МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

В модели множественной регрессии переменная y зависит от нескольких переменных x.

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

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

МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

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

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

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

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

Модель строим с помощью Сервис – Анализ данных - регрессия

Пример: Имеются данные о потреблении мяса в США *B* в 1980 – 2007 годах (фунты на душу населения), и его зависимости от цены *P* (центы за фунт) и личного располагаемого дохода *YD* (тысячи долларов в расчете на душу населения).

Построим модель зависимости потребления мяса от цены и дохода

	A	B	C
1	P	YD	B
2	20,4	6,036	91,9
3	20,2	6,113	98,3
4	21,3	6,271	99,8
5	19,9	6,378	107,3
6	18	6,727	115,4
7	19,9	7,027	113,5
8	22,2	7,28	113,1
9	22,3	7,513	111,4
10	23,4	7,728	109,2
11	26,2	7,891	106,7
12	27,1	8,134	109,8

Модель строим с помощью Сервис – Анализ данных - регрессия

Пример: Имеются данные о потреблении мяса в США *B* в 1980 – 2007 годах (фунты на душу населения), и его зависимости от цены *P* (центы за фунт) и личного располагаемого дохода *YD* (тысячи долларов в расчете на душу населения).

Построим модель зависимости потребления мяса от цены и дохода

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересеч	57,34433	6,483396	8,844799	3,59E-09
P	-0,93789	0,106374	-8,8169	3,82E-09
YD	9,891514	1,137905	8,692744	5E-09

B=57.34-0.938P+9.892YD

ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

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:

Модель строим с помощью Сервис – Анализ данных - регрессия

Пример: Имеются данные о потреблении мяса в США *B* в 1980 – 2007 годах (фунты на душу населения), и его зависимости от цены *P* (центы за фунт) и личного располагаемого дохода *YD* (тысячи долларов в расчете на душу населения).

Построим модель зависимости потребления мяса от цены и дохода

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересеч	57,34433	6,483396	8,844799	3,59E-09
P	-0,93789	0,106374	-8,8169	3,82E-09
YD	9,891514	1,137905	8,692744	5E-09

B=57.34-0.938P+9.892YD

При увеличении цены на мясо на 1 цент за фунт потребление сократится на 0,938 фунтов на душу населения

Модель строим с помощью Сервис – Анализ данных - регрессия

Пример: Имеются данные о потреблении мяса в США *B* в 1980 – 2007 годах (фунты на душу населения), и его зависимости от цены *P* (центы за фунт) и личного располагаемого дохода *YD* (тысячи долларов в расчете на душу населения).

Построим модель зависимости потребления мяса от цены и дохода

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересеч	57,34433	6,483396	8,844799	3,59E-09
P	-0,93789	0,106374	-8,8169	3,82E-09
YD	9,891514	1,137905	8,692744	5E-09

B=57.34-0.938P+9.892YD

Какой фактор цена или доход влияет сильнее на потребление мяса?

Сравнение влияния на зависимую переменную различных объясняющих переменных

Расчет средних эластичностей

$$E_P = a_P \frac{\overline{P}}{\overline{B}}$$

Средняя эластичность по цене. Показывает на сколько % изменится потребление мяса, если цена увеличится на 1% процент.

x_j

Сравнение влияния на зависимую переменную различных объясняющих переменных

Расчет средних эластичностей

$$E_{YD} = a_{YD} \frac{\overline{YD}}{\overline{B}}$$

Средняя эластичность по доходу. Показывает на сколько % изменится потребление мяса, если доход увеличится на 1% процент.

Чем больше эластичность по абсолютной величине, тем сильнее влияние

Как оценить качество построенной модели?

Вычисляем прогноз по модели

B=57.34-0.938P+9.892YD

	A	B	C	D
1	P	YD	B	Прогноз B
2	20,4	6,036	91,9	97,92
3	20,2	6,113	98,3	98,87
4	21,3	6,271	99,8	99,40
5	19,9	6,378	107,3	101,77
6	18	6,727	115,4	107,00
7	19,9	7,027	113,5	108,19
8	22,2	7,28	113,1	108,53
9	22,3	7,513	111,4	110,74
10	23,4	7,728	109,2	111,84
11	26,2	7,891	106,7	110,83
12	27,1	8,134	109,8	112,39

Как оценить качество построенной модели?

Вычисляем остатки $e = y - \hat{y}$

	A	B	C	D	E
1	P	YD	B	Прогноз B	e
2	20,4	6,036	91,9	97,92	-6,02
3	20,2	6,113	98,3	98,87	-0,57
4	21,3	6,271	99,8	99,40	0,40
5	19,9	6,378	107,3	101,77	5,53
6	18	6,727	115,4	107,00	8,40
7	19,9	7,027	113,5	108,19	5,31
8	22,2	7,28	113,1	108,53	4,57

Как оценить качество построенной модели?

