Прогнозирование по модели МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Предположим, что нас интересует зависимость некоторого экономического показателя от ряда факторов.

Пример: Множественная регрессия

Мы хотим определить связь между потреблением некоторого продукта , среднедушевым доходом и ценой на этот продукт, иными словами построить функцию спроса на некоторый продукт.

- y – потребительские расходы.
- x_1 – среднедушевой доход
- x_2 – цена на продукт

$$y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon$$

МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon \qquad B = c + a \cdot YD + b \cdot P + \varepsilon$$

Для оценки необходима ***выборка (наблюдения за тремя переменными за несколько месяцев или лет)***

	A	B	C
1	B	P	YD
2	85,1	20,4	6,036
3	87,8	20,2	6,113
4	88,9	21,3	6,271
5	94,5	19,9	6,378
6	99,9	18	6,727
7	99,5	19,9	7,027
8	104,2	22,2	7,28
9	106,5	22,3	7,513
10	109,7	23,4	7,728
11	110,8	26,2	7,891
12	113,7	27,1	8,134
13	113	29	8,322
14	116	33,5	8,562

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

МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$B = c + a \cdot YD + b \cdot P + \varepsilon$$

Для оценки необходима ***выборка (наблюдения за тремя переменными за несколько месяцев или лет)***

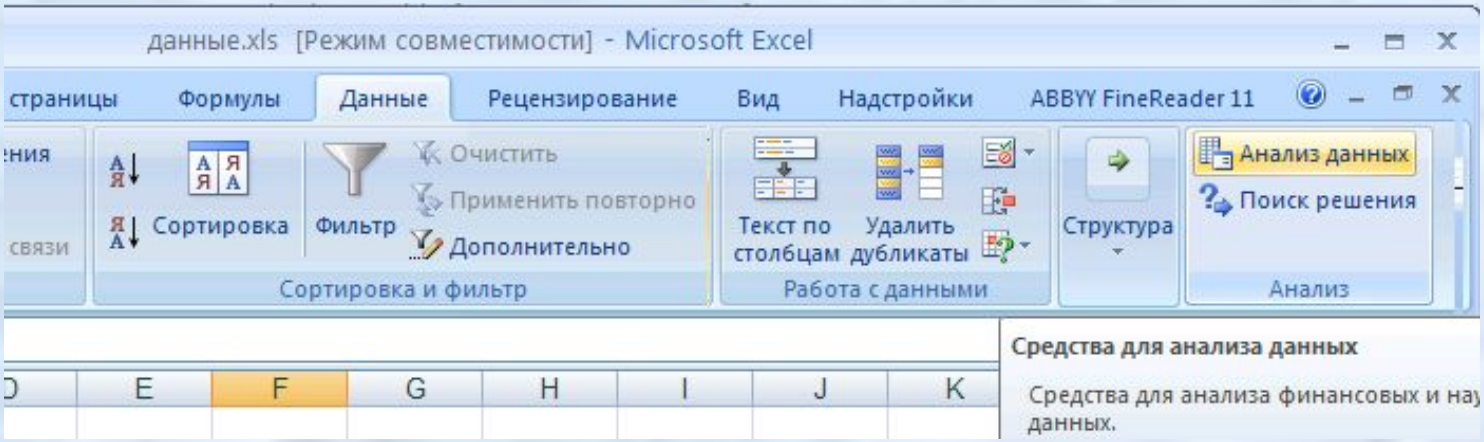
	A	B	C
1	B	P	YD
2	85,1	20,4	6,036
3	87,8	20,2	6,113
4	88,9	21,3	6,271
5	94,5	19,9	6,378
6	99,9	18	6,727
7	99,5	19,9	7,027
8	104,2	22,2	7,28
9	106,5	22,3	7,513
10	109,7	23,4	7,728
11	110,8	26,2	7,891
12	113,7	27,1	8,134
13	113	29	8,322
14	116	33,5	8,562

