Регрессио́нный анализ. Примеры применения регрессионного анализа.

- Моделирование потоков миграции в зависимости от таких факторов как средний уровень зарплат, наличие медицинских, школьных учреждений, географическое положение и др.
- Моделирование дорожных аварий как функции скорости, дорожных условий, погоды и т.д.
- Моделирование потерь от пожаров как функции от таких переменных как количество пожарных станций, время обработки вызова, или цена собственности.

Суть регрессионного анализа заключается в нахождении наиболее важных факторов, которые влияют на зависимую переменную!

Регрессионный анализ используют для решения задач:

- **Установления формы зависимости между переменными** (линейная-нелинейная, отрицательная-положительная и т.д.).

- **Определения функции регрессии.**
  Важно выяснить, каково было бы действие на зависимую переменную главных факторов, если бы прочие факторы не изменялись и если бы были исключены случайные элементы.

Цель регрессионного анализа - по значениям одной переменной, выбранной в качестве аргумента, предсказать соответствующее значение другой (функции).

Регрессио́нный анализ

Регрессио́нный анализ - статистический метод исследования влияния одной или нескольких независимых переменных $X_1, X_2, \ldots, X_i$ на зависимую переменную Y.

Уравнение регрессии - это математическая формула, применяемая к независимым переменным, чтобы лучше спрогнозировать зависимую переменную, которую необходимо смоделировать $y = f(x_1, x_2, \ldots, x_i) + \varepsilon$

f - заранее не известная функция, подлежащая определению;

ε - ошибка аппроксимации данных.

Уравнение множественной линейной регрессии

$$y = a_0 + b_1x_1 + b_2x_2, \ldots + b_ix_i$$

Измерение экспериментальных данных

Зависимая переменная (Y) - это переменная, описывающая процесс, который мы пытаемся предсказать или понять.

Независимые переменные (X) это переменные, используемые для моделирования или прогнозирования значений зависимых переменных.

Существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как случайные ошибки ε.

Линейная регрессия

В линейной регрессионной модели функция имеет вид $f(x_i) = a + bx_i$

А сама модель имеет следующий вид:

