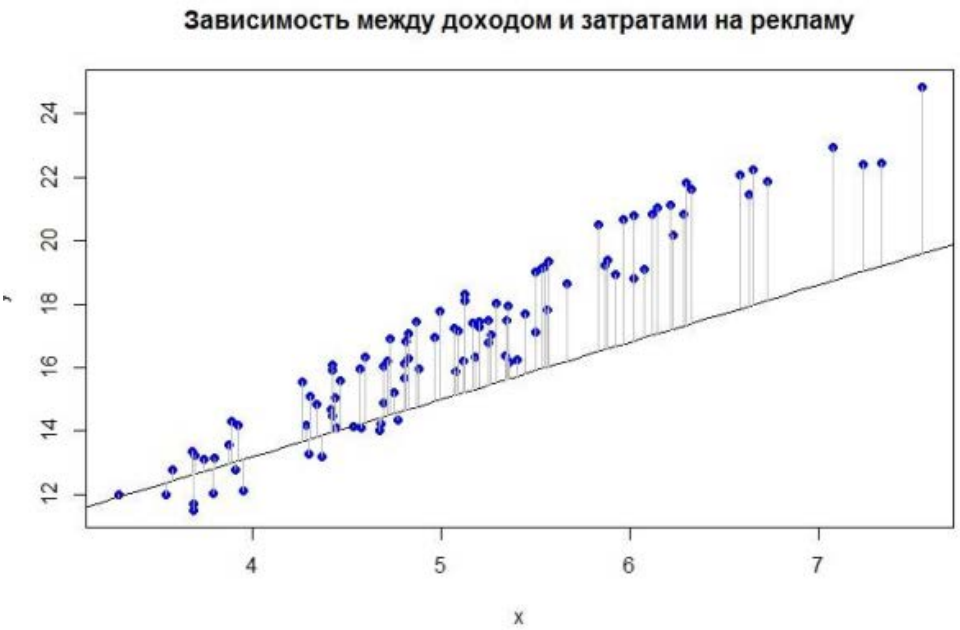


Рисунок 3.

Найдём решение задачи (2). Целевую функцию задачи (2) обозначим через $\varphi(\beta_0, \beta)$. Очевидно, $\varphi(\beta_0, \beta)$ – дифференцируемая функция двух переменных. Найдём её частные производные:

$$\begin{aligned} \frac{\partial \varphi}{\partial \beta_0} &= 2\sum_{i=1}^{n}\left(\beta_0 + \beta x_i - y_i\right) \\ \frac{\partial \varphi}{\partial \beta} &= 2\sum_{i=1}^{n}\left(\beta_0 + \beta x_i - y_i\right) x_i \end{aligned} \tag{3}$$

и запишем систему для поиска стационарной точки:

$$\begin{cases} 2\sum_{i=1}^{n}\left(\beta_0 + \beta x_i - y_i\right) = 0 \\ 2\sum_{i=1}^{n}\left(\beta_0 + \beta x_i - y_i\right) x_i = 0. \end{cases} \tag{4}$$

После несложных преобразований системы (4) получим

$$\begin{cases} n\beta_0 + \beta\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 \\ \beta_0\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0. \end{cases} \tag{5}$$

Введём обозначения: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ и $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

Поделив оба уравнения системы (5) на n, получим

$$\begin{cases} \beta_0 + \beta\bar{x} - \bar{y} = 0 \\ \beta_0\bar{x} + \beta\overline{x^2} - \overline{xy} = 0. \end{cases} \quad (6)$$

Нетрудно показать, что система (5) имеет единственное решение (β_0^*, β^*), , где

$$\beta^* = \frac{\bar{x}\cdot\bar{y} - xy}{\bar{x}^2 - \overline{x^2}}, \quad \beta_0^* = \bar{y} - \beta^*\bar{x}. \quad (7)$$

Учитывая свойства функции $\varphi(\beta_0, \beta)$, нетрудно показать также, что это решение (т.е., стационарная точка функции $\varphi(\beta_0, \beta)$) является точкой минимума функции $\varphi(\beta_0, \beta)$. Иными словами, определяемые формулами (7) значения β_0^* и β^*обеспечивают получение наилучшей (в смысле задачи (2)) линейной функции, отражающей зависимость переменной отклика y от фактора x . График этой линейной зависимости называется прямой регрессии (y на x). На приведённом ниже рисунке 4 эта прямая имеет красный цвет.