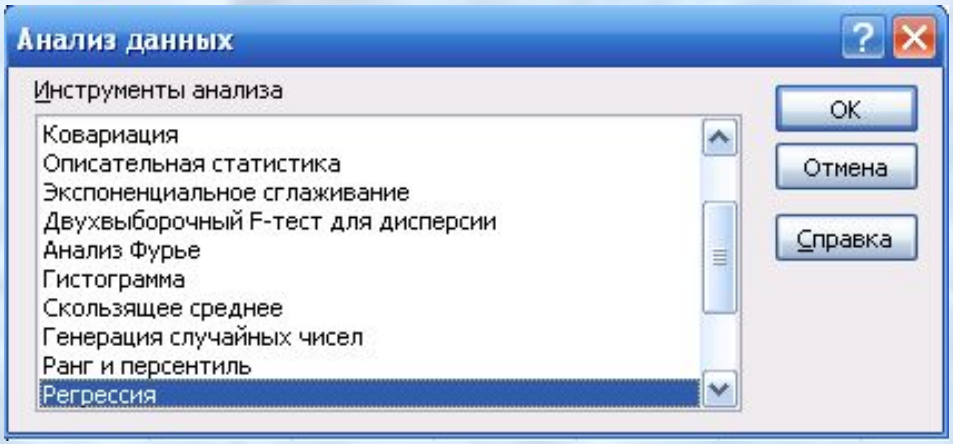
По этим данным нужно подобрать коэффициенты модели a,b,c «наилучшим образом». Наилучшим означает так, чтобы отклонение прогнозируемого по модели спроса у отличалось от наблюдаемого спроса как можно меньше.

$$B = c + a \cdot YD + b \cdot P + \varepsilon$$

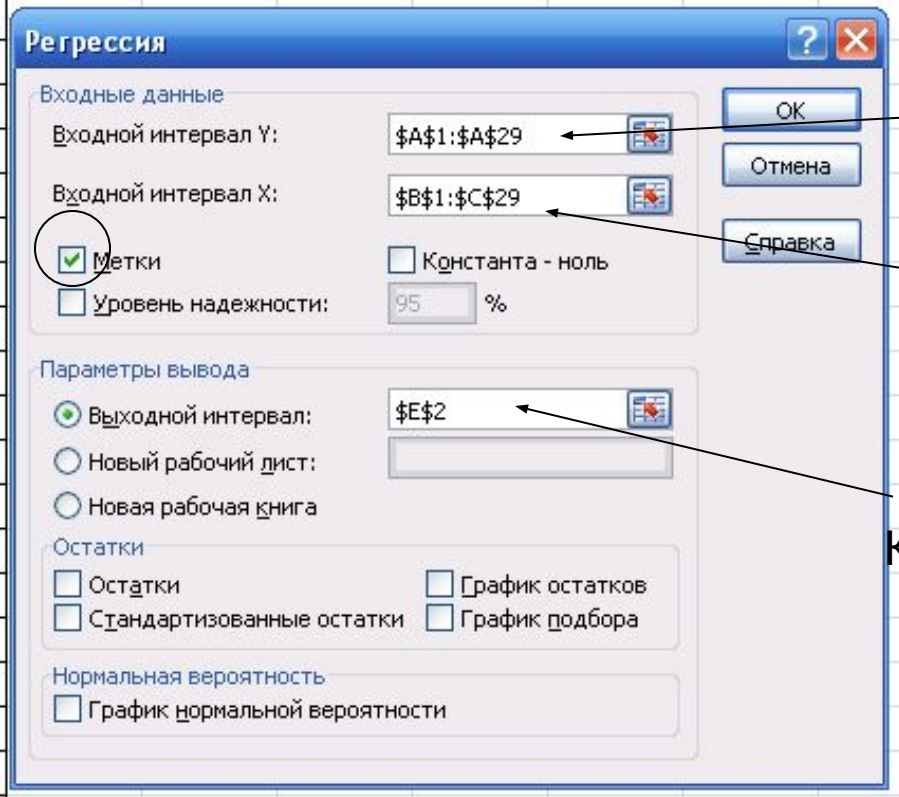
Используя специальные математические алгоритмы, Excel Подбирает наилучшие параметры автоматически.

На вкладке Данные выбираем Анализ данных.





	A	B	C
1	B	P	YD
2	85,1	20,4	6,036
3	87,8	20,2	6,113
4	88,9	21,3	6,271
5	94,5	19,9	6,378
6	99,9	18	6,727
7	99,5	19,9	7,027
8	104,2	22,2	7,28
9	106,5	22,3	7,513
10	109,7	23,4	7,728
11	110,8	26,2	7,891
12	113,7	27,1	8,134
13	113	29	8,322
14	116	33,5	8,562
15	108,7	42,8	9,042
16	115,4	35,6	8,867
17	118,9	32,2	8,944
18	127,4	33,7	9,175
19	123,5	34,4	9,381
20	117,9	48,5	9,735
21	105,4	66,1	9,829



столбец B вместе с названием

столбцы P и YD вместе с названием

Куда выводить результат

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множеств	0,81119
R-квадрат	0,65803
Нормиров	0,630672
Стандартн	6,080646
Наблюден	28

