Линейная регрессия

Модель

Данные: пары чисел $\{ x_i , y_i \}$

Гипотеза: имеется линейная статистическая зависимость между переменными X и Y

$$Y = a + b \cdot X + \varepsilon$$

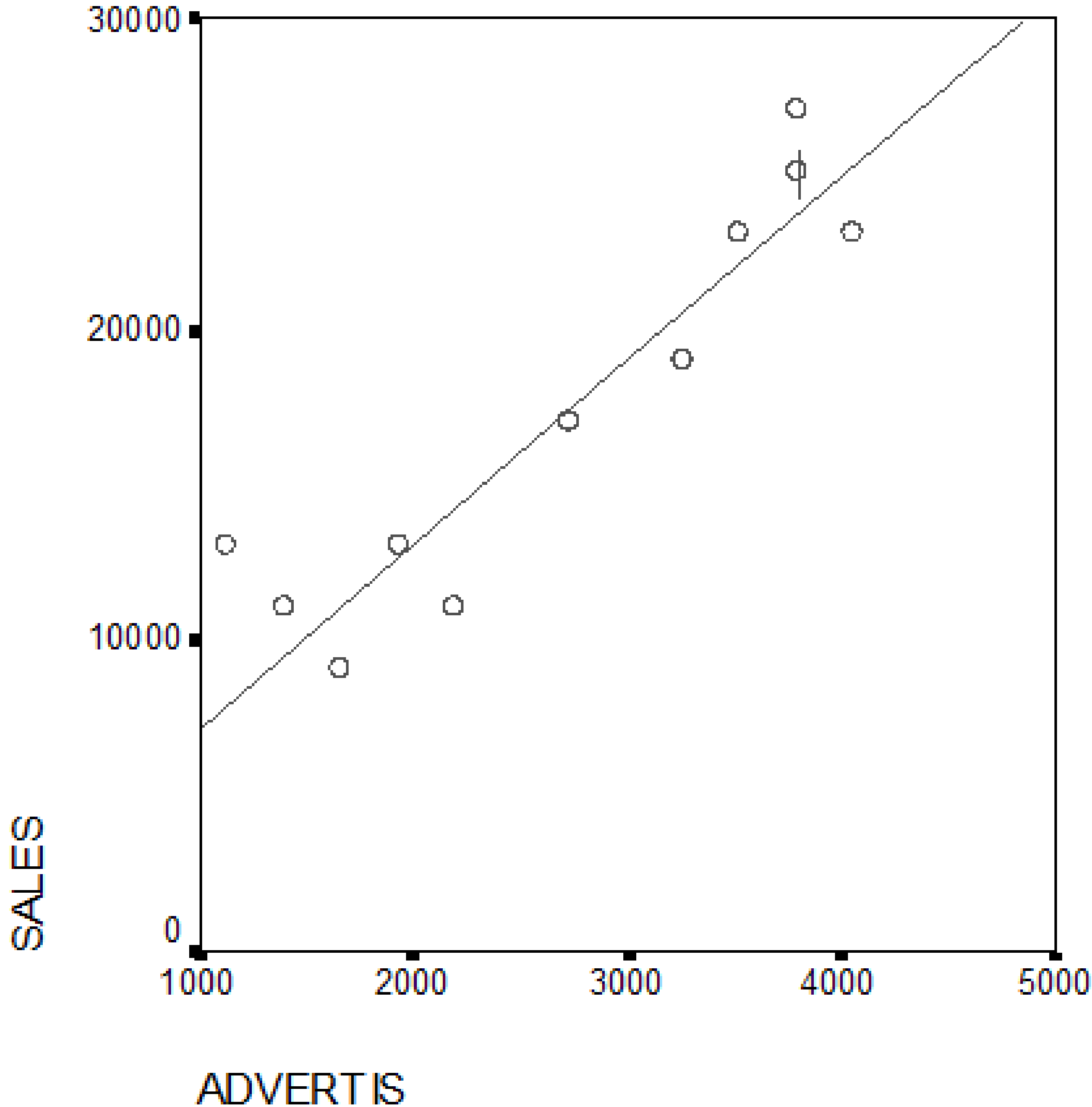
Надо найти

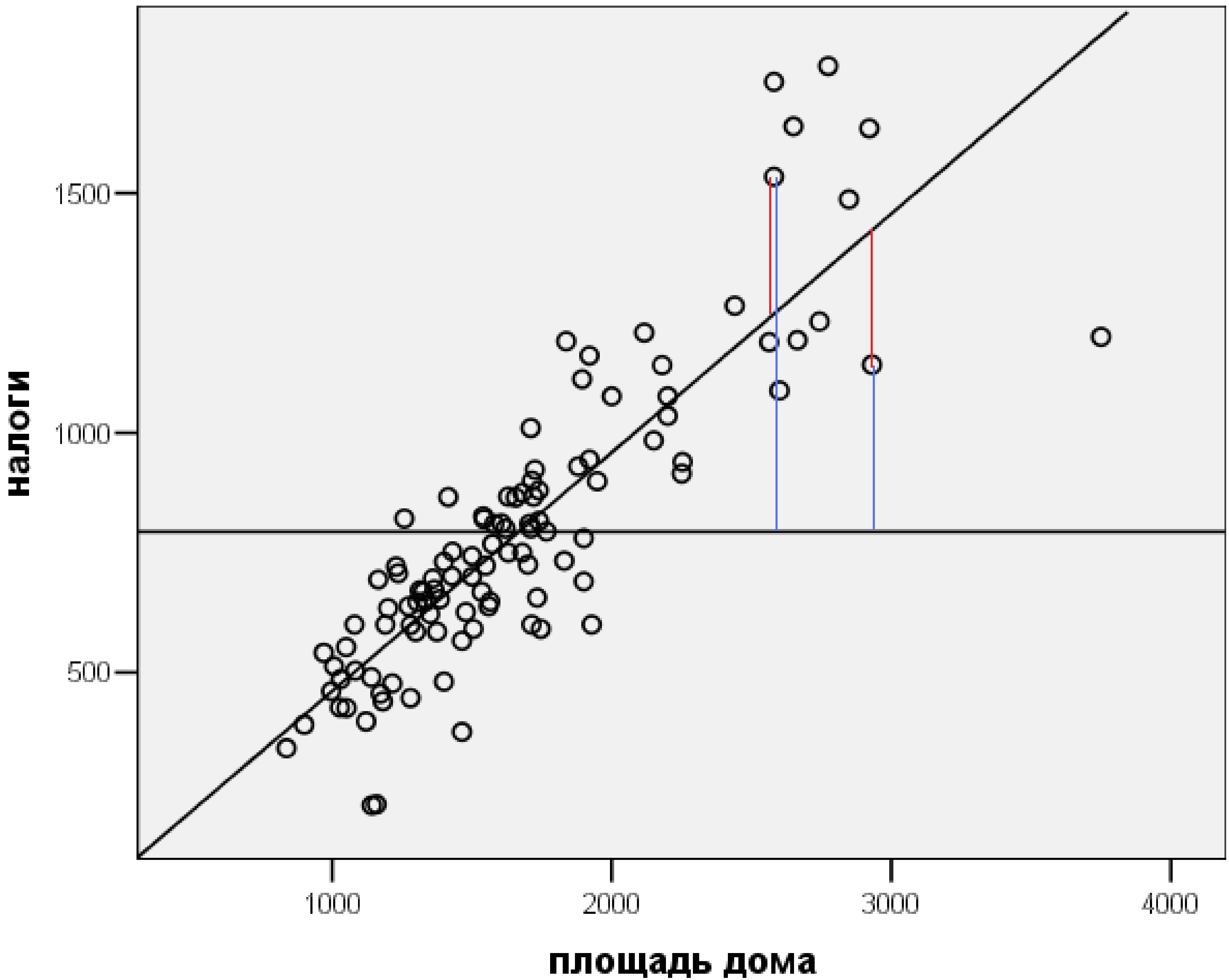
оценки коэффициентов *a* и *b* уравнения регрессии:

$$Y = a + b \cdot X$$

Геометрический взгляд

Уравнение определяет прямую, наиболее близко проходящую ко всем точкам с координатами x_i ,y_i


30000
20000
10000
0
SALES
1000
2000
3000
4000
5000
ADVERTIS



1500
1000
500
1000
2000
3000
4000
налоги
площадь дома

Алгебраический взгляд

Значения *a* и *b* находятся по методу наименьших квадратов, т.е. так, чтобы минимизировать величину

$$SSE=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-(a+b\cdot x_i)\right)^2$$

Терминология

- Y — отклик / зависимая переменная
- X — предиктор / независимая переменная
- $\varepsilon_i = y_i - (a + b \cdot x_i)$ - ошибка / невязка /
-

Измеряем качество модели

Регрессионная модель

Сумма квадратов отклонений для регрессионной модели

Зависит от единиц измерения

$$SSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-(a+b\cdot x_i))^2$$

Измеряем качество модели

Базовая модель

Сумма квадратов отклонений для базовой модели

Зависит от единиц измерения

$$SSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

Коэффициент детерминации

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - (a + b \cdot x_i)\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}_n\right)^2}$$

Коэффициент детерминации

указывает, какой процент вариации у объясняется влиянием предикторов.

Недостатки множественного коэффициента детерминации

Пример «Квартет Анскомба»

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician 27 (1): 17–21.

Квартет Анскомба: 4 набора данных

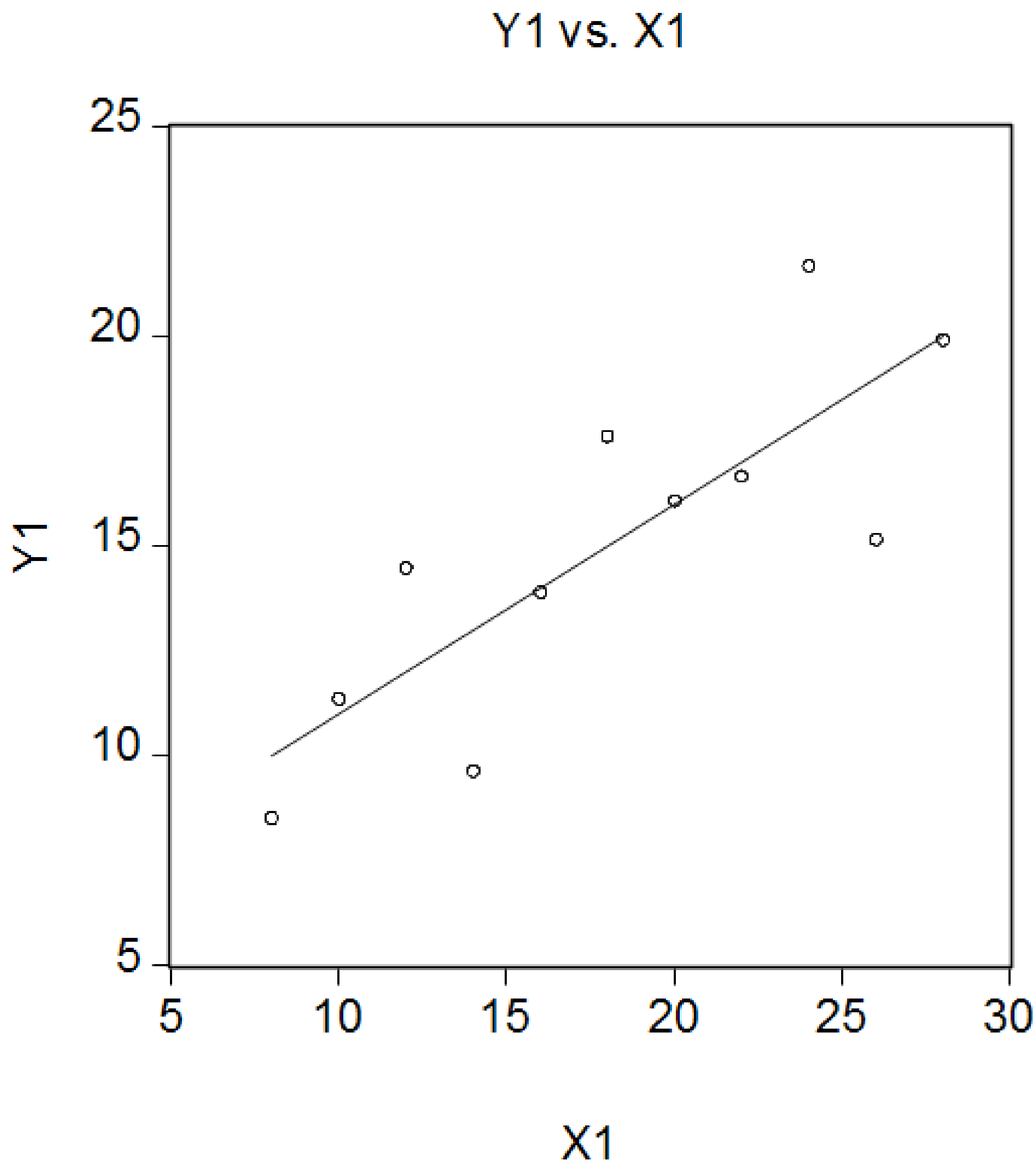
Характеристика	Значение
Mean of x in each case	9
Sample variance of x in each case	11
Mean of y in each case	7.50
Sample variance of y in each case	4.122 or 4.127
Correlation between x and y in each case	0.816
Linear regression line in each case	y = 3.00 + 0.500x

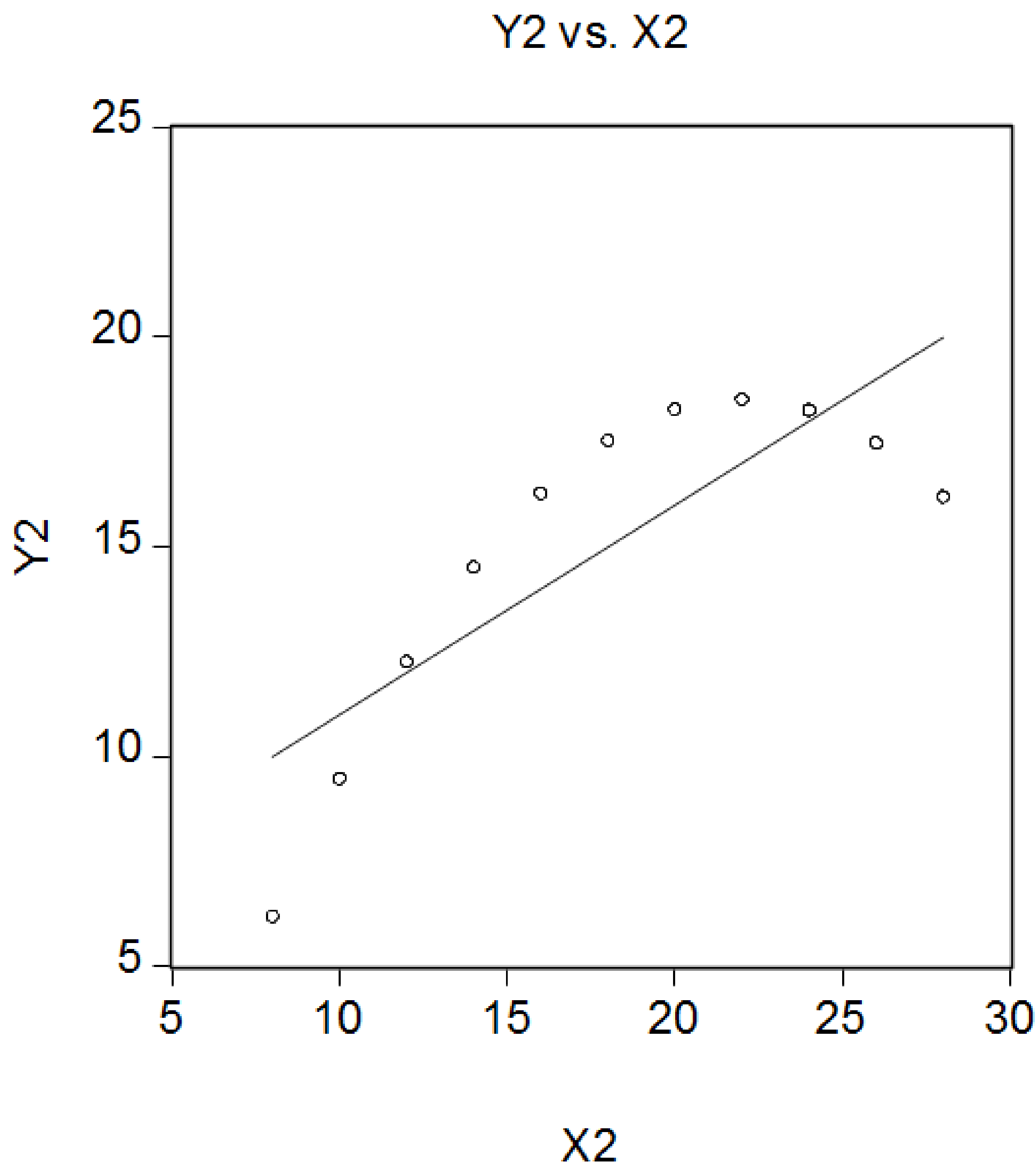
Квартет Анскомба в R

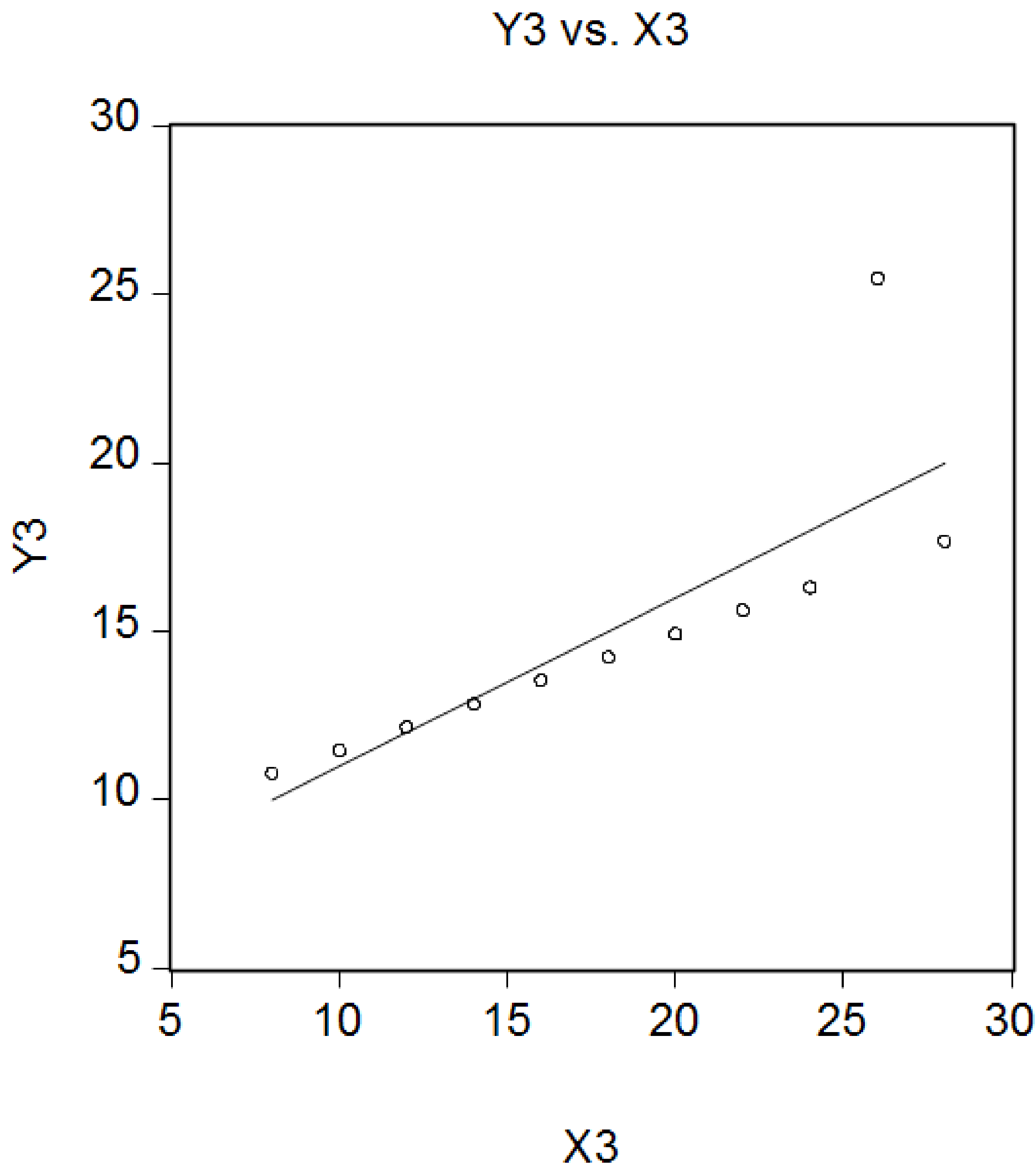
- help(anscombe)

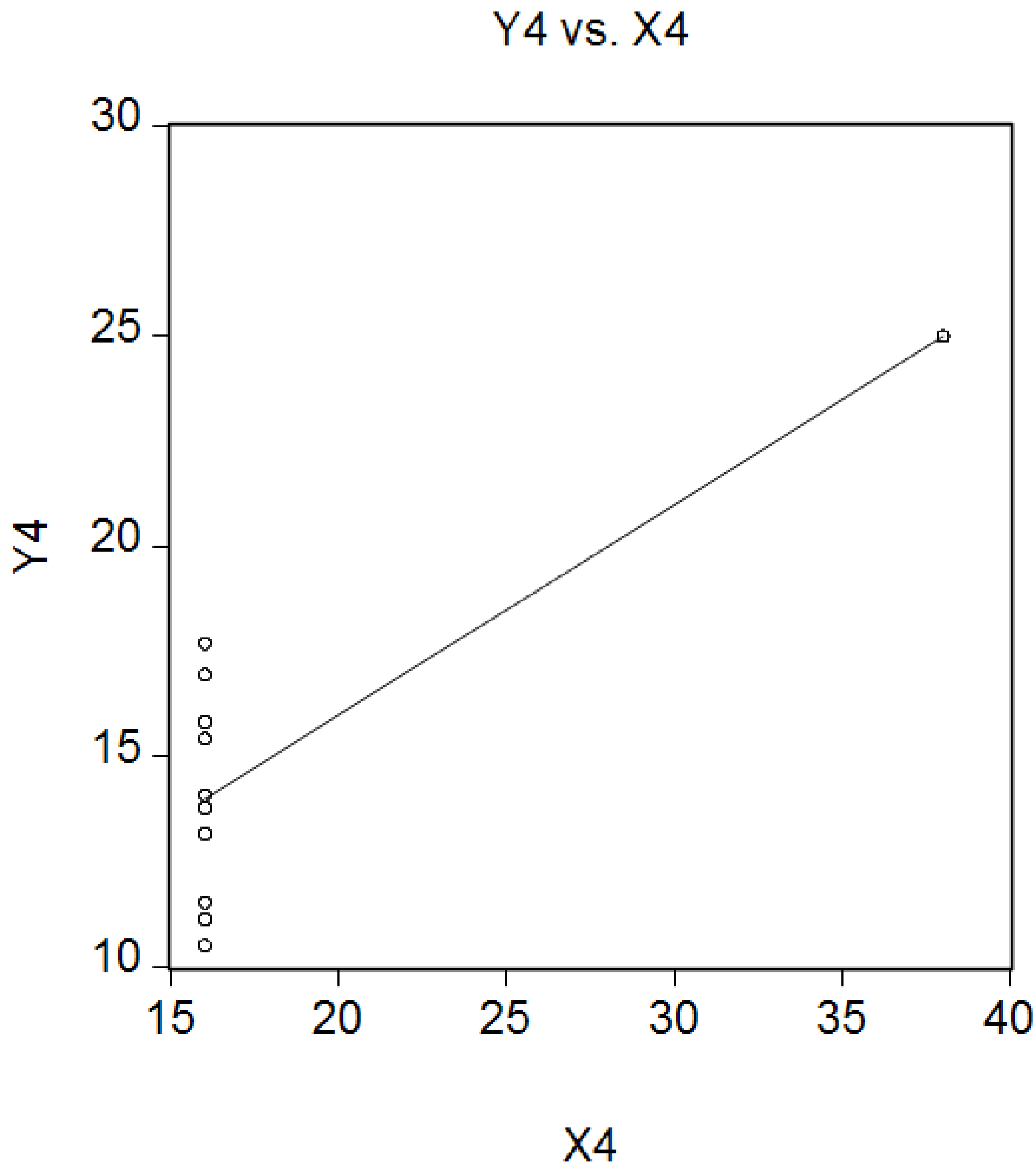
```
ans <- anscombe
res1 <- lm(ans$y1~ans$x1)
plot(ans$x1,ans$y1)
abline(res1)
```

-
-


Y1 vs. X1
25
20
15
10
5
Y1
5
10
15
20
25
30
X1


Y2 vs. X2
25
20
15
10
5
Y2
5
10
15
20
25
30
X2


Y3 vs. X3
Y3
X3
30
25
20
15
10
5
5
10
15
20
25
30


Y4 vs. X4
30
25
20
15
10
Y4
15
20
25
30
35
40
X4

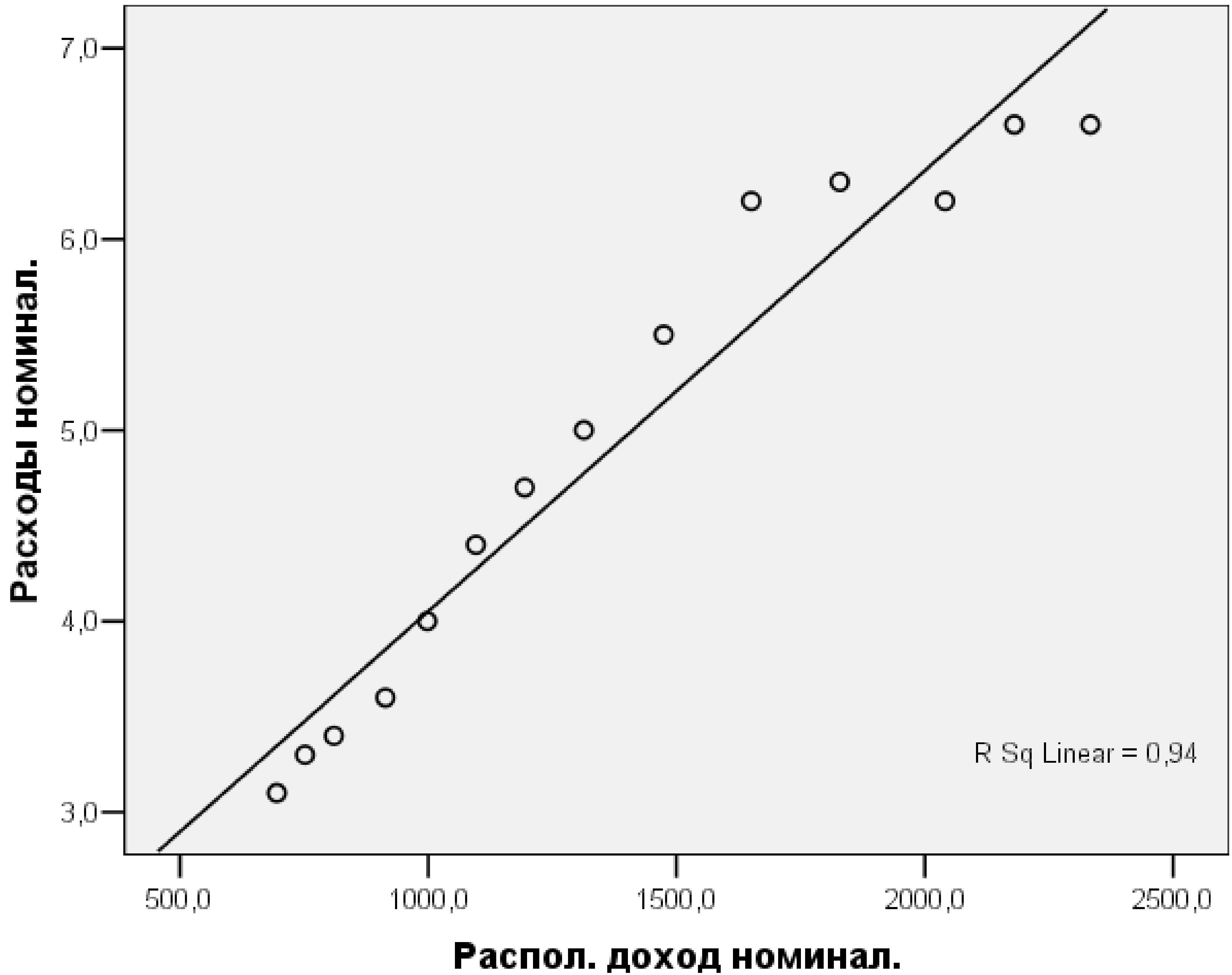
Влияние инфляции на модель

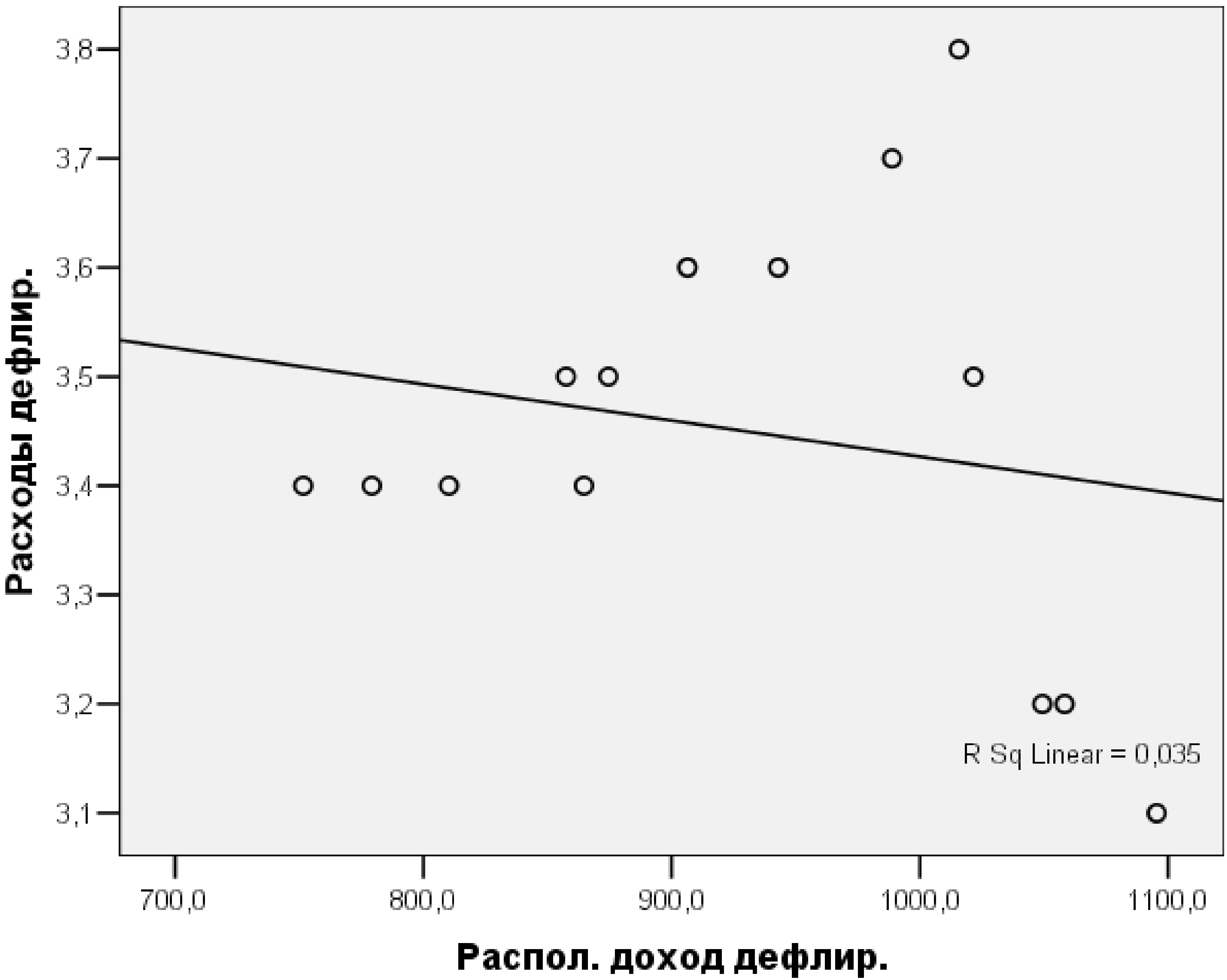
Пример

совокупный располагаемый доход и
совокупные личные расходы на местный транспорт
в США за период с 1970 по 1983 год.

Данные представлены как в текущих долларах США, так и в долларах 1972 года — пересчет к последним выполнен с учетом динамики индекса потребительских цен в указанном периоде.

(Уровень цен в 1972 г. принят за 100%.)


Расходы номинал.
7,0
6,0
5,0
4,0
3,0
R Sq Linear = 0,94
500,0
1000,0
1500,0
2000,0
2500,0
Распол. доход номинал.