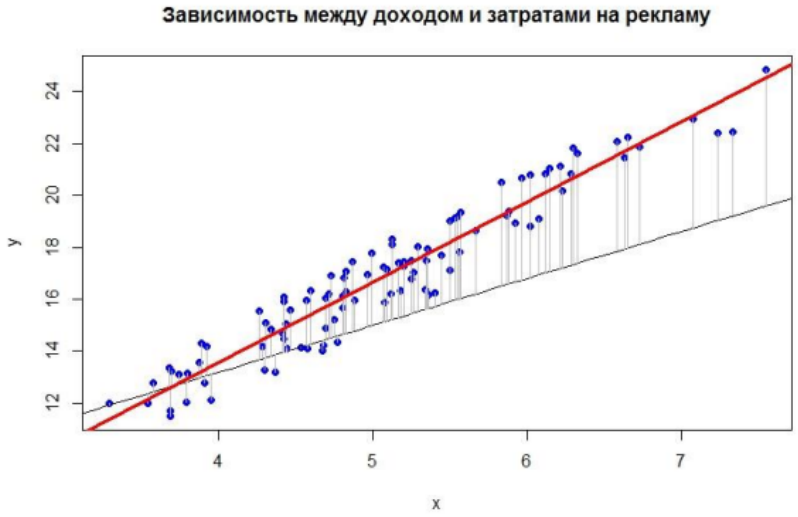


Рисунок 4

Найденная линейная функция позволяет прогнозировать значение зависимого признака (y) по заданным значениям независимого фактора (x).

Пример – однофакторная линейная регрессия в среде R

Рассмотрим решение задачи построения, анализа и использования уравнения однофакторной линейной регрессии на примере данных о моллюсках вида abalone.

В данных содержится информация о 8 признаков моллюсков, это –длина особи, её вес, диаметр раковины и т.д. Число особей, для которых были произведены измерения, равно 4177.

Найдём зависимость диаметра раковины от длины особи (будем предполагать, что зависимость линейна и искать коэффициенты линейной функции).

```
# Прочитаем содержимое файла в переменную A
A=read.table("abalone.txt",header=FALSE,sep=",")
A
# Сохраним данные по интересующим нас признакам в переменных
# Второй столбец - длина моллюска dlina=A[,2]
# Третий столбец - диаметр раковины diam=A[,3]
```

```
# Построим облако точек
plot(dlina, diam, col="blue", type="p", pch=16, xlab="dlina", ylab="diam", main="Зависимость между длиной моллюска и диаметром раковины")
```

Обратите внимание: здесь мы указываем параметры функции plot:

col – цвет символов (синий);

type – тип символа («p» – (от англ.: «point»— точка);

pch – вид точки (от англ.: point character) (16 – закрашенная, 1 – пустой кружок и т.д.);

xlab, ylab – надписи на осях;

main – заголовок графика.

Получим график, показанный на рисунке

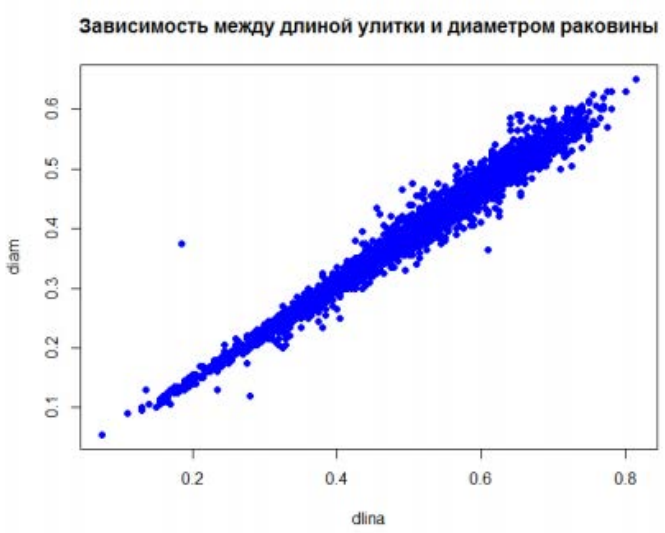


Рисунок 5. – Облако точек, характеризующих зависимость диаметра раковины от длины моллюска.

Как видно из рисунка 5, есть основания полагать, что между признаками наблюдается положительная линейная зависимость. Для нахождения коэффициентов линейной регрессии воспользуемся встроенной функцией `lm` для нахождения коэффициентов этой зависимости. Заметим, что результатом вызова функции `lm` является некоторый объект. Назовём его, например, `myregress`. Обратите внимание на формат вызова функции `lm` – формулу нужно понимать так: линейная функция, отражающая зависимость переменной «`diam`» от переменной «`dlina`»:

```
myregress=lm(formula = diam ~ dlina)
myregress
```

Получим результат , показанный на рисунке 6.

```
Coefficients:
(Intercept)        dlina
   -0.01941      0.81546
```

Рисунок 6. – Результат вызова функции lm

Как видно из Рис.6, свободный коэффициент модели (т.е. значение линейной функции в нуле) равен примерно –0,019, коэффициент при переменной «`dlina`» (т.е. угол наклона прямой) равен примерно 0,815. Построим прямую регрессии и совместим её с облаком точек:

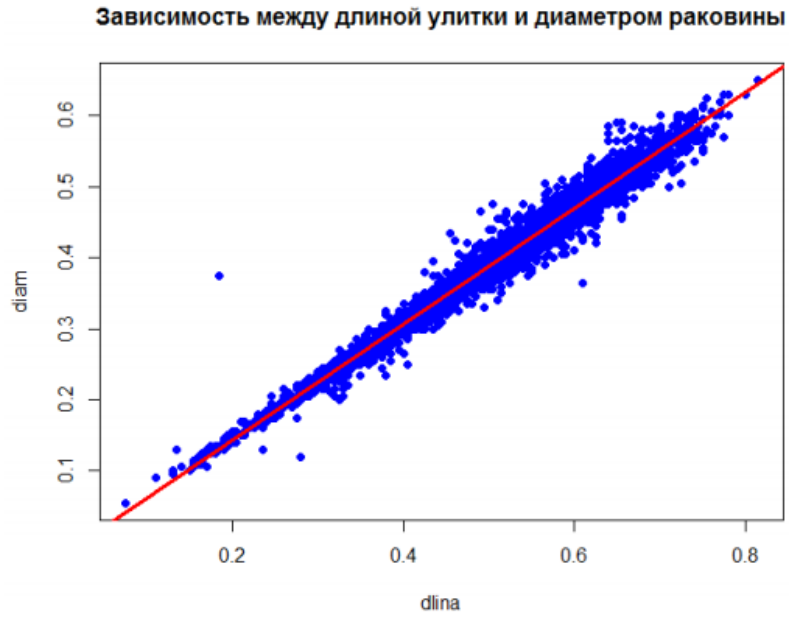
```
# Построим график уравнения однофакторной линейной регрессии (с найденными коэффициентами)
```

```
abline(regress,col="red",lwd="3",add=TRUE)
```

Здесь параметр `lwd` задаёт толщину линии, параметр `add` со значением `TRUE` указывает на то, что график должен быть построен на том же рисунке, что и предыдущий графический объект (т.е. в нашем случае он будет совмещён с облаком точек). Получим результат, показанный на Рис.7

Рисунок 7. – Облако точек, характеризующих зависимость диаметра раковины от длины моллюска, и прямая регрессии

Заметим, что свободный коэффициент (`Intercept`), рассчитанный с помощью функции `lm`, оказался равным –0,01941. В рассматриваемом примере естественно считать, что этот параметр равен 0 (т.к. особь нулевой длины имеет нулевой диаметр раковины). Пакет R позволяет



задать значения параметров, если они известны априори. Так, чтобы указать, что параметр `Intercept` тождественно равен нулю, нужно изменить формулу, которую мы передаём функции `lm`, следующим образом:

```
# Положим Intercept = 0
myregress1=lm(formula = diam ~ -1 + dlina)
myregress1
```

Получим результат , показанный на рис.8:

```
Coefficients:
 dlina
0.7803
```

Рисунок 8. – Результат вызова функции `lm` при заданном (нулевом) значении параметра `Intercept`.

Как видим из Рис. 5 и 7, фиксация параметра `Intercept` (в нуле) привела к изменению второго коэффициента уравнения регрессии (отвечающего за наклон прямой): вместо 0,81546 этот коэффициент теперь стал равен 0,7803. Построим график функции линейной регрессии с нулевым коэффициентом `Intercept` зелёным цветом и совместим его с предыдущим графиком:

```
# Построим прямую "исправленной" регрессии:
abline(myregress1,col="green",lwd="3",add=TRUE)
```

Полученный результат показан на Рис.9

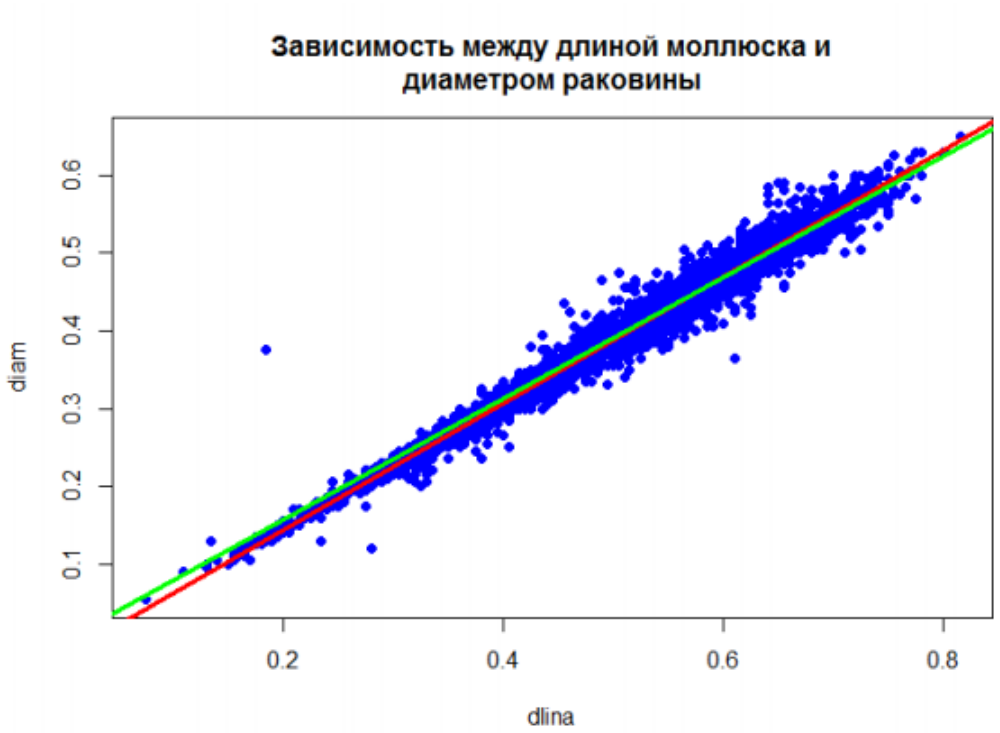


Рисунок 9. – Зелёным цветом показана «исправленная» регрессионная прямая.

Исследуем теперь эту модель – выясним, насколько сильна статистическая зависимость между признаками (длиной особи и диаметром раковины).

Для исследования модели вызовем функцию `summary`, выводящую всю информацию о модели:

```
summary(myregress)
```

Получим результат, показанный на рис.10:

```
Call:
lm(formula = diam ~ dlina)

Residuals:
      Min        1Q    Median        3Q       Max
-0.113017 -0.008703 -0.000549  0.008678  0.243553

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.019414   0.001113  -17.44   <2e-16 ***
dlina        0.815461   0.002070  393.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01607 on 4175 degrees of freedom
Multiple R-squared:  0.9738,    Adjusted R-squared:  0.9738
F-statistic: 1.552e+05 on 1 and 4175 DF,  p-value: < 2.2e-16
```

Рисунок 10. – Сводная информация о линейной регрессионной модели.

Среди прочих характеристик, здесь приведены результаты проверки гипотезы о том, что признаки (в нашем случае это – длина особи и диаметр раковины) независимы, т.е. коэффициент

корреляции ρ между этими признаками равен 0. Для проверки этой гипотезы используется критерий Стьюдента (или, как его ещё называют, Т-критерий). Напомним, что Т-критерий имеет вид:

$$T=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}, \tag{8}$$

где n – объём выборки, r – выборочный коэффициент корреляции:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \tag{9}$$

Здесь $\bar{x}$ и $\bar{y}$ – выборочные средние, т.е.

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i. \tag{10}$$

В пакете R выборочный коэффициент корреляции вычисляется с помощью функции `cor`:

```
r=cor(dlina,diam)
r
```

Для нашего примера выборочный коэффициент корреляции r между фактором (длиной улитки) и переменной отклика (диаметром раковины) равен 0.9868116. Поскольку r близок к 1, можем предположить наличие сильной линейной зависимости между признаками. Вычислим значение Т-критерия по формуле (1):

```
# Сначала найдём объём выборки n
n=length(dlina)
# Вычислим и выведем значение Т-критерия
t=r/sqrt(1-r^2)*sqrt(n-2)
t
```

Получим число 393.9017.

Заметим, что это число присутствует в таблице «`Coefficients`» (столбец «`t-value`») на Рис.10, т.е. объект `myregress` содержит информацию о наблюдаемом значении критерия Стьюдента. Поскольку наблюдаемое значение критерия Стьюдента (393.9017) сильно отличается от нуля, что мало вероятно, если считать, что наша гипотеза о независимости признаков верна (т.е. ρ=0), у нас есть основания полагать, что гипотеза H_0: $\rho = 0$ ложна (т.е. на самом деле $\rho \neq 0$, а значит, существует линейная зависимость между признаками).

Чтобы сделать вывод о том, насколько сильно наблюдаемое значение Т-критерия (т.е. 393.9017) отлично от нуля, найдём вероятность получить (в условиях нулевой гипотезы) ещё большее значение, т.е. вероятность P(T>393.9017). Для этого воспользуемся тем известным фактом, что с.в. (8) имеет распределение Стьюдента с n-2 степенями свободы:

$$P(T>393.9017) = 1 - F(393.9017),$$

где F – функция распределения Стьюдента с (n-2) степенями свободы.

В пакете R найдём:

```
1-pt(t,n-2).
```

Здесь `pt` – функция распределения Стьюдента. Получим число, очень близкое к 0. Заметим, что на Рис.10. вероятность превысить наблюдаемое значение критерия Стьюдента (в условиях истинности нулевой гипотезы) находится в столбце `Pr(>|t|)`. Оно меньше, чем $2\cdot10^{-16}$, т.е. практически равно нулю. Итак, наша гипотеза об отсутствии зависимости между

признаками не подтвердилась, а значит, линейная зависимость существует, и найденное уравнение регрессии может использоваться для целей прогноза.

Замечание. Проверить гипотезу о равенстве коэффициента корреляции (между признаками *x* и *y*) нулю в пакете R можно с помощью функции `cor.test(x,y).` Результатом вызова этой функции также является величина `Pr(>|t|)` (которую следует сравнить с заранее заданным уровнем значимости ε (обычно его полагают равным 0,05)).

Когда в регрессионной модели есть одна зависимая и одна независимая переменная, такой подход называется простой линейной регрессией. Когда есть одна зависимая переменная, но в модель входят ее степени (например, X, X^2, X^3), это называется полиномиальной регрессией. Если есть больше одной независимой переменной, это называется множественной регрессией.

Множественная линейная регрессия

Если существует больше одной независимой переменной, простая линейная регрессия превращается во множественную линейную регрессию, а ход вычислений становится более сложным.

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:

$$y = b_0 + b_1x_1 + b_2x_2... + b_kx_k .$$

Если существует больше одной независимой переменной, то регрессионные коэффициенты показывают, на сколько увеличится значение зависимой переменной при изменении данной независимой переменной на единицу при условии, что все остальные независимые переменные останутся неизменными.

Важным дополнительным условием является некоррелированность объясняющих переменных между собой (отсутствие мультиколлинеарности). Считается, что явление мультиколлинеарности наблюдается тогда, когда коэффициент корреляции между объясняющими переменными превышает по модулю 0,7.

Пошаговая множественная линейная регрессия

Существуют две методики построения множественной регрессии – пошаговая вперед и пошаговая назад.

Пошаговая вперед заключается в то, что первоначально строится модель с одной экзогенной переменной. Затем добавляется следующая и строится новая модель. Модели сравниваются и, в зависимости от того ухудшилась или улучшилась модель, введенная переменная либо остается в модели, либо заменяется на другую. Таким образом, перебираются различные комбинации экзогенных переменных, в результате получается наилучшая модель.

Пошаговая назад начинается с того, что рассчитывается множественная регрессия на всем множестве факторов. Затем построенная модель исследуется с точки зрения статистической значимости модели в целом, статистической значимости коэффициентов регрессии, оценивается коэффициент детерминации. Затем из модели удаляется один из влияющих факторов.

Его выбор можно осуществить следующим образом:

1. Строится матрица парных коэффициентов корреляции между переменными.

2. Выбираются две экзогенные переменные, между которыми наибольший коэффициент парной корреляции.

3. Из этих двух переменных выбирается та, которая оказывает меньшее влияние на эндогенную переменную, и исключается из модели.

Затем строится новая модель, исследуется ее качество. Также проводится тест на лучшую из двух моделей: с меньшим или большим числом переменных. В конце получается наилучшая модель.

Результат применения метода пошаговой регрессии зависит от критериев включения или удаления переменных. При помощи функции `stepAIC()` из пакета MASS можно провести построение пошаговой регрессии с использованием точного критерия AIC (Akaike Information Criterion – информационный критерий Акаике).

При расчете этого критерия учитывается статистическое соответствие модели данным и число необходимых для достижения этого соответствия параметров. Предпочтение нужно отдавать моделям с **меньшими** значениями AIC, указывающими на хорошее соответствие данным при использовании меньшего числа параметров**.**

Мультиколлинеарность

Наибольшие затруднения в использовании аппарата множественной регрессии возникают при наличии мультиколлинеарности факторных переменных, когда более чем два фактора связаны между собой линейной зависимостью.

Мультиколлинеарностью для линейной множественной регрессии называется наличие линейной зависимости между факторными переменными, включёнными в модель.

Мультиколлинеарность – нарушение одного из основных условий, лежащих в основе построения линейной модели множественной регрессии.

Мультиколлинеарность можно выявить на начальном этапе моделирования (до построения регрессии). О ней могут свидетельствовать:

1. Большие (по абсолютной величине) парные коэффициенты корреляции между независимыми переменными.

2. Высокие (>10) значения коэффициента *VIF*.

Коэффициент *VIF* (variance inflation factor) характеризует силу мультиколлинеарности. Вычисление коэффициента выполняется с помощью функции `vif()`

Симптомами присутствия мультиколлинеарности в уже построенной модели являются:

1. Небольшое изменении исходных данных, приводит к существенному изменению оценок коэффициентов.

2. Каждая переменная в отдельности является незначимой, а уравнение в целом имеет высокий R^2 (коэффициент детерминации) и является значимым.

3. Оценки коэффициентов имеют неправильные с точки зрения экономической теории знаки или неоправданно большие значения

Пример – множественная линейная регрессия в среде R

Рассмотрим нахождение коэффициентов линейной регрессионной модели и (на их основе) прогнозирование значения переменной отклика по формуле:

$$\hat{y}_i = X_i\hat{\theta},\ i = n+1,\ \ldots$$

Файл с данными имеет вид:

TotalSquare (m2)	LivingSquare (m2)	DistCenter (km)	DistMetro (km)	Price
80	53	17	2,1	14 612 000,00 ₽
76	51	1	0,7	16 931 128,00 ₽
96	72	16	1,3	18 905 472,00 ₽
56	37	16	2,2	14 829 304,00 ₽
75	56	6	2,8	19 214 025,00 ₽
75	56	11	1,1	19 582 950,00 ₽
97	65	12	1,4	19 123 259,00 ₽
30	24	14	2,3	6 035 280,00 ₽
84	63	7	1,9	20 058 696,00 ₽
50	33	11	2,4	13 807 800,00 ₽
55	44	6	1,7	13 087 745,00 ₽
94	71	10	1,7	17 337 266,00 ₽
91	68	5	1,4	17 189 900,00 ₽
32	26	7	1,8	6 405 792,00 ₽
86	65	4	0,3	19 267 698,00 ₽
55	41	2	1,6	13 827 495,00 ₽
65	49	18	1,7	11 242 920,00 ₽
45	34	9	2,2	12 004 470,00 ₽
47	38	4	1,3	10 586 844,00 ₽

Скопируем данные (без заголовков) в текстовый файл, предварительно преобразовав данные из денежного формата. В текстовом файле заменим десятичные запятые точками. Сохраним текстовый файл как flats.txt. Выполним расчёты двумя способами:

1) явным, по формулам;

2) с помощью библиотечной функции lm.

Ниже приведён код R-программы для вычисления коэффициентов линейной регрессионной модели и для получения прогноза, т.е. оценки стоимости квартиры по заданным значениям факторов.

R-код

```
#
# МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ
#
# Назначим текущий директорий (в котором находится наш файл с данными) setwd("D://Olga/Multiple_Regression")
#
N=50  # Число наблюдений (квартир)
M=4 # Число факторов
# Создадим "шаблоны" матриц
Data = matrix(1 : (N*(M+1)), ncol=(M+1)) # Данные из файла
X = matrix(1 : (N*M), ncol=M) # Матрица значений факторов, расширенная единичным столбцом
Y = matrix(1 : N, ncol=1) # Столбец значений переменной отклика (цен на квартиры)
T = matrix(1 : (M+1)*(M+1), ncol = (M+1) ) Teta = matrix(1 : (M+1), ncol = 1) # Вектор коэффициентов ЛРМ
 # Прочтём файл с характеристиками квартир и ценами и запишем данные в заготовленную матрицу
Data = read.table("flats.txt", sep = '\t') #edit(Data)
# Разделим значения факторов и переменной отклика
X = Data[, 1:M]
Y = Data[, (M+1)]
# Создадим вектор из единиц (как набор из N единиц)
odin=vector(length=N,mode='numeric')
odin = rep(1,N)
```

```
# Заготовим матрицу нужного размера
X1 = matrix(1 : (N*(M+1)),ncol=(M+1))
# Заполним первый столбец матрицы X1 единицами
X1[,1] = odin #edit(X1)
# Остальные столбцы возьмём из матрицы
X for(i in 1:M) X1[,i+1] = X[,i]
# Найдём вектор параметров множественной линейной регрессии Teta
# Вычислим сначала матрицу t(X1)
T1 = t(X1)
# Вычислим матрицу T=t(X1)%*%X1
T=T1%*%X1
# Подключим библиотеку для работы с матрицами
library(MASS)
# Найдём обратную матрицу для T
obr = ginv(T)
# Вычислим вектор Teta - вектор коэффициентов линейной модели
Teta=obr%*%T1%*%Y
Teta=Teta[,1]
print(Teta)
# ============================================================
# Введём данные нашей квартиры (на первое место сразу поставим единицу)
myx = c(1, 65.7, 46.2, 9, 1.5)
myx
myX=matrix(myx,ncol=M+1,byrow=TRUE)
tmyX=t(myX)
# Оценим стоимость нашей квартиры
myprice=Teta%*%tmyX print(myprice)
# ============================================================
```

Получили результат: 14539.25.

Таким образом, согласно построенной модели, квартира со следующими характеристиками: общая площадь – 65.7 м ^{2} , жилая – 46,2 м^2 , находящаяся на расстоянии 9 км от центра города и 1,5 км от метро, оценивается в 14 млн. 539 тыс. 250 рублей.

Пример – построение множественной линейной регрессионной модели в пакете R с помощью функции lm

Код	Результат

`# МНОГОФАКТОРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ`
`# Назначим текущий директорий (в котором находится наш файл с данными)`
`setwd("C://Olga/КФУ/кэк/Анализ_данных/Proba")`	

`# Прочтём файл с характеристиками квартир и ценами`
`myData = read.csv("myflats.txt", sep = '\t')`
`myData`
	`> myData`
	totsquare livesquare floor height distcenter distmetro price
	1 83 64 8 10 5.0 500 10.200
	2 72 58 4 5 1.5 1500 8.300
	3 64 45 1 5 10.0 2000 7.950
	4 85 65 4 12 2.0 100 11.000
	5 36 28 2 5 6.0 1000 5.100
	6 98 82 2 16 8.0 350 15.020
	7 74 64 9 16 6.5 800 8.760
	8 56 44 3 10 9.0 3000 6.900
	9 28 21 12 12 15.0 3500 4.985
	10 68 56 4 12 8.0 800 8.120
	11 88 79 2 4 4.0 300 11.900
	Это – фрагмент консоли, где выведен объект myData

```
# Построим регрессионную модель

mymodel = lm(price ~ -1+ totsquare + livesquare + floor + height + distcenter + distmetro, data = myData)

summary(mymodel)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.6560 -1.7738 -0.7540  0.6493  9.3139

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
totsquare   0.0519874  0.1177009   0.442    0.665
livesquare  0.1386667  0.1465748   0.946    0.360
floor      -0.0089039  0.2659979  -0.033    0.974
height     -0.2524347  0.2125690  -1.188    0.255
distcenter  0.1461910  0.1682269   0.869    0.399
distmetro  -0.0002076  0.0009206  -0.225    0.825

Residual standard error: 3.273 on 14 degrees of freedom
Multiple R-squared:  0.9395,    Adjusted R-squared:  0.9136
F-statistic: 36.25 on 6 and 14 DF,  p-value: 9.547e-08
```

```
# Введём данные нашей квартиры (чтобы оценить её стоимость)

myflat=c(46,38,10,10,11,1.5)

myflat=data.frame(t(myflat))

colnames(myflat)<-c("totsquare","livesquare","floor","height","distcenter","distmetro")

myflat
```

```
myflat
totsquare livesquare floor height distcenter dismetro
       46         38    10     10          _      11      1.5
```

```
# Спрогнозируем цену нашей квартиры по нашей модели

newprice = predict(mymodel, myflat)

newprice
```

```
6.65516
```

Задание и порядок выполнения лабораторной работы №1

Задание 1.

Используя тестовые данные для решения задач по теме «Однофакторная линейная регрессия» или данные, смоделированные Вами для двух признаков, выполните следующее:

1. Постройте на графике облако точек;
2. Найдите коэффициенты линейной регрессии;
3. Совместите на графике линию регрессии с облаком точек;
4. Оцените визуально характер зависимости признаков.
5. Сделайте выводы.

Задание 2.

Подберите реальные данные для задачи однофакторной линейной регрессии. Выполните следующее:

1. Постройте на графике облако точек.
2. Найдите уравнение регрессии. (Укажите численные значения коэффициентов регрессии)
3. Совместите уравнение регрессии с облаком точек.
4. Сделайте выводы относительно зависимости исследуемых признаков.
5. Выберите произвольно несколько значений независимого признака x и вычислите ожидаемые (согласно полученному уравнению регрессии) значения признака y.
6. Покажите на графике точки, соответствующие сделанному прогнозу.
7. Оформите отчёт. Включите в отчёт построенный график, запишите численно коэффициенты регрессии, коэффициент корреляции, значение критерия Стьюдента, расчётные значения прогнозируемого признака. Обоснуйте выводы относительно зависимости признаков и качества прогноза.

Задание 3.

Изучите теоретический материал по теме «Множественная линейная регрессия». Рассмотрите примеры решения задачи в пакете R.

Задание 4.

Ознакомьтесь с материалами:

ЛЗ_2_Ввод и редактирование данных с помощью редактора данных.pdf

ЛЗ_2_Исследование и улучшение линейной регрессионной модели.pdf

ЛЗ_2_Предварительная обработка данных.pdf

Задание 5.

Решите в пакете R задачу построения и анализа уравнения множественной линейной регрессии:

1. Подберите данные для задачи;
2. Постройте с помощью пакета R уравнение линейной регрессии (сделайте это 2мя способами:
   - по явным формулам;
   - с помощью функции lm.
3. Сравните результаты с помощью таблицы следующего вида:

Коэффициенты уравнения множественной линейной регрессии, вычисленные		
	по явным формулам	**с помощью функции lm**
beta_0		
beta_1		
...		
beta_m		

4. Приведите в отчёте матрицу попарных корреляций и сводную информацию (summary) модели.
5. Проанализируйте степень влияния факторов на переменную отклика согласно каждой из полученных моделей;
6. Сделайте прогноз (с помощью каждой из полученных моделей. если они различны) - на вход модели подайте не менее 10 значений из обучающей выборки и сравните:
   - истинное значение переменной отклика;
   - прогноз, полученный с помощью модели, коэффициенты которой вычислены явно;
   - прогноз, полученный с помощью модели, коэффициенты которой вычислены с помощью функции lm.
7. Проанализируйте и объясните полученные результаты.

Контрольные вопросы

1. Линейная однофакторная регрессия
2. Множественная линейная регрессия
3. Построение множественной линейной регрессионной модели в пакете R
4. Построение линейной однофакторной регрессионной модели в пакете R

Библиография

1. Алексей Шипунов и др. Наглядная статистика. Используем R! – М.: ДМК Пресс, 2014. – 298 с. [Электронный ресурс]. Режим доступа: http://ashipunov.info/shipunov/school/books/rbook.pdf.
2. Зарядов И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. М.: Издательство Российского университета дружбы народов, 2010. – 207 с.
3. Роберт И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил.
4. Оффициальный сайт RStudio. Режим доступа: https://www.rstudio.com.
5. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. Режим доступа: http://machinelearning.ru.
6. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга. Режим доступа: http://r-analytics.blogspot.com