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	2	1778,674	889,3368	24,05287	1,5E-06
Остаток	25	924,3564	36,97426		
Итого	27	2703,03			

	Коэффициен	*дартная о*	*татистик*	*P-Значени*	*ижние 95%*	*ерхние 95%*
Y-пересеч	37,53605	10,0402	3,738575	0,000966	16,85787	58,21423
P	-0,88262	0,16473	-5,35798	1,48E-05	-1,22189	-0,54335
YD	11,89115	1,762162	6,748045	4,51E-07	8,261908	15,52039

B=37.54-0.882P+11.891YD

ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Интерпретация: коэффициент регрессии при переменной x показывает на сколько единиц изменится переменная y при изменении переменной x 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	37,53605	10,0402	3,738575	0,000966
P	-0,88262	0,16473	-5,35798	1,48E-05
YD	11,89115	1,762162	6,748045	4,51E-07

B=37.54-0.882P+11.891YD

При увеличении цены на мясо на 1 цент за фунт потребление сократится на 0,882 фунтов на душу населения (при неизменном доходе)

При увеличении дохода на 1 тысячу долларов на душу населения потребление мяса увеличится на 11,891 фунтов на душу населения (при неизменной цене)

Как оценить качество построенной модели?

Одной из характеристик качества является коэффициент детерминации
Коэффициент детерминации это доля вариации зависимой переменной, объясненная уравнением.
Коэффициент детерминации принимает значения от 0 до 1.
Чем ближе к 1, тем выше качество модели.

ВЫВОД ИТОГОВ	
Регрессионная статистика	
Множественный R	0,81119
R-квадрат	0,65803
Нормированный R-квадрат	0,630672
Стандартная ошибка	6,080646
Наблюдения	28

66% вариации потребления мяса объясняется доходом и ценой и 34% иными факторами.
Модель среднего качества, так как коэффициент детерминации не близок к 1.

Как оценить качество построенной модели?

Еще одной характеристикой качества является **средняя ошибка аппроксимации.**

Вычисляем прогноз по модели

B=37.54-0.882P+11.891YD

	A	B	C	D
1	B	P	YD	Прогноз B
2	85,1	20,4	6,036	91,31
3	87,8	20,2	6,113	92,40
4	88,9	21,3	6,271	93,31
5	94,5	19,9	6,378	95,81
6	99,9	18	6,727	101,64
7	99,5	19,9	7,027	103,53
8	104,2	22,2	7,28	104,51
9	106,5	22,3	7,513	107,19
10	109,7	23,4	7,728	108,78
11	110,8	26,2	7,891	108,24
12	113,7	27,1	8,134	110,34
13	113	29	8,322	110,90

Как оценить качество построенной модели?

Вычисляем остатки

$$e = B - \hat{B}$$

	A	B	C	D	E
1	B	P	YD	Прогноз B	e
2	85,1	20,4	6,036	91,31	-6,21
3	87,8	20,2	6,113	92,40	-4,60
4	88,9	21,3	6,271	93,31	-4,41
5	94,5	19,9	6,378	95,81	-1,31
6	99,9	18	6,727	101,64	-1,74
7	99,5	19,9	7,027	103,53	-4,03
8	104,2	22,2	7,28	104,51	-0,31
9	106,5	22,3	7,513	107,19	-0,69
10	109,7	23,4	7,728	108,78	0,92
11	110,8	26,2	7,891	108,24	2,56
12	113,7	27,1	8,134	110,34	3,36

Как оценить качество построенной модели?

Находим относительную ошибку аппроксимации

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

	A	B	C	D	E	F
1	B	P	YD	Прогноз B	e	A
2	85,1	20,4	6,036	91,31	-6,21	7,29%
3	87,8	20,2	6,113	92,40	-4,60	5,24%
4	88,9	21,3	6,271	93,31	-4,41	4,96%
5	94,5	19,9	6,378	95,81	-1,31	1,39%
6	99,9	18	6,727	101,64	-1,74	1,74%
7	99,5	19,9	7,027	103,53	-4,03	4,05%
8	104,2	22,2	7,28	104,51	-0,31	0,30%
9	106,5	22,3	7,513	107,19	-0,69	0,65%
10	109,7	23,4	7,728	108,78	0,92	0,84%

Процентный формат

Как оценить качество построенной модели?