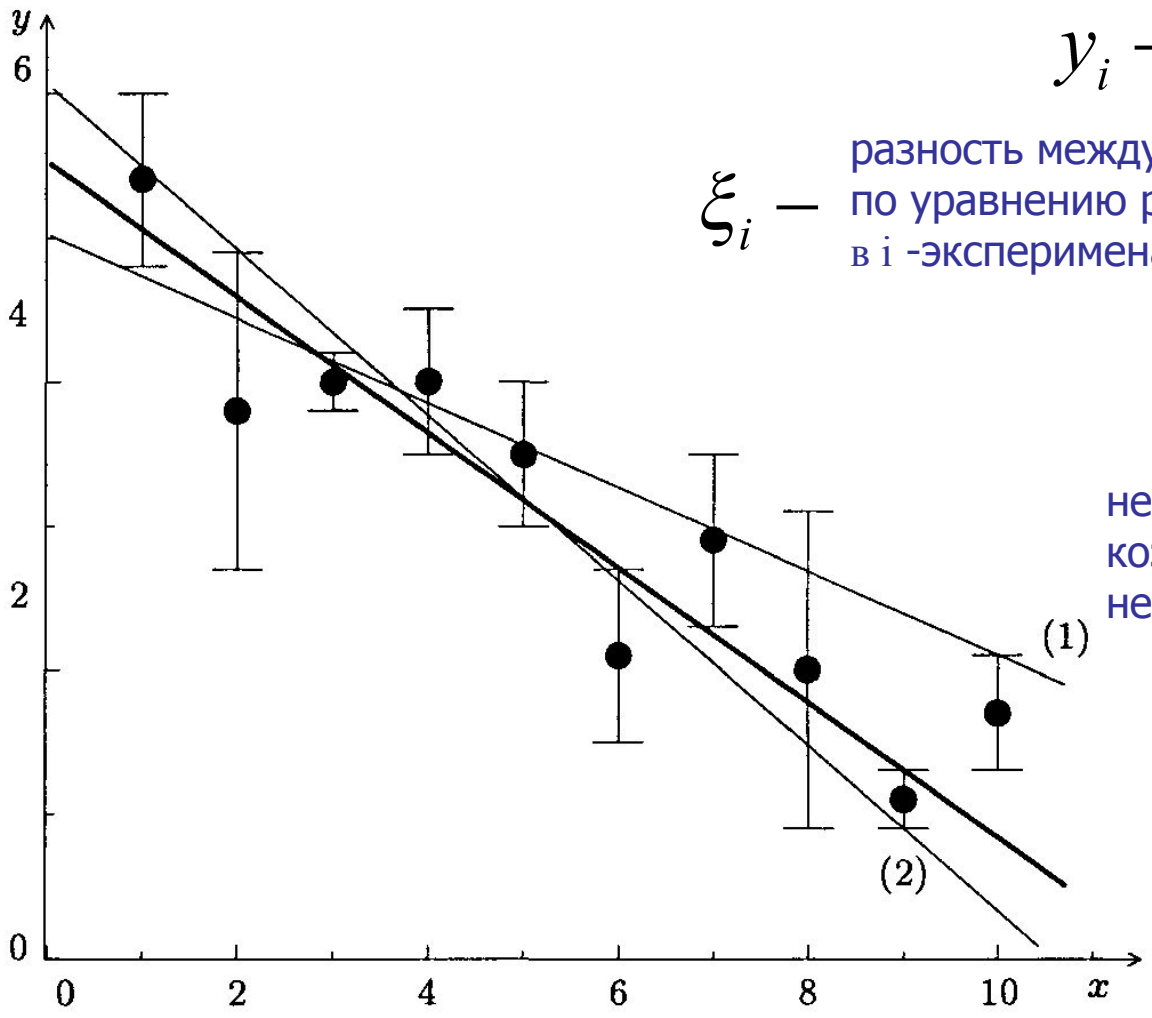
$$y_i = a + bx_i$$

При данных a и b **сумма квадратов отклонений** экспериментальных данных от найденной прямой **будет наименьшей.**

Расчет коэффициентов регрессии (МНК)

$$y_i - a - bx_i = 0$$

все экспериментальные точки лежат строго на прямой линии



$$y_i - a - bx_i = \xi_i$$

ξ_i — разность между экспериментальным и вычисленным по уравнению регрессии значениями величины y в i -эксперименальной точке (**невязка**)

$$U = \sum_{i=1}^{n} \xi_i^2 = \min$$

необходимо найти такие коэффициенты регрессии, при которых невязки будут минимальны

Расчет коэффициентов регрессии (МНК)

$$U = \sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \min$$

$$\frac{dU}{da} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\frac{dU}{db} = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

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

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

Коэффициент корреляции Пирсона

$$R = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} \left((x_i - \bar{x})(y_i - \bar{y}) \right)}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Условия применения коэффициента корреляции Пирсона:

1. Переменные x и y должны быть распределены нормально.
2. Переменные x и y должны быть измерены в интервальной шкале или шкале отношений.
3. Количество значений в исследуемых переменных x и y должно быть одинаковым.

Значение коэффициента корреляции не зависит от масштаба измерения.

Коэффициент корреляции Пирсона

Коэффициент корреляции принимает значения от **–1,0** (строгая отрицательная корреляция) до **+1,0** (строгая положительная корреляция). Значение 0,0 означает отсутствие корреляции.

Связи между переменными могут быть слабыми и сильными (тесными). Их критерии можно оценивать по различным шкалам, из которых наиболее часто применяют шкалы Чеддока и Е.П.Голубкова

Шкала Чеддока		**Шкала Е.П.Голубкова**	
R	**Интерпретация**	**R**	**Интерпретация**
0,1 – 0,3	Слабая	0,00 - 0,20	Отсутствует
0,3 – 0,5	Умеренная	0,21 - 0,40	Очень слабая
0,5 – 0,7	Заметная	0,41 - 0,60	Слабая
0,7 – 0,9	Высокая	0,61 - 0,80	Умеренная
0,9 – 1,0	Весьма высокая	0,81 - 1,00	Сильная

Расчет СКО найденных коэффициентов a и b в уравнении

$$\sigma_0 = \sqrt{\frac{1}{n(n-2)} \sum_{i=1}^{n} (y_i - a - bx_i)^2}$$

$$\sigma_a = \sigma_0 \sqrt{\frac{n \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}}$$

$$\sigma_b = \sigma_0 \sqrt{\frac{n}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}}$$

$$\Delta a = \frac{\sigma_a \, t_{P,n-1}}{\sqrt{n}}$$

$$\Delta b = \frac{\sigma_b \, t_{P,n-1}}{\sqrt{n}}$$

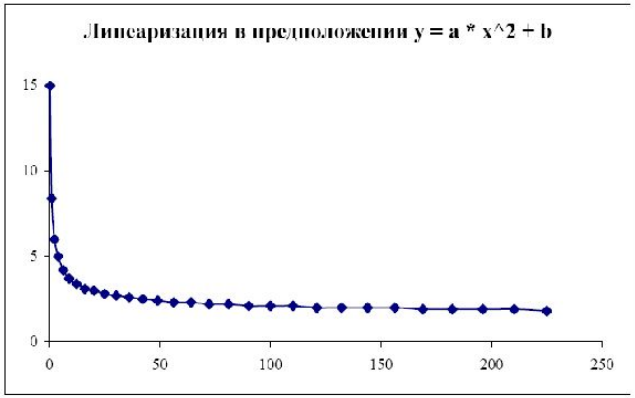
Анализ нелинейных зависимостей. Линеаризация зависимостей

