Основы анализа данных. Регрессионный анализ

Лекция 6

КМАИ

Секции

Определения, термины и примеры применения

Виды регрессионного анализа

Коэффициенты регрессии и детерминации

Линейная регрессия на корреляции

Примеры применения

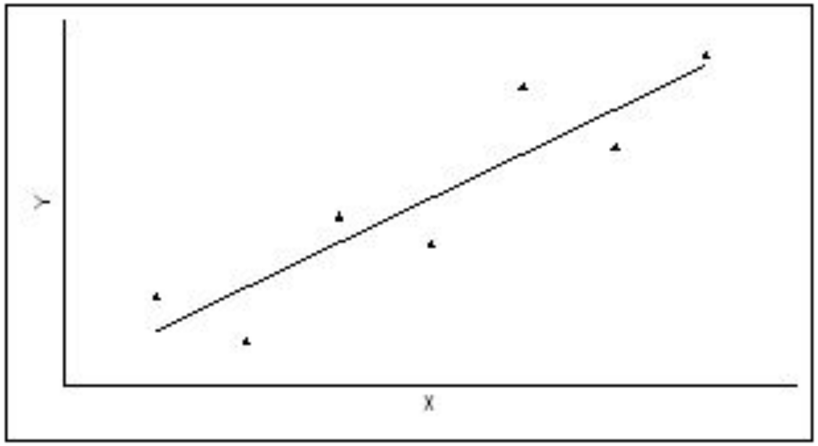
1. Моделирование числа поступивших в университет для лучшего понимания факторов, удерживающих детей в том же учебном заведении.

2. Моделирование потоков миграции в зависимости от таких факторов как средний уровень зарплат, наличие медицинских, школьных учреждений, географическое положение.

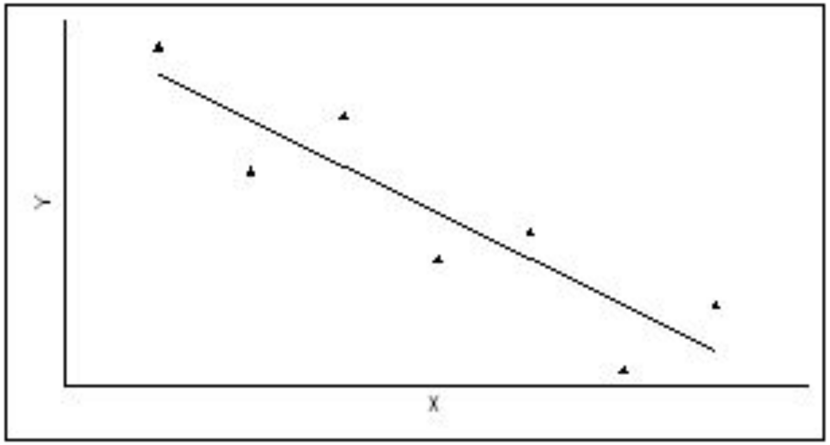
3. Моделирование дорожных аварий как функции скорости, дорожных условий, погоды и т.д.,

4. Моделирование потерь от пожаров как функции от таких переменных как количество пожарных станций, время обработки вызова, или цена собственности.