Находим среднюю относительную ошибку аппроксимации

	A	B	C	D	E	F
25	105,7	55,5	9,93	106,63	-0,93	0,88%
26	105,5	57,3	10,419	110,86	-5,36	5,08%
27	106,5	53,7	10,625	116,48	-9,98	9,37%
28	107,3	52,6	10,905	120,78	-13,48	12,57%
29	103,3	61,1	10,97	114,05	-10,75	10,41%
30			Средняя ошибка аппроксимации			4,27%

среднее по столбцу

В среднем прогноз отличается от наблюдаемого значения на 4,27%

Проверка значимости коэффициентов модели регрессии

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

Необходимо проверить значимость коэффициентов a и b
Значимость коэффициента говорит о том, что он существенно отличается от нуля. Незначимость говорит о том, что можно считать, что коэффициент равен 0.
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

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	37,53605	10,0402	3,738575	0,000966
P	-0,88262	0,16473	-5,35798	1,48E-05
YD	11,89115	1,762162	6,748045	4,51E-07

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

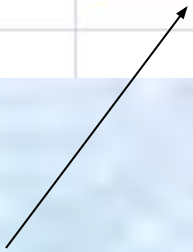
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	1778,674	889,3368	24,05287	1,5E-06
Остаток	25	924,3564	36,97426		
Итого	27	2703,03			

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

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	1778,674	889,3368	24,05287	1,5E-06
Остаток	25	924,3564	36,97426		
Итого	27	2703,03			



В нашем случае уравнение регрессии значимо.

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-пересечение	37,53605	10,0402	3,738575	0,000966
P	-0,88262	0,16473	-5,35798	1,48E-05
YD	11,89115	1,762162	6,748045	4,51E-07

B=37.54-0.883P+11.891YD

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

Модели множественной нелинейной регрессии

Построим теперь степенную модель зависимости потребления мяса от цены и дохода:

$$B = c \cdot YD^{a} \cdot P^{b}$$

Модели множественной нелинейной регрессии

Построим теперь степенную модель зависимости потребления мяса от цены и дохода:

$$B = c \cdot YD^{a} \cdot P^{b}$$

Прологарифмируем модель

Модели множественной нелинейной регрессии

Построим теперь степенную модель зависимости потребления мяса от цены и дохода:

$$B = c \cdot YD^{a} \cdot P^{b}$$

Прологарифмируем модель

$$\ln B = \ln c + a \ln YD + b \ln P$$

Модели множественной нелинейной регрессии

Построим теперь степенную модель зависимости потребления мяса от цены и дохода:

$$B = c \cdot YD^{a} \cdot P^{b}$$

Прологарифмируем модель

$$\ln B = \ln c + a \ln YD + b \ln P$$

Теперь построим линейную модель с помощью сервиса Анализ данных, задав в качестве зависимой переменной ln(B), а в качестве регрессоров ln(YD) и ln(P)

Модели множественной нелинейной регрессии

P	YD	B	ln(P)	ln(YD)	ln(B)
20,4	6,036	85,1	3,016	1,798	4,444
20,2	6,113	87,8	3,006	1,810	4,475
21,3	6,271	88,9	3,059	1,836	4,488
19,9	6,378	94,5	2,991	1,853	4,549
18	6,727	99,9	2,890	1,906	4,604
19,9	7,027	99,5	2,991	1,950	4,600
22,2	7,28	104,2	3,100	1,985	4,646
22,3	7,513	106,5	3,105	2,017	4,668
23,4	7,728	109,7	3,153	2,045	4,698
26,2	7,891	110,8	3,266	2,066	4,708
27,1	8,134	113,7	3,300	2,096	4,734
29	8,322	113	3,367	2,119	4,727

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересеч	3,594	0,141	25,439	0,000
ln(P)	-0,344	0,062	-5,536	0,000
ln(YD)	1,071	0,148	7,216	0,000

$$\ln(B)=3.594-0.344\ln(P)+1.071\ln(YD)$$

Модели множественной нелинейной регрессии

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересеч	3,594	0,141	25,439	0,000
ln(P)	-0,344	0,062	-5,536	0,000
ln(YD)	1,071	0,148	7,216	0,000

ln(B)=3.594-0.344ln(P)+1.071ln(YD)

$$\ln B = \ln c + a \ln YD + b \ln P$$

$$a = 1.071$$

$$b = -0.344$$

$$\ln c = 3.594$$

Модели множественной нелинейной регрессии

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересеч	3,594	0,141	25,439	0,000
ln(P)	-0,344	0,062	-5,536	0,000
ln(YD)	1,071	0,148	7,216	0,000

ln(B)=3.594-0.344ln(P)+1.071ln(YD)

$$\ln B = \ln c + a \ln YD + b \ln P$$

$$\ln c = 3.594$$

$$c = \exp(3.594) = