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.

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:

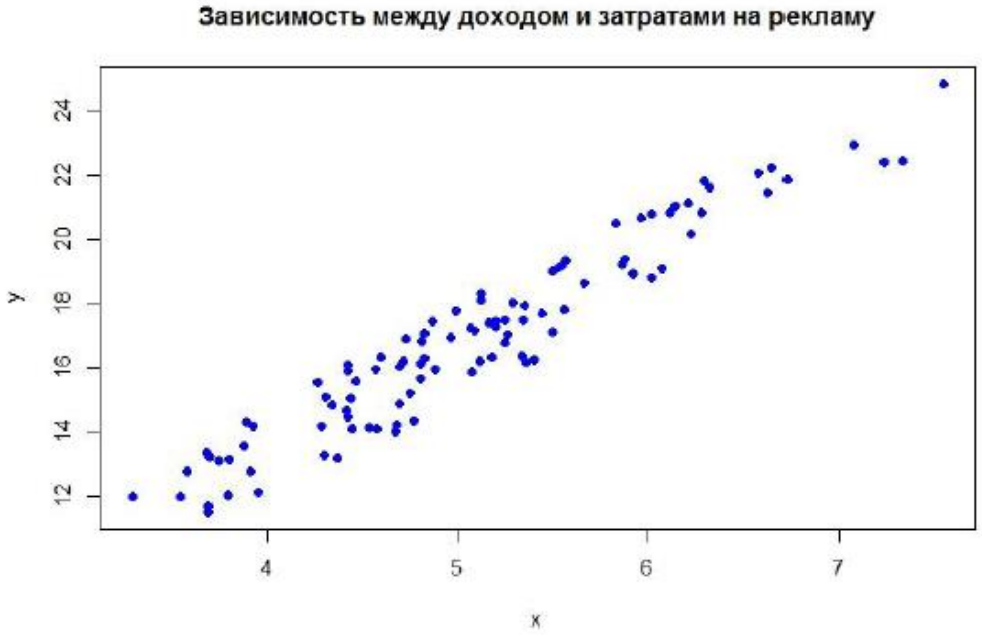
№ наблюдения, i	Значения фактора, x_i	Значения переменной отклика, y_i
1	x_1	y_1
…		
i	x_i	y_i
…		
n	x_n	y_n

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) $(i = \overline{1,n})$ 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.



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

Пусть Y зависит от одной переменной x . При этом предполагается, что Y принимает заданные (фиксированные) значения, а зависимая переменная x имеет случайный разброс из-за ошибок измерения, влияния неучтенных факторов и других причин. Каждому значению x соответствует некоторое вероятностное распределение случайной величины (СВ) Y . Предположим, что СВ Y «в среднем» линейно зависит от значений переменной x . Это означает, что условное математическое ожидание

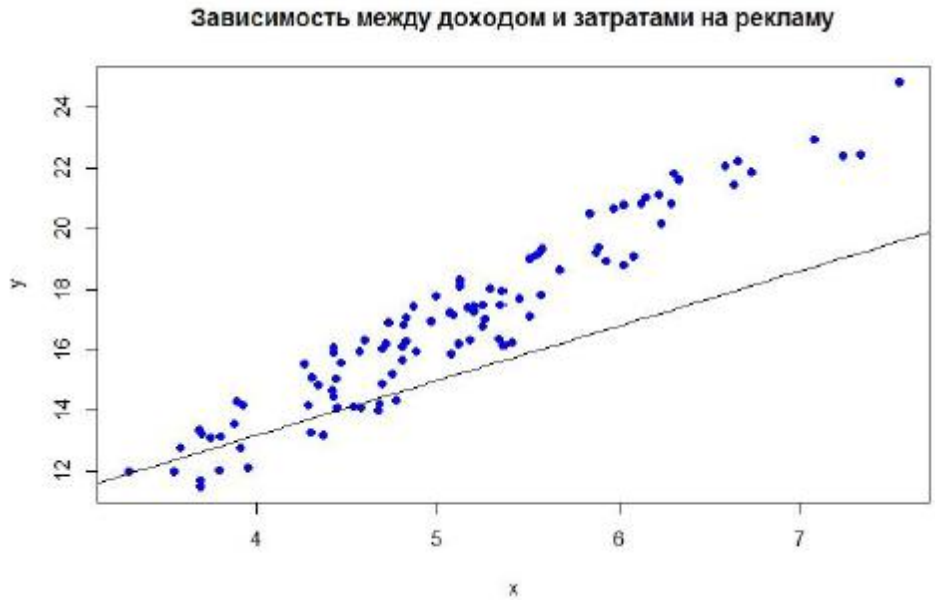
случайной величины Y при заданном значении переменной x имеет вид (y среднее значение всех значений y для данного x)

$$M(Y|x) = \beta_0 + \beta_1 x$$

Функция переменной x , определяемая правой частью этой формулы, называется линейной регрессией Y на x , а параметры β_0, β_1 - параметрами линейной регрессии.

β_1 – тангенс угла наклона графика этой функции к оси OX (английский термин: «*slope*»), β_0 – ордината точки пересечения этой прямой с осью *OY* (англ.: «*intercept*»). Задача состоит в том, чтобы *найти такие значения переменных* β_0, β_1, *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_1 и построим соответствующую прямую:



Очевидно, построенная «наугад» прямая является не самой лучшей для данного облака точек. Формализуем понятие «качества» модели. При фиксированных β_0 и β_1 «ожидаемое» (согласно (1)) значение y при $x = x_i$ составляет $\beta_0 + \beta_1 \cdot x, \quad i = 1,..n$ (т.е. точка $(x_i, \beta_0 + \beta_1 \cdot x_i)$ лежит на построенной прямой). Но фактическое значение переменной y при $x = x_i$ составляет y_i, т.е. «ошибка» составляет $((\beta_0 + \beta_1 \cdot x_i) - y_i)$.

На практике параметры линейной регрессии неизвестны и их оценки определяются по результатам наблюдений переменных Y и x . Пусть проведено n независимых наблюдений случайной величины Y при значениях переменной X: $x_1, x_2, ..., x_n$. При этом измерения величины Y дали следующие результаты: $y_1, y_2, ..., y_n$. Так как эти значения имеют разброс относительно линейной регрессии, то связь между переменными Y и x можно записать в виде линейной (по параметрам β_0, β_1) регрессионной модели:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

где ε - случайная ошибка наблюдений, причем $M(\varepsilon) = 0$, $D(\varepsilon) = \sigma^2$. Значение дисперсии ошибок σ^2 неизвестно, и оценка ее определяется по результатам наблюдений.

Задача линейного регрессионного анализа состоит в том, чтобы по результатам наблюдений (x_i, y_i), $i = \overline{1,n}$:

1.Получить наилучшие точечные и интервальные оценки неизвестных параметров β_0, β_1 и σ^2 линейной регрессионной модели.

2) Проверить статистические гипотезы о параметрах модели.

3) Проверить, достаточно ли хорошо модель согласуется с результатами наблюдений (адекватность модели результатам наблюдений).

В соответствии с моделью результаты наблюдений зависимой переменной Y : $y_1, y_2, ..., y_n$ являются реализациями случайных величин $\beta_0 + \beta_1 x_i + \varepsilon_i$, обозначаемых y_i, $i = \overline{1,n}$.

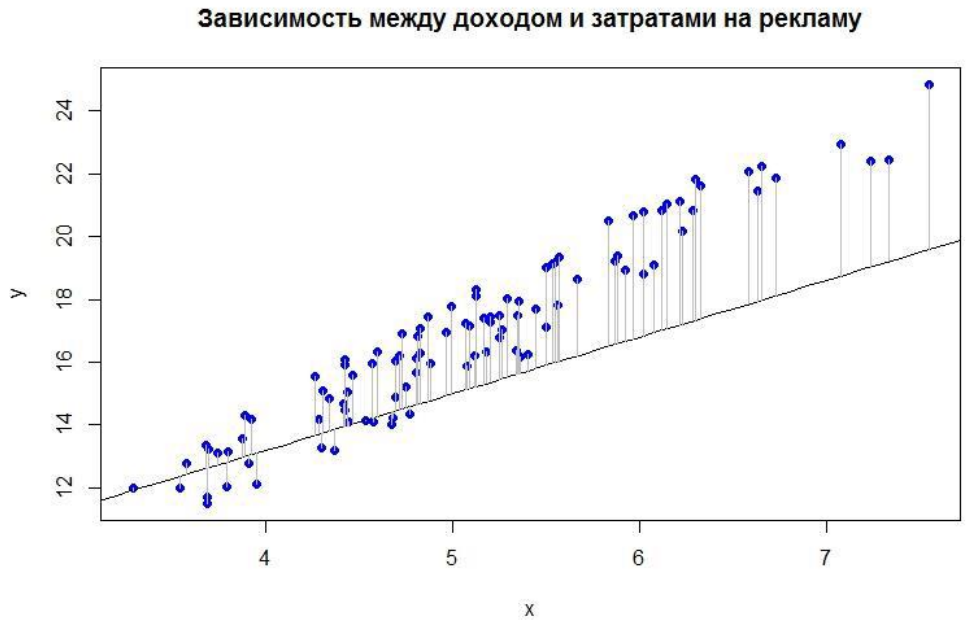
Задача линейного регрессионного анализа решается в предположении, что случайные ошибки наблюдений ε_i и ε_j не коррелированы, имеют математические ожидания равные нулю, и одну и ту же дисперсию, равную σ^2, т.е. $M(\varepsilon_i) = 0$, $K_{\varepsilon_i \varepsilon_j} = \begin{cases} 0, i \neq j \\ \sigma^2, i = j \end{cases}$, $i = \overline{1,n}$. При статистическом анализе регрессионной модели предполагается также, что случайные ошибки наблюдений $\varepsilon_i, i = \overline{1,n}$ имеют нормальное распределение, т.е. $\varepsilon_i \sim N(0, \sigma^2)$, $i = \overline{1,n}$. В этом случае ошибки наблюдений также являются независимыми случайными величинами. Можно определить величину ошибки для всех отмеченных точек. Линейная модель, которая наилучшим образом аппроксимирует данные – одна из тех, для которых общая ошибка выборки имеет наименьшее значение. Чтобы рассчитать ее нужно избежать позитивных и негативных значений. Это можно сделать, возведя все ошибки в квадрат и делая их положительными величинами. Линия наилучшего подбора – та, которая минимизирует квадраты разниц между рассматриваемыми значениями y и соответствующими значениями x, рассчитанными с помощью линии наилучшего подбора. Эта линия называется линией регрессии, полученной методом наименьших квадратов. Для нахождения оценок параметров модели по результатам наблюдений используется метод наименьших квадратов. По этому методу выбирают такие оценки β_0, β_1, которые минимизируют сумму квадратов отклонений наблюдаемых значений случайных величин Yi от их математических ожиданий, т.е.

$$\sum_{i=1}^{n}((\beta_0 + \beta_1 \cdot x_i) - y_i)^2 \to \min \qquad (2)$$

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 следующем рисунке.



Значение целевой функции задачи (2) при фиксированных значениях β_0 и β_1 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).

Найдём решение задачи (2). Целевую функцию задачи (2) обозначим через $\varphi(\beta_0, \beta_1)$. Очевидно, $\varphi(\beta_0, \beta_1)$– дифференцируемая функция *двух* переменных. Найдём её частные производные:

$$\frac{\partial \varphi}{\partial \beta_0} = 2\sum_{i=1}^{n}((\beta_0 + \beta_1 \cdot x_i) - y_i) \qquad (3)$$

$$\frac{\partial \varphi}{\partial \beta_1} = 2\sum_{i=1}^{n}((\beta_0 + \beta_1 \cdot x_i) - y_i)x_i$$

и запишем систему для поиска стационарной точки:

$$\begin{cases} \sum_{i=1}^{n}\left(\left(\beta_0+\beta_1\cdot x_i\right)-y_i\right)=0 \\ \sum_{i=1}^{n}\left(\left(\beta_0+\beta_1\cdot x_i\right)-y_i\right)x_i=0 \end{cases} \qquad (4)$$

После несложных преобразований системы (4) получим

$$\begin{cases} n\beta_0+\beta_1\cdot\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i=0 \\ \beta_0\sum_{i=1}^{n}x_i+\beta_1\cdot\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_iy_i=0 \end{cases} \qquad (5)$$

Введём обозначения: $\bar{x}=\frac{1}{n}\sum_{n=1}^{n}x_i$, $\bar{y}=\frac{1}{n}\sum_{n=1}^{n}y_i$, $\overline{x\cdot y}=\frac{1}{n}\sum_{n=1}^{n}x_iy_i$ и $\overline{x^2}=\frac{1}{n}\sum_{n=1}^{n}x_i^2$.

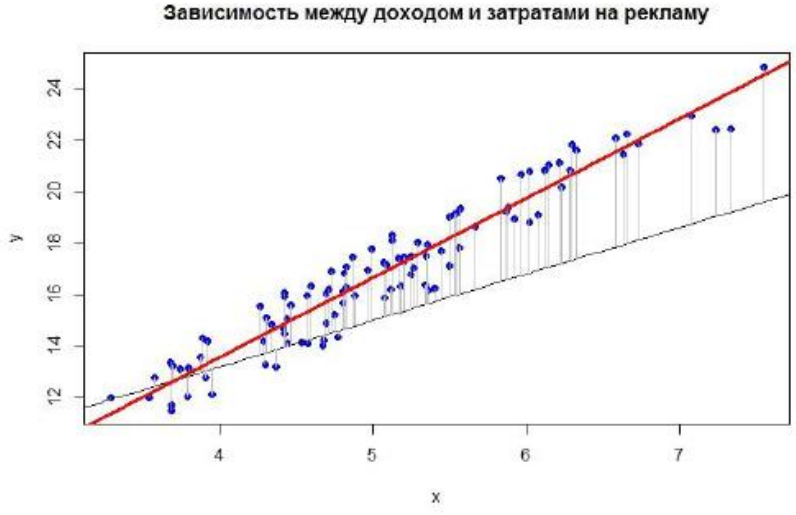
Поделив оба уравнения системы (5) на n, получим

$$\begin{cases} \beta_0+\beta_1\cdot\bar{x}-\bar{y}=0 \\ \beta_0\bar{x}+\beta_1\cdot\overline{x^2}-\overline{x\cdot y}=0 \end{cases} \qquad (6)$$

Нетрудно показать, что система (5) имеет единственное решение $\left(\beta_0^*,\beta_1^*\right)$, где

$$\beta_1^*=\frac{\bar{x}\cdot\bar{y}-\overline{x\cdot y}}{\bar{x}^2-\overline{x^2}} \qquad \beta_0^*=\bar{y}-\beta_1^*\cdot\bar{x} \qquad (7)$$

Учитывая свойства функции $\varphi(\beta_0,\beta_1)$, нетрудно показать также, что это решение (т.е., стационарная точка функции $\varphi(\beta_0,\beta_1)$ является *точкой минимума* функции $\varphi(\beta_0,\beta_1)$. Иными словами, определяемые формулами (7) значения β_0^* и β_1^* 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 эта прямая имеет красный цвет.



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

Указания к выполнению задания

1. Задать векторы x и y с помощью функции «c» (сокр. от англ. Combine):

```
> x<–c(7,8,9,10,11,12,13,14,15,16)
> y<–c(6.4,7.9,8.5,8.1,7.4,7.9,8.2,9.5,9.0,10.0)
> x
 [1]  7  8  9 10 11 12 13 14 15 16
> y
 [1] 6.4  7.9  8.5  8.1  7.4  7.9  8.2  9.5  9.0 10.0
```

2. Построить график экспериментальных значений, используя функцию plot:

```
plot(x,y).
```

3. Построить уравнение парной линейной регрессии, используя команду lm:

```
mod<–lm(y~x)
```

4. Получить полный набор расчетов по модели командой summary:

```
> mod<–lm(y~x)
> summary(mod)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.75091 -0.52364 -0.01818  0.50045  0.90545

Coefficients:
            Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  5.09091    0.83366     6.107    0.000287 ***
x            0.27818    0.07033     3.955    0.004205 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6388 on 8 degrees of freedom
Multiple R-squared:  0.6617,    Adjusted R-squared:  0.6194
F-statistic: 15.64 on 1 and 8 DF,  p-value: 0.004205
```

Команда R summary выводит:

− оценки коэффициентов регрессии (Estimate):

(Intercept) 5.09091 = β_0

x 0.27818 = β_1

− стандартные ошибки коэффициентов (Std. Error);

− наблюдаемые значения t-критерия при проверке значимости коэффициентов регрессии (t value);

− P-значения для коэффициентов регрессии (P-value).

Звездочками или точками в столбце справа R показывает значимость или незначимость коэффициентов:

*** — значимы на уровне значимости менее 0.001;

** — значимы на уровне значимости 0.001;

* — значимы на уровне значимости 0.01;

. — значимы на уровне значимости 0.05 и т. д.

Эти обозначения приведены в разделе Signif. codes.

Коэффициент детерминации (Multiple R-squared) равен 0.6617; скорректированный коэффициент детерминации (Adjusted R-squared) равен 0.6194.

Наблюдаемое значения F-критерия проверки значимости уравнения в целом и P-значение: F-statistic: 15.64 on 1 and 8 DF, p-value: 0.004205

Таким образом, уравнение регрессии получилось значимым.

Коэффициент β_1 является мерой эффекта: изменение величины x на 1 вызывает изменение y в среднем на β_1.

Коэффициент детерминации характеризует качество построенного уравнения регрессии; чем ближе коэффициент детерминации R^2 к единице, тем выше качество полученного уравнения регрессии. Максимальное значение коэффициента детерминации $R^2 = 1$ достигается в том случае, когда все остатки = 0, а уравнение прямой регрессии проходит точно через все точки y_i. Однако при включении в модель нескольких независимых переменных, с такой интерпретацией R^2 следует быть очень осторожным. Дело в том, что значение R^2 *всегда будет возрастать* при увеличении числа предикторов в модели, даже если некоторые из этих предикторов не имеют тесной связи с зависимой переменной. Соответственно, простой коэффициент детерминации будет отдавать предпочтение т.н.переобученным моделям, что крайне нежелательно. Выход заключается в использовании *скорректированного коэффициента детерминации* (англ. *adjusted R-squared*):

$$R^2_{adj} = R^2 - \left(1 - R^2\right)\frac{p}{n - p - 1},$$

где R^2 - исходный коэффициент детерминации, p число параметров модели, а n - объем выборки. Как следует из приведенной формулы, поправка сводится к наложению "штрафа" на число параметров модели - чем больше параметров, тем больше этот "штраф" и, как результат, тем меньше значение скорректированного коэффициента детерминации.

Получить оценки β_0, β_1 можно командой

> coef_mod<−coef(mod)

```
> coef_mod
(Intercept)           x
  5.0909091   0.2781818

> b0<-coef_mod[1]
> b0
(Intercept)
  5.090909
> b1<-coef_mod[2]
> b1
       x
0.2781818
```

5. Доверительные интервалы для зависимой переменной.

Одной из центральных задач моделирования является предсказание (прогнозирование) значений зависимой переменной при определенных значениях объясняющих переменных. При этом решаются две основные задачи: либо предсказывается условное математическое ожидание зависимой переменной при определенных значениях объясняющих переменных (предсказание среднего значения), либо прогнозируется конкретное значение зависимой переменной (предсказание конкретного значения).

Доверительный интервал y (зависимой переменной) также называется стандартной ошибкой предсказания среднего [standard error of mean prediction]. Примерно в 95 случаях из 100 истинное среднее y будет находиться внутри границ доверия вокруг наблюдаемого среднего из n выборочных оценок. Следует отметить, что доверительный интервал y относится к среднему, а не к отдельному случаю y. Кроме того, доверительный интервал уже, чем интервал предсказания, который относится к отдельным случаям. Доверительный интервал для значений зависимой переменной: строится для каждого значения X, причём наименьшая ошибка получается для среднего Y. Ширина доверительного интервала зависит от нескольких факторов. При заданном уровне значимости возрастание амплитуды колебаний вокруг линии регрессии, измеренное с помощью среднеквадратичной ошибки, приводит к увеличению ширины интервала. С другой стороны, увеличение объема выборки сопровождается сужением интервала. Кроме того, ширина интервала изменяется в зависимости от значений X_i. Если значение переменной Y предсказывается для величин X, близких к среднему значению $\overline{X}$, доверительный интервал оказывается уже, чем при прогнозировании отклика для значений, далеких от среднего.

Команда в R:

```
predict (mod, interval="confidence")
```

Интервал предсказания y. Примерно в 95 случаях из 100 случай с данными значениями независимых переменных будет находиться внутри рассчитанных границ предсказания. Интервал предсказания будет шире (менее определенным), чем доверительный интервал, поскольку он имеет дело с оценкой интервалов отдельных случаев, а не средних значений (изменчивость при прогнозировании индивидуальных значений намного больше, чем при оценке математического ожидания).

Команда в R:
predict (mod, interval="prediction")

6. Построить на одном графике точечный график пар точек (x_i, y_i) $\left(i = \overline{1,n}\right)$, линию регрессии $y = \beta_0 + \beta_1 \cdot x$, доверительные интервалы для линии регрессии и доверительные интервалы для предсказанных значений. Эта информация наглядно представляет результаты подгонки модели линейной регрессии.

Задаем новые значения *x*
> pred.frame <− data.frame (x= seq(5, 20,by =1))

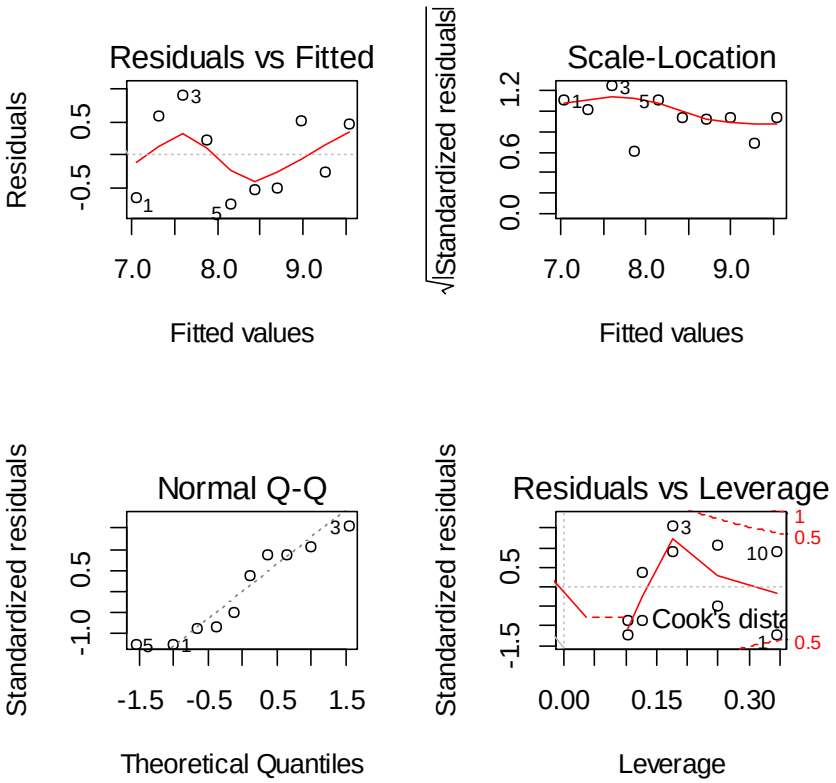
Вычисляем доверительные интервалы для новых значений *x*
> pc <− predict(mod, int="c", newdata=pred.frame)
> pp <− predict(mod, int="p", newdata=pred.frame)

Создаем график и наносим на него экспериментальные точки
plot (x, y, cex = 1.2, pch = 19, col = 'gray', xlim=c(4,21))

добавляем интервалы
matlines (pred.x, pc, lty = 2, lwd = 2, col ='blue')
matlines (pred.x, pp, lty = 2, lwd = 2, col ='red')

добавляем линию модели
abline(mod, col="black", lwd = 2)

7. Анализ модели
#Создание окна, в котором будут выводиться 4 графика (2 – вверху, 2 - внизу)
layout(matrix(1:4,2,2))
графики модели
plot(mod)

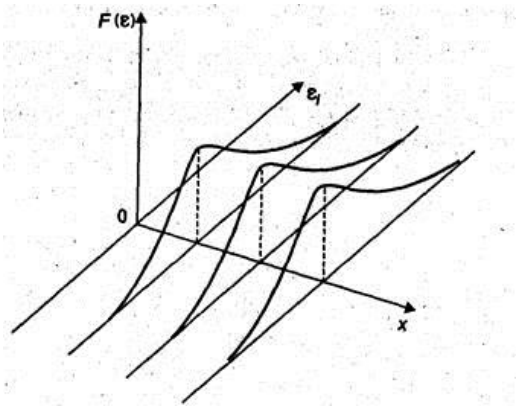


1 график - Точечный график предсказанных значений и остатков позволяет обнаружить закономерности распределения последних. Остатки появляются на оси y, а установленные значения отображаются на оси x. Если линейная модель является точной, разности, или остатки, будут носить случайный характер и их сумма будет близка к нулю. К графику добавлена сглаживающая линия, облегчающая выявление закономерностей в расположении остатков. Примерно горизонтальное расположение этой линии указывает на отсутствие каких-либо проблем.

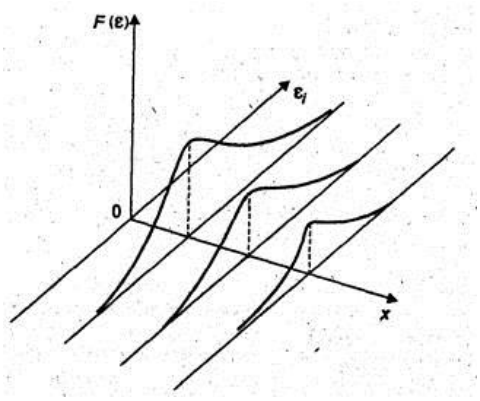
2- График квантилей (Q-Q) помогает определить нормальность остатков. Остатки следуют прямой линии или они сильно отклоняются? Хорошо, если остатки хорошо выровнены по прямой пунктирной линии.

3 - График “Разброс-положение” даёт возможность оценить монотонность распределения остатков (гетероскедастичность). Хорошо, если вы видите горизонтальную линию с одинаковыми (случайными) точками распространения

Гетероскедастичность (англ. *heteroscedasticity*) — понятие, означающее неоднородность наблюдений, выражающуюся в неодинаковой (непостоянной) дисперсии случайной ошибки регрессионной модели. Гетероскедастичность противоположна *гомоскедастичности*, означающей однородность наблюдений, то есть постоянство дисперсии случайных ошибок модели.



Гомоскедастичность остатков



Гетероскедастичность остатков

4. - С помощью графика “Остатки-влияние” можно определить точки с большим влиянием на модель. Этот сюжет помогает нам найти влиятельные случаи, если таковые имеются. Не все выбросы влияют на линейный регрессионный анализ (независимо от уровня выбросов). Несмотря на то, что данные имеют экстремальные значения, они могут не влиять на определение линии регрессии. Это означает, что результаты не будут сильно отличаться, если мы включим или исключим их из анализа. В большинстве случаев они следуют тенденции, и они не имеют большого значения; они не влияют. С другой стороны, некоторые случаи могут быть очень влиятельными, даже если они выглядят в пределах разумного диапазона значений. Они могут быть крайними случаями против линии регрессии и могут изменять результаты, если мы исключаем их из анализа. Другой способ сказать, что они не справляются с тенденцией в большинстве случаев. Мы следим за значениями в верхнем правом углу или в правом нижнем углу. Эти пятна - места, где случаи могут влиять на линию регрессии. Ищите случаи за пределами пунктирной линии, расстояние Кука. Когда случаи находятся за пределами расстояния Кука (что означает, что они имеют высокие оценки Кука), случаи влияют на результаты регрессии. Результаты регрессии будут изменены, если мы исключим эти случаи.

Код модели линейной регрессии

```
x<−c(7,8,9,10,11,12,13,14,15,16)
y<−c(6.4,7.9,8.5,8.1,7.4,7.9,8.2,9.5,9.0,10.0)
x
y
plot(x,y)
mod<−lm(y~x)
summary(mod)
coef_mod<−coef(mod)
coef_mod
```

```
b0<-coef_mod[1]
b0
b1<-coef_mod[2]
b1
#new data
pred.frame <− data.frame (x= seq(5, 20,by =1))
#Confidence bands
pc <−predict(mod, int="c", newdata=pred.frame)
# Prediction bands
pp <− predict(mod, int="p", newdata=pred.frame)
plot (x, y, cex = 1.2,pch = 19, col = 'gray', xlim=c(4,21))
matlines (pred.x, pc, lty = 2, lwd = 2, col ='blue')
matlines (pred.x, pp, lty = 2, lwd = 2, col ='red')
abline(mod, col="black", lwd = 2)
layout(matrix(1:4,2,2))
plot(mod)
```