МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_{r-1} x_{r-1} + a_r + \xi$$

y – зависимая или объясняемая переменная

$x_1, x_2 \ldots , x_{r-1}$ - независимые или объясняющие переменные

ξ - случайная составляющая.

Задача множественного регрессионного анализа – оценить $a_1, a_2 \ldots a_r$

Пример: Множественная регрессия

Мы хотим определить связь между потреблением, доходом семьи, финансовыми активами семьи и размером семьи.

- y – потребительские расходы.
- x_1 – доход семьи
- x_2 – финансовые активы семьи
- x_3 – размер семьи

$$y = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 + \xi$$

МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$y = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 + \xi$$

Для оценки необходима ***выборка (большое количество семей)***

№ семьи	y потребительские расходы	x_1 доход семьи	x_2 финансовые активы семьи	x_3 размер семьи
1	100	20	300	2
2	120	30	50	3
3	230	100	400	4
4	150	80	200	1
5	340	170	140	3

n – объем выборки

y_i доходы i-й семьи

x_{i1} потребительские расходы i-й семьи

x_{i2} доход i-й семьи

x_{i3} размер i-й семьи

$$i = 1 \ldots \ n$$

Уравнение для i-й семьи

$$y_i = a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + a_4 + \xi_i$$

Чтобы подобрать наилучшие a_1, a_2, a_3, a_4

$$S(a_1, a_2, a_3, a_4) = \sum_{i=1}^{n} \left(y_i - a_1 x_{i1} - a_2 x_{i2} - a_3 x_{i3} - a_4 \right)^2$$

$$\min_{a_1, a_2 \ldots a_r} S(a_1, a_2 \ldots a_r)$$

$$y = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 + \xi$$

Для оценки необходима ***выборка (большое количество семей)***

№ семьи	y потребительские расходы	x1 доход семьи	x2 финансовые активы семьи	x3 размер семьи	const
1	100	20	300	2	1
2	120	30	50	3	1
3	230	100	400	4	1
4	150	80	200	1	1
5	340	170	140	3	1

вектор Y

матрица X

№ семьи	y потребительские расходы	x1 доход семьи	x2 финансовые активы семьи	x3 размер семьи	const
1	100	20	300	2	1
2	120	30	50	3	1
3	230	100	400	4	1
4	150	80	200	1	1
5	340	170	140	3	1

вектор Y

матрица X

$$a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} \qquad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix}$$

$$Y = Xa + \xi$$

Оценки наименьших квадратов (ОНК) в КЛММР

$$S(a_1, a_2, a_3, a_4) = \sum_{i=1}^{n} \left(y_i - a_1 x_{i1} - a_2 x_{i2} - a_3 x_{i3} - a_4 \right)^2$$

$$\min_{a_1, a_2 \ldots a_r} S(a_1, a_2 \ldots a_r)$$

$$\hat{a} = \left(X^T X \right)^{-1} X^T Y$$

оценка наименьших квадратов (ОНК) параметров линейной множественной регрессии

№ семьи	y потребительские расходы	x1 доход семьи	x2 финансовые активы семьи	x3 размер семьи	const
1	100	20	300	2	1
2	120	30	50	3	1
3	230	100	400	4	1
4	150	80	200	1	1
5	340	170	140	3	1

вектор Y

матрица X

$$\vec{a} = \left(X^T X\right)^{-1} X^T Y$$

Пример оценки параметров в модели зависимости заработной платы от числа лет обучения и опыта работы

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,93164811	4,523407834	-5,95384	4,73E-09
N	2,674036105	0,231999296	11,52605	1,28E-27
Nrab	0,59409725	0,137923673	4,307435	1,96E-05

Zpl = 2.67*N + 0.59*NRab -26.93

Условия Гаусса-Маркова для модели линейной множественной регрессии

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_{r-1} x_{ir-1} + a_r + \xi_i \qquad i = \overline{1,n}$$

$$Y = Xa + \xi$$

1. $M\xi_i = 0 \quad \forall i = \overline{1,n}$

На самом деле это требование несущественно, если в модель включена константа

Условия Гаусса-Маркова для модели линейной множественной регрессии

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_{r-1} x_{ir-1} + a_r + \xi_i \qquad i = \overline{1,n}$$

$$Y = Xa + \xi$$

2. $D\xi_i = \sigma^2 \ \forall i = \overline{1,n}$ условие гомоскедастичности (постоянства дисперсии)

Условия Гаусса-Маркова для модели линейной множественной регрессии

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_{r-1} x_{ir-1} + a_r + \xi_i \qquad i = \overline{1,n}$$

$$Y = Xa + \xi$$

3. $\operatorname{cov}\left(\xi_i, \xi_j\right) = 0 \quad \forall i \neq j$ автокорреляция отсутствует

Условия Гаусса-Маркова для модели линейной множественной регрессии

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_{r-1} x_{ir-1} + a_r + \xi_i \qquad i = \overline{1,n}$$

$$Y = Xa + \xi$$

4. **Случайные ошибки** ξ_i **не зависят от объясняющих переменных** $x_1, x_2 \ldots , x_{r-1}$

Условия Гаусса-Маркова для модели линейной множественной регрессии

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_{r-1} x_{ir-1} + a_r + \xi_i \qquad i = \overline{1,n}$$

$$Y = Xa + \xi$$

5. $n > r, \quad rangX = r$ **– число наблюдений больше числа оцениваемых параметров и все r столбцов матрицы X линейно независимы.**

Для обеспечения статистической надежности должно выполняться условие: $n > 3r$

Условия Гаусса-Маркова

Модель $Y = Xa + \xi$ **, удовлетворяющая условиям 1-5 называется <u>классической линейной моделью множественной регрессии</u> (КЛММР)**

Условия Гаусса-Маркова

Если к 5-ти условиям добавляют шестое

6) Нормальность ошибок: $\xi_i \sim N\left(0, \sigma^2\right)$

То модель $Y = Xa + \xi$ **называется**

<u>классической нормальной линейной моделью</u>

<u>множественной регрессии</u> (КНЛММР)

СВОЙСТВА ОЦЕНОК КОЭФФИЦИЕНТОВ в КЛММР (ТЕОРЕМА ГАУССА-МАРКОВА)

В КЛАССИЧЕСКОЙ ЛИНЕЙНОЙ МОДЕЛИ МНОЖЕСТВЕННОЙ РЕГРЕССИИ (выполнены 5 условий Гаусса-Маркова) ОЦЕНКИ НАИМЕНЬШИХ КВАДРАТОВ

$$\hat{a} = \left(X^T X\right)^{-1} X^T Y$$

ЯВЛЯЮТСЯ НЕСМЕЩЕННЫМИ , СОСТОЯТЕЛЬНЫМИ, ЭФФЕКТИВНЫМИ

Если модель является нормальной (выполнены 6 условий Гаусса-Маркова), то оценки наименьших квадратов имеют нормальное распределение. Это позволяет проверять гипотезы и строить прогнозы с заданным уровнем надежности.

ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

$$\hat{y} = \hat{a}_1 x_1 + \hat{a}_2 x_2 + \ldots + \hat{a}_{r-1} x_{r-1} + \hat{a}_r$$

Интерпретация: коэффициент регрессии при переменной x_i показывает на сколько единиц изменится переменная y при изменении переменной x_i на 1 единицу, при условии постоянства других переменных:

Пример оценки параметров в модели зависимости заработной платы от числа лет обучения и опыта работы

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,93164811	4,523407834	-5,95384	4,73E-09
N	2,674036105	0,231999296	11,52605	1,28E-27
Nrab	0,59409725	0,137923673	4,307435	1,96E-05

Zpl = 2.67*N + 0.59*NRab -26.93

Пример оценки параметров в модели зависимости заработной платы от числа лет обучения и опыта работы

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,93164811	4,523407834	-5,95384	4,73E-09
N	2,674036105	0,231999296	11,52605	1,28E-27
Nrab	0,59409725	0,137923673	4,307435	1,96E-05

Zpl = 2.67*N + 0.59*NRab -26.93

Каждый дополнительный год обучения при данном опыте работы увеличивает часовой заработок на 2,67$

Каждый дополнительный год опыта работы при данной продолжительности обучения увеличивает часовой заработок на 0,59$

-26,93 не имеет содержательной интерпретации.

ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Пример y – затраты на питание (млрд. $)

x_1 – личный располагаемый доход (млрд. $)

x_2 – индекс цен на продукты питания (%)

$$\hat{y} = 0,112x_1 - 0,739x_2 + 116,7$$

ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Пример y – затраты на питание (млрд. $)

x_1 – личный располагаемый доход (млрд. $)

x_2 – индекс цен на продукты питания (%)

$$\hat{y} = 0,112x_1 - 0,739x_2 + 116,7$$

При увеличении личного располагаемого дохода на 1 млрд. $ (при сохранении неизменной цены) расходы на питание увеличатся на 112 млн.$

При увеличении индекса цен на 1 процентный пункт (при сохранении постоянных доходов) расходы на питание сократятся на 739 млн.$

116,7 не интерпретируется, т.к. x_1 и x_2 не могут быть равными 0.

Сравнение влияния на зависимую переменную различных объясняющих переменных

Пример y – затраты на питание (млрд. $)

x_1 – личный располагаемый доход (млрд. $)

x_2 – индекс цен на продукты питания (%)

$$\hat{y} = 0,112x_1 - 0,739x_2 + 116,7$$

При увеличении личного располагаемого дохода на 1 млрд. $ (при сохранении неизменной цены) расходы на питание увеличатся на 112 млн.$

При увеличении индекса цен на 1 процентный пункт (при сохранении постоянных доходов) расходы на питание сократятся на 739 млн.$

116,7 не интерпретируется, т.к. x_1 и x_2 не могут быть равными 0.

Какой фактор (доход или цена) оказывают большее влияние на расходы на питание?

Сравнение влияния на зависимую переменную различных объясняющих переменных

Расчет средних эластичностей

$$E_j = a_j \frac{\overline{x}_j}{\overline{y}}$$

Средняя эластичность j-го фактора. Показывает на сколько % изменится среднее значение фактора y при увеличении среднего значения фактора x_j на 1% от среднего значения

Сравнение влияния на зависимую переменную различных объясняющих переменных

Пример y – затраты на питание (млрд. $)

x_1 – личный располагаемый доход (млрд. $)

x_2 – индекс цен на продукты питания (%)

$$\hat{y} = 0,112x_1 - 0,739x_2 + 116,7$$

$$\overline{y} = 128 \text{ì ëðä\$} \quad \overline{x_1} = 780 \text{ì ëðä\$} \quad \overline{x_2} = 120$$

Сравнение влияния на зависимую переменную различных объясняющих переменных

Пример y – затраты на питание (млрд. $)

x_1 – личный располагаемый доход (млрд. $)

x_2 – индекс цен на продукты питания (%)

$$\hat{y} = 0,112x_1 - 0,739x_2 + 116,7$$

$$\overline{y} = 128 \text{ì ëðä\$} \quad \overline{x_1} = 780 \text{ì ëðä\$} \quad \overline{x_2} = 120$$

$$E_1 = 0,112 \cdot \frac{780}{128} = 0,68$$

$$E_2 = -0,739 \cdot \frac{120}{128} = -0,69$$