3,8
3,7
3,6
3,5
3,4
3,3
3,2
3,1
Расходы дефлир.
R Sq Linear = 0,035
700,0
800,0
900,0
1000,0
1100,0
Распол. доход дефлир.

Коллинеарность-1

Допустим, что задана зависимость

$$Z = 6X + 8Y+5$$

Одновременно известно, что

$$Y = 2X$$

Коллинеарность-2

Если верно $Z = 6X + 8Y+5$,

То верно

$Z = 2X + 10Y+5$

$Z = 22X + 0*Y+5$

$Z = 0*X + 11*Y+5$

И так далее…

(На самом деле все еще хуже...)

tolerance
пороговые значения 0.1 или 0.2

$$tolerance = 1 - R_j^2$$

VIF variance inflation factor

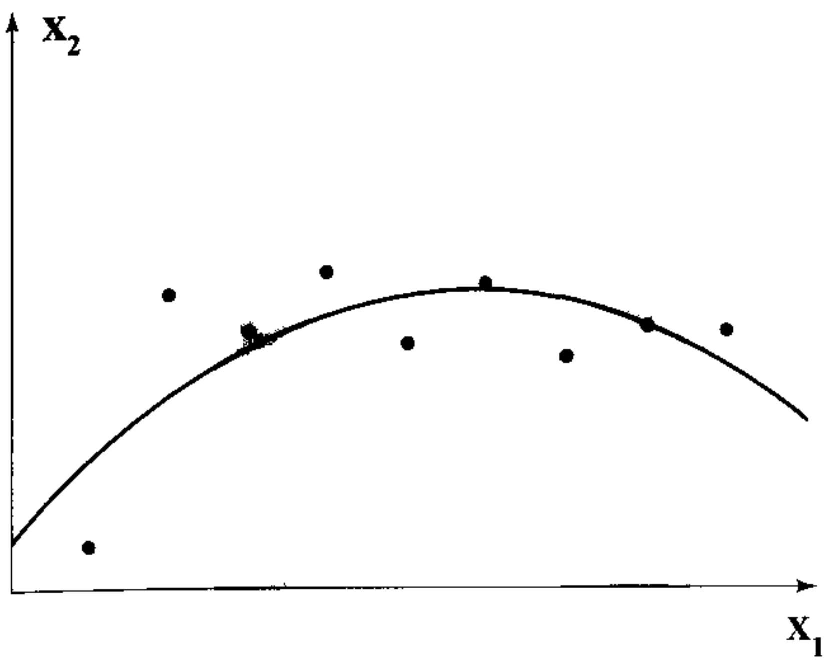
пороговые значения 5 или 10

$$VIF = \frac{1}{tolerance}$$

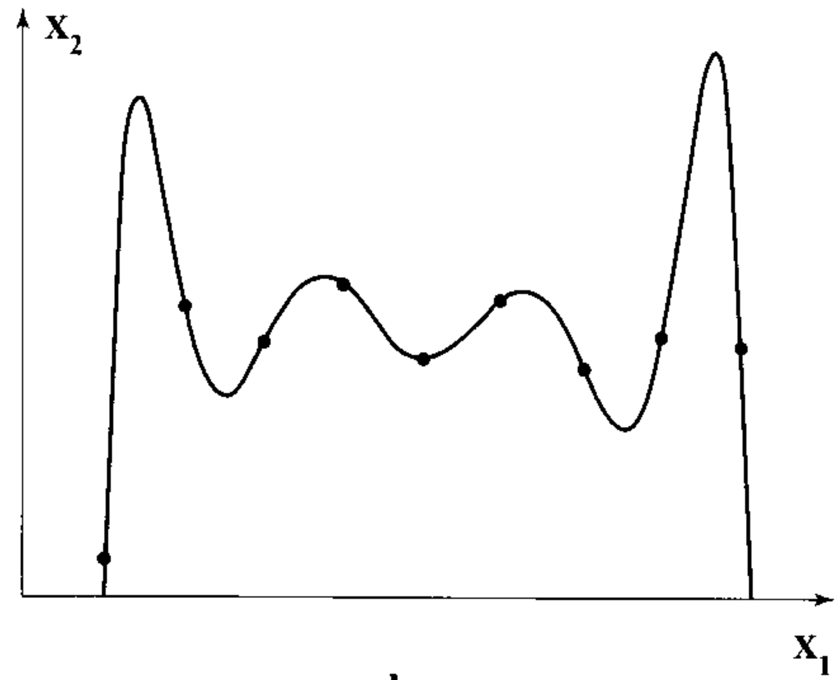
Вычисления VIF в R

- library(car)
- # пример Albuquerque Home Prices
- vif(fit)   # variance inflation factors
- vif(itog1)
x$SQFT	x$AGE	x$FEATS	x$NE	x$CUST	x$COR	x$TAX
6.197	1.692	1.459	1.374	1.385	1.108	6.476

Переобучение



a



b

Non-independence of Errors

-
-
- #   Test for Autocorrelated Errors
- durbinWatsonTest(fit)