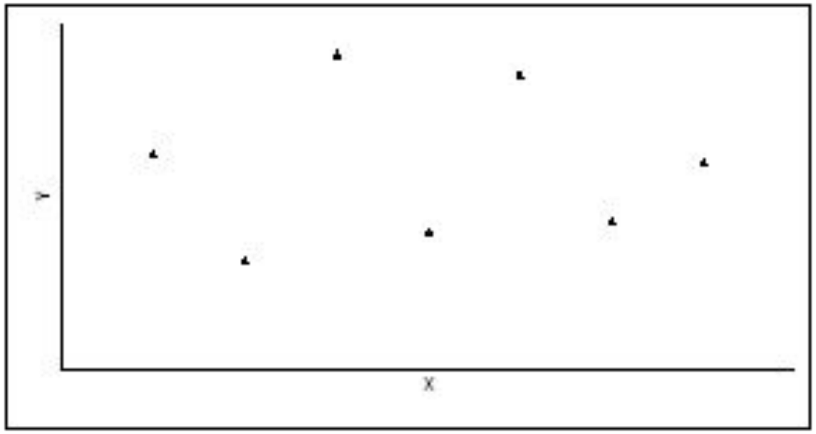
Связь между переменными



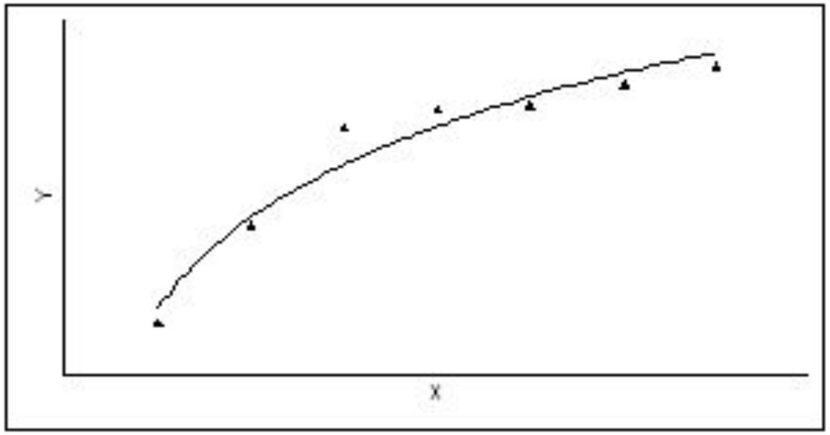
Линейная положительная связь



Линейная отрицательная связь



Связь отсутствует



Нелинейная связь

Определения

Регрессионный анализ — статистический метод исследования влияния одной или нескольких независимых переменных X 1 , X 2 , . . . , X p на зависимую переменную Y .

Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные — критериальными.

Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

$$Y = F\left(X_1, X_2, X_3, \dots, X_p\right) + \varepsilon$$

Цели регрессионного анализа

1. Определение степени детерминированности вариации критериальной (зависимой) переменной предикторами (независимыми переменными).

2. Предсказание значения зависимой переменной с помощью независимой(-ых).

3. Определение вклада отдельных независимых переменных в вариацию зависимой.

$$Y = F\left(X_1, X_2, X_3, \ldots, X_p\right) + \varepsilon$$

Регрессионный анализ нельзя использовать для определения наличия связи между переменными, поскольку наличие такой связи и есть предпосылка для применения анализа.

Термины

Уравнение регрессии Математическая формула, применяемая к независимым переменным, чтобы лучше спрогнозировать зависимую переменную, которую необходимо моделировать.

$$y_i = \underbrace{a_1 x_{1i} + a_2 x_{2i} + \cdots + a_n x_{ni}}_{\text{Неслучайная часть}} + \underbrace{a_{n+1}}_{\text{Свободный коэф.}} + \underbrace{\varepsilon_i}_{\text{Случайный остаток}}$$

$a_1, \dots, a_{n+1}$ - Коэффициенты регрессии

$x_1, \dots, x_n$ - Зависимые переменные

ε - Ошибка регрессии

Термины

Формирование уравнения регрессии – процедура минимизации случайного остатка.

$$y_i = a_1 x_{1i} + a_2 x_{2i} + \cdots + a_n x_{ni} + a_{n+1} + \varepsilon_i$$

Выделение зависимых переменных

Выделение смещения мат. ожидания

$$\sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2 \rightarrow min$$

Определения

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.

Коэффициенты регрессии (a) — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

Невязки — существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как случайные ошибки **ε**

Секции

Определения, термины
и примеры применения

Виды регрессионного анализа

Коэффициенты регрессии
и детерминации

Линейная регрессия на корреляции

Виды регрессионного анализа

	Линейные по переменным	Не линейные по переменным
Линейные по параметрам	$y = ax + b + \varepsilon$	$y = ax^2 + b + \varepsilon$ $y = \frac{a}{x} + b + \varepsilon$
Не линейные по параметрам	$y = ax^c + b + \varepsilon$ $y = ab^x + \varepsilon$	$y = \frac{ax^{2c} + b}{c^x} + \varepsilon$ Не надо использовать

Линейный регрессионный анализ

$$y_i = a_1 x_{1i} + a_2 x_{2i} + \cdots + a_n x_{ni} + a_{n+1} + \varepsilon_i$$

Запись в матричной форме: $Y = AX + U$

$$Y = \begin{bmatrix} y_1 \\ \cdots \\ y_M \end{bmatrix} \quad X = \begin{bmatrix} x_{11} & \cdots & x_{n1} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{1M} & \cdots & x_{nM} & 1 \end{bmatrix} \quad \boxed{A = \begin{bmatrix} a_1 \\ \cdots \\ a_{n+1} \end{bmatrix}}$$

Оценка параметров:

$$Y = AX + U \longrightarrow A = X^{-1}Y$$

Получаем в явном виде набор уравнений:

$$\begin{bmatrix} y_1 \\ \dots \\ y_M \end{bmatrix} \cdot \begin{bmatrix} x_{11} & \cdots & x_{n1} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{1M} & \cdots & x_{nM} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} a_1 \\ \dots \\ a_{n+1} \end{bmatrix}$$

МНК

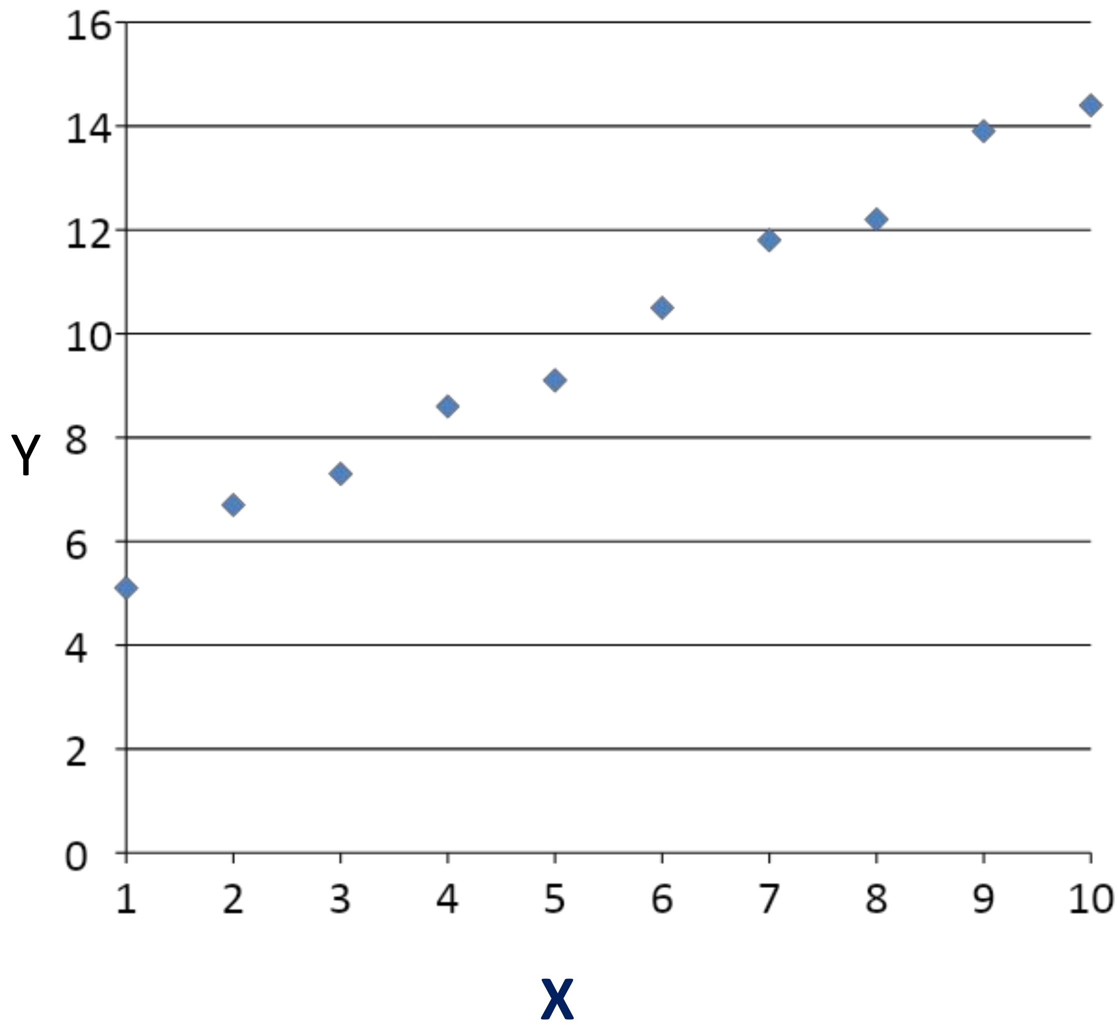
:

$$A = (X^T X)^{-1} X^T Y$$

Линейный регрессионный анализ

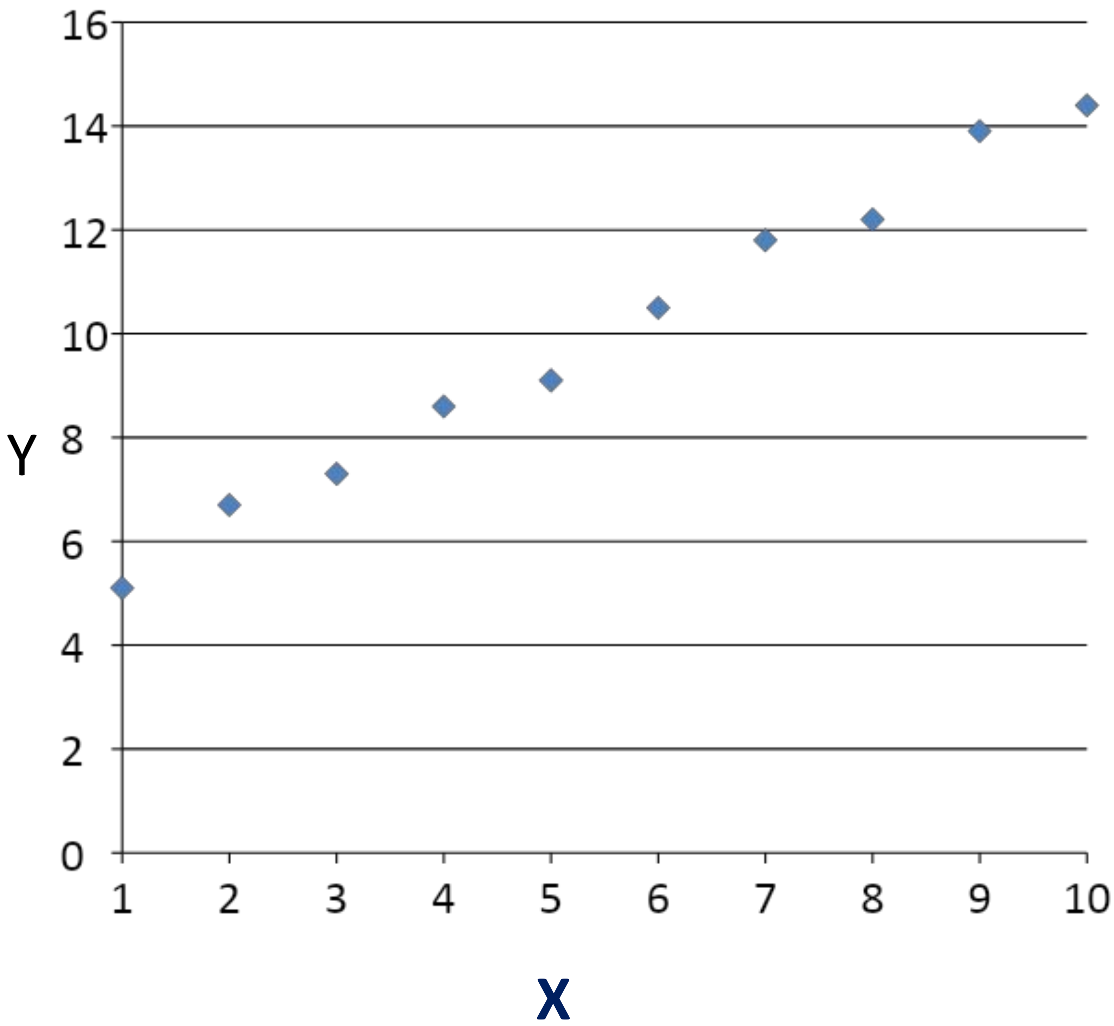
Пример:

Y	X
5,1	1
6,7	2
7,3	3
8,6	4
9,1	5
10,5	6
11,8	7
12,2	8
13,9	9
14,4	10



Пример:

Y	X
5,1	1
6,7	2
7,3	3
8,6	4
9,1	5
10,5	6
11,8	7
12,2	8
13,9	9
14,4	10



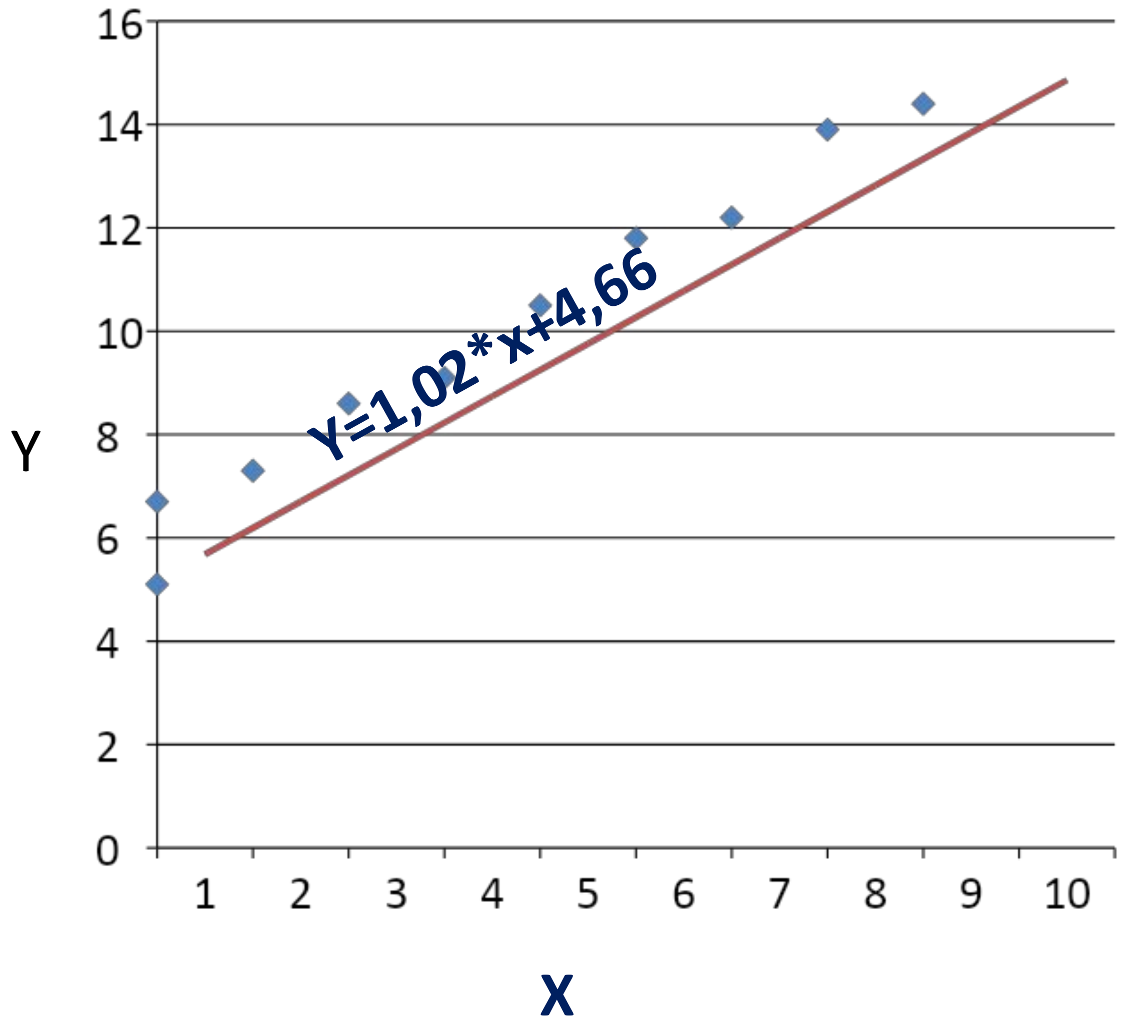
Линейный регрессионный анализ

Пример:

Y	X
6,7	2
11,8	7

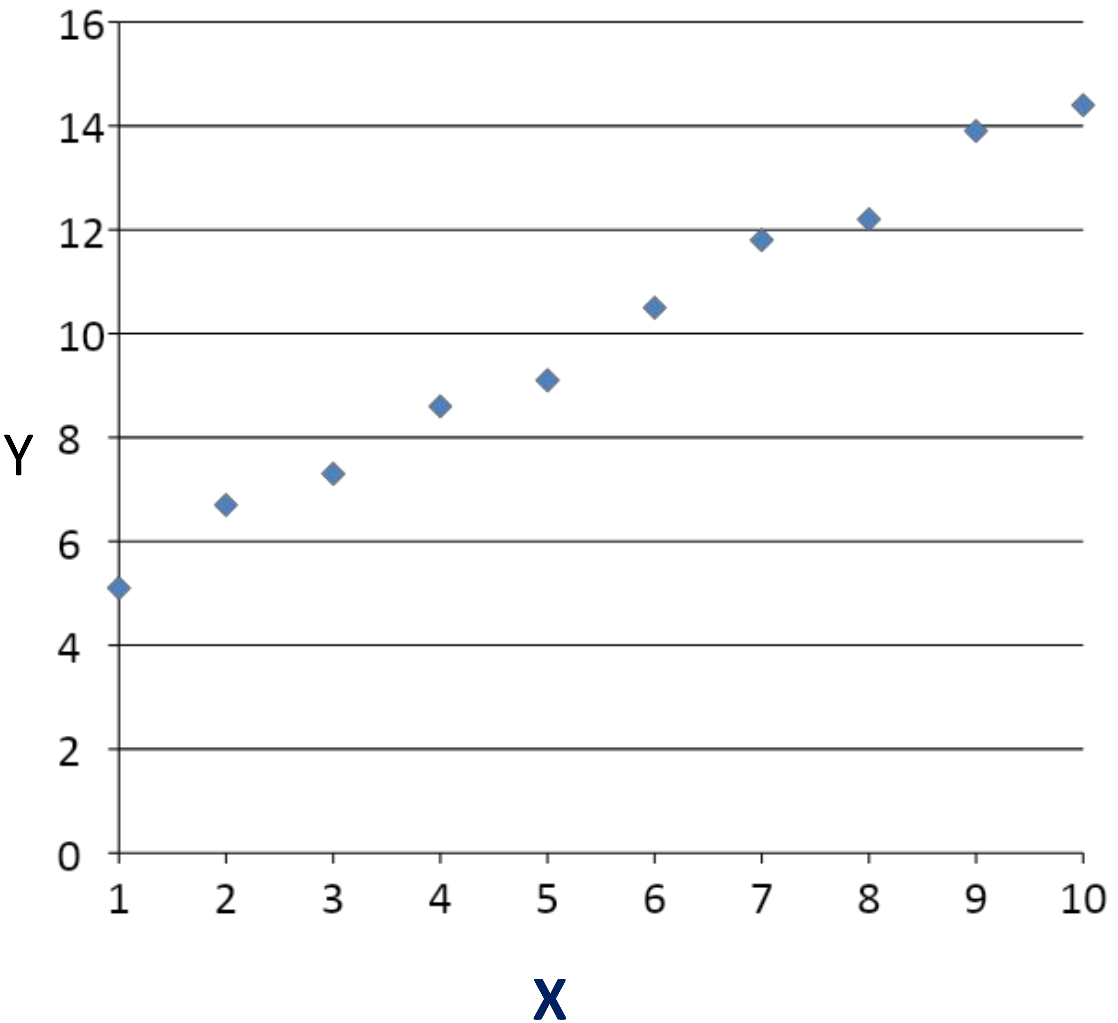
$$\begin{cases} 6{,}7 = a_1 2 + a_2 \\ 11{,}8 = a_1 7 + a_2 \end{cases}$$

$$\begin{cases} a_1 = 1{,}02 \\ a_2 = 4{,}66 \end{cases}$$



Задание:

Y	X
5,1	1
6,7	2
7,3	3
8,6	4
9,1	5
10,5	6
11,8	7
12,2	8
13,9	9
14,4	10



$$\sum_{1}^{10}(y - (a_1 x + a_2))^2 \to min$$

Другие виды регрессионного анализа

Полиномиальная функция регрессии:

$$y_i = a_1 x_{1i} + a_2 x_{2i} + a_3 x_{1i} x_{2i} + x_{1i} + a_2 {x_{1i}}^2 + a_{n+1} + \varepsilon_i$$

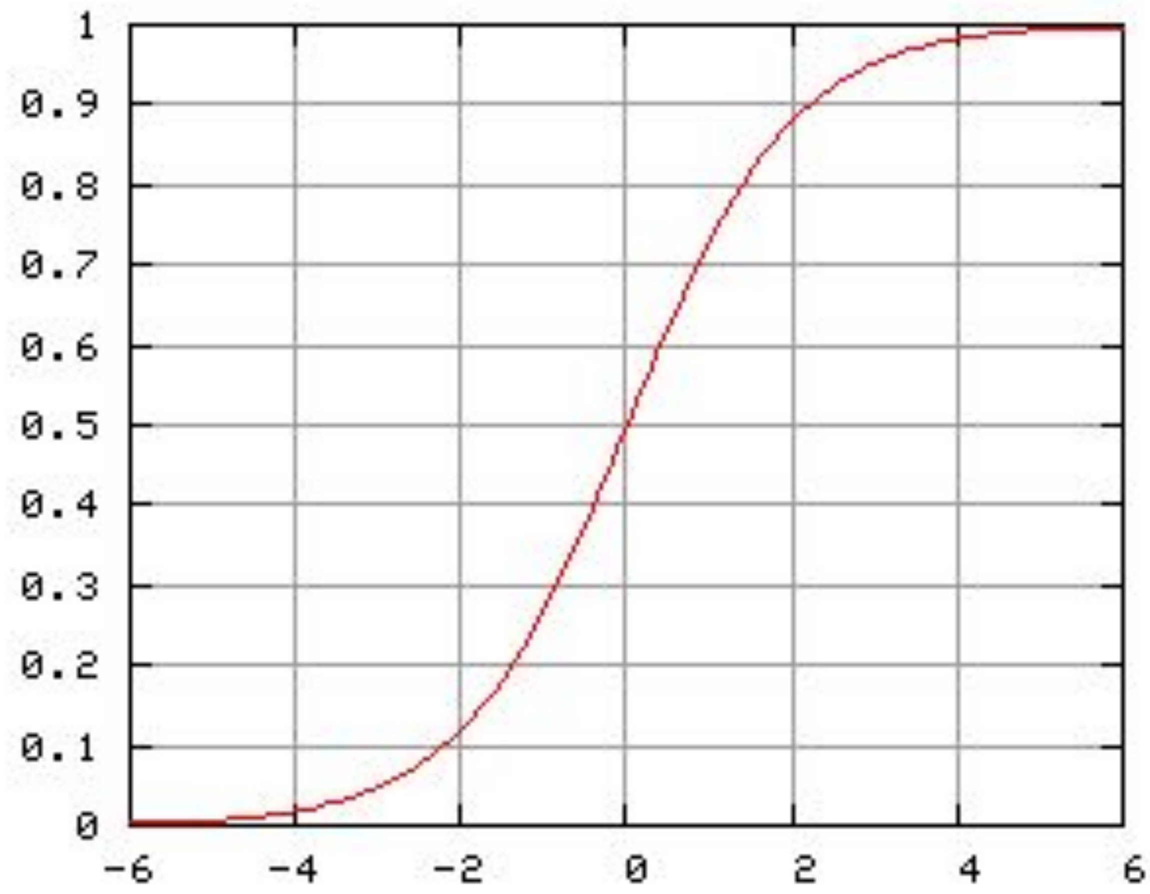
Запись в матричной форме: $Y = AX + U$

$$Y = \begin{bmatrix} y_1 \\ \dots \\ y_M \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{21} & x_{11}x_{21} & {x_{11}}^2 & 1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{1M} & x_{2M} & x_{1M}x_{2M} & {x_{1M}}^2 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} a_1 \\ \dots \\ a_{n+1} \end{bmatrix}$$