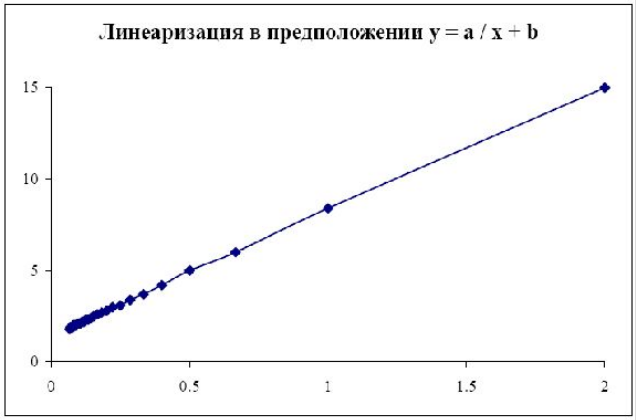
Вид зависимости	Замены				Ограничения
	v	u	k	z	
Парабола второго (или высшего) порядка $y = ax^2 + b$	y	x^2	a	b	
Гипербола $y = \frac{a}{x} + b$	y	$\frac{1}{x}$	a	b	$x \neq 0$
Логарифмическая функция $y = a \ln x + b$	y	$\ln x$	a	b	$x > 0$
Показательная функция $y = ba^x$	$\ln y$	x	$\ln a$	$\ln b$	$y > 0$ $a > 0$ $b > 0$
Степенная функция $y = bx^a$	$\ln y$	$\ln x$	a	$\ln b$	$y > 0$ $x > 0$ $b > 0$
Экспоненциальная функция $y = be^{ax}$	$\ln y$	x	a	$\ln b$	$y > 0$ $b > 0$
$y = \frac{x}{ax + b}$	$\frac{1}{y}$	$\frac{1}{x}$	a	b	$y \neq 0$ $x \neq 0$

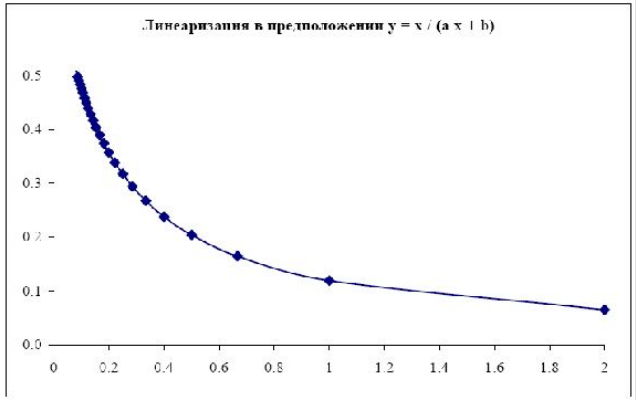
Задание 5. Вариант 47

Исходные данные

0,5	15,0
1	8,4
1,5	6,0
2	5,0
2,5	4,2
3	3,7
3,5	3,4
4	3,1
4,5	3,0
5	2,8
5,5	2,7
6	2,6
6,5	2,5
7	2,4
7,5	2,3
8	2,3
8,5	2,2
9	2,2
9,5	2,1
10	2,1
10,5	2,1
11	2,0
11,5	2,0
12	2,0
12,5	2,0
13	1,9
13,5	1,9
14	1,9
14,5	1,9
15	1,8







Расчет коэффициента корреляции Пирсона

$$R = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

представление	R
$y = a x^2 + b$	-0,5126
$y = a / x + b$	0,9998
$y = \frac{x}{a x + b}$	-0,8259

Наблюдается сильная корреляция экспериментальных данных в представлении $y = a / x + b$

Расчет коэффициентов регрессии

$$a = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 6{,}8405 \qquad b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 1{,}4281$$

Расчет среднеквадратичных отклонений

$$\sigma_0 = \sqrt{\frac{1}{n(n-2)}\sum_{i=1}^{n}(y_i - b - a x_i)^2} = 0{,}0098$$

$$\sigma_a = \sigma_0 \sqrt{\frac{n}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}} = 0{,}0047$$

$$\sigma_b = \sigma_0 \sqrt{\frac{n\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}} = 0{,}0120$$

Доверительный интервал коэффициентов регрессии при P=0,95

$$\Delta a = \frac{\sigma_a t_{P,n-1}}{\sqrt{n}} = 0{,}0018 \qquad \Delta b = \frac{\sigma_b t_{P,n-1}}{\sqrt{n}} = 0{,}0045$$

Уравнение регрессии

$$y = \frac{(a \pm \Delta a)}{x} + (b \pm \Delta b)$$