Находим относительную ошибку аппроксимации

$$A = \frac{\left| y - \hat{y} \right|}{y}$$

	A	B	C	D	E	F
1	P	YD	B	Прогноз B	e	*Относительная ошибка аппроксимации*
2	20,4	6,036	91,9	97,92	-6,02	6,55%
3	20,2	6,113	98,3	98,87	-0,57	0,58%
4	21,3	6,271	99,8	99,40	0,40	0,40%
5	19,9	6,378	107,3	101,77	5,53	5,16%
6	18	6,727	115,4	107,00	8,40	7,28%
7	19,9	7,027	113,5	108,19	5,31	4,68%
8	22,2	7,28	113,1	108,53	4,57	4,04%
9	22,3	7,513	111,4	110,74	0,66	0,59%
10	23,4	7,728	109,2	111,84	-2,64	2,42%

Процентный формат

Как оценить качество построенной модели?

Находим среднюю относительную ошибку аппроксимации

22	62,4	9,722	95,5	94,99	0,51	0,54%
23	58,6	9,769	95	99,01	-4,01	4,23%
24	56,7	9,725	95,1	100,36	-5,26	5,53%
25	55,5	9,93	103,3	103,51	-0,21	0,21%
26	57,3	10,419	108,9	106,66	2,24	2,05%
27	53,7	10,625	115,4	112,08	3,32	2,88%
28	52,6	10,905	120,6	115,88	4,72	3,91%
29	61,1	10,97	111,7	108,55	3,15	2,82%
30		Средняя относительная ошибка аппроксим				2,91%

среднее по столбцу

В среднем прогноз отличается от наблюдаемого значения на 2,91%

Как оценить качество построенной модели?

Еще один показатель качества – коэффициент детерминации
Для его вычисления вычисляем сумму квадратов остатков ESS (Error Sum of Squares). Его можно вычислить также как для линейной модели или просто посмотреть в таблице вывода результатов

ВЫВОД ИТОГОВ	
Регрессионная статистика	
Множественный R	0,873936
R-квадрат	0,763765
Нормированный R-квадра	0,744866
Стандартная ошибка	3,926538
Наблюдения	28

Проверка значимости коэффициентов модели регрессии

Построено уравнение $\hat{B} = aYD + bP + c$

Необходимо проверить значимость коэффициентов a и b
Если коэффициент a незначим, то потребление мяса не зависит от цены. Если коэффициент b незначим, то потребление мяса не зависит от дохода

Проверка значимости коэффициентов модели регрессии

Построено уравнение $\hat{B} = aYD + bP + c$

Необходимо проверить значимость коэффициентов a и b
Если коэффициент a незначим, то потребление мяса не зависит от цены. Если коэффициент b незначим, то потребление мяса не зависит от дохода

Для проверки значимости коэффициентов рассчитываются величины T

$$T_a = \frac{a}{s_a}, T_b = \frac{b}{s_b}$$

Проверка значимости коэффициентов модели регрессии

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	57,34433	6,483396	8,844799	3,59E-09
P	-0,93789	0,106374	-8,8169	3,82E-09
YD	9,891514	1,137905	8,692744	5E-09

Проверка значимости коэффициентов модели регрессии

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	57,34433	6,483396	8,844799	3,59E-09
P	-0,93789	0,106374	-8,8169	3,82E-09
YD	9,891514	1,137905	8,692744	5E-09

На основе t-статистик рассчитывают P-значения

P-значение - это вероятность того, что переменная не значима. При P-значении меньще 0,05 обычно считают, что соответствующая переменная значима, т.е. y зависит от этой x

В этом примере обе переменные P и YD значимы, т.е. и цена, и доход влияют на потребление мяса

Проверка значимости уравнения регрессии в целом

Уравнение регрессии считается незначимым, если ни одна из переменных, включенных в уравнение не влияет на переменную y

Для проверки значимости уравнения регрессии в целом, рассчитывается F-статистика

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	1246,162	623,0809	40,41335	1,47E-08
Остаток	25	385,4425	15,4177		
Итого	27	1631,604			

Проверка значимости уравнения регрессии в целом

Уравнение регрессии считается незначимым, если ни одна из переменных, включенных в уравнение не влияет на переменную y

Для проверки значимости уравнения регрессии в целом, рассчитывается F-статистика

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	1246,162	623,0809	40,41335	1,47E-08
Остаток	25	385,4425	15,4177		
Итого	27	1631,604			

Значимость F показывает вероятность того, что уравнение незначимо, т.е y не зависит от включенных в уравнение переменных x. Обычно считают, что если Значимость F<0.05, то уравнение регрессии значимо, т.е. хотя бы одна из включенных в уравнение переменных влияет на y.

Проверка значимости уравнения регрессии в целом

Уравнение регрессии считается незначимым, если ни одна из переменных, включенных в уравнение не влияет на переменную y

Для проверки значимости уравнения регрессии в целом, рассчитывается F-статистика

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	1246,162	623,0809	40,41335	1,47E-08
Остаток	25	385,4425	15,4177		
Итого	27	1631,604			

В нашем случае уравнение регрессии значимо.