Логистическая регрессии:

$$y = \frac{1}{1 + e^{-x}}$$



Секции

Определения, термины и примеры применения

Виды регрессионного анализа

Коэффициенты регрессии и детерминации

Линейная регрессия на корреляции

Смысл коэффициента регрессии

Свойства коэффициента регрессии

- Коэффициент регрессии может принимать любые значения.
- Коэффициент регрессии не симметричен , т.е. изменяется, если X и Y поменять местами.
- Единицей измерения коэффициента регрессии является отношение единицы измерения Y к единице измерения X: ([Y] / [X]).
- Коэффициент регрессии изменяется при изменении единиц измерения X и Y .

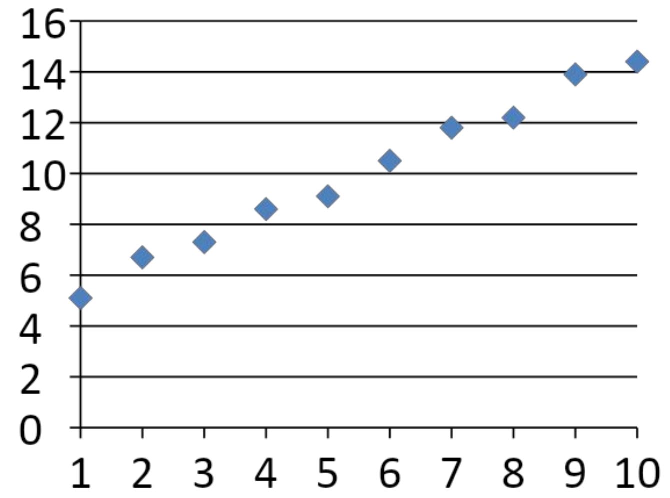
Например, результативный признак Y измеряется в рублях, а факторный признак X в количестве рабочих (чел.), то коэффициент регрессии измеряется в рублях на человека (руб. / чел.)

Коэффициент детерминации

Коэффициент детерминации рассматривают, как правило, в качестве основного показателя, отражающего меру качества регрессионной модели, описывающей связь между зависимой и независимыми переменными модели. Коэффициент детерминации показывает, какая доля вариации объясняемой переменной y учтена в модели и обусловлена влиянием на нее факторов, включенных в модель:

$$R^2 = 1 - \frac{\sum_{i=1}^{M}(y_i - \overline{f(x,a)})^2}{\sum_{i=1}^{M}(y_i - \bar{y})^2}$$

Достоинства и недостатки



Достоинства:

1. Простота вычислительных алгоритмов.
2. Наглядность и интерпретируемость результатов (для линейной модели)

Недостатки:

1. Невысокая точность прогноза (в основном - интерполяция данных).
2. Субъективный характер выбора вида конкретной зависимости (формальная подгонка модели под эмпирический материал).
3. Отсутствие объяснительной функции (невозможность объяснения причинно -следственной связи).

Секции

Определения, термины и примеры применения

Виды регрессионного анализа

Коэффициенты регрессии и детерминации

Линейная регрессия на корреляции

Линейная регрессия на корреляции

Линейная регрессия на корреляции — частный случай линейной регрессии. Применяется для построения простейших регрессионных моделей для прогнозирования временных рядов.

$$y_i = \sigma(y) \cdot \sum_{j=1}^{n} \frac{x_{ij} - \overline{x_j}}{\sigma(x_j)} \cdot corr(y, x_j) + \bar{y}$$

Вопросы

1. Виды зависимостей. Регрессионный анализ. Цели регрессионного анализа.
2. Зависимая и независимая переменные, коэффициент регрессии, невязка.
3. Виды регрессионного анализа.
4. Линейный регрессионный анализ.
5. Полиномиальные регрессионный анализ.
6. Экспоненциальный регрессионный анализ.
7. Коэффициент детерминации.
8. Достоинства и недостатки регрессионного анализа.
9. Когда нельзя применять регрессионный анализ.
10. Регрессия на коэффициентах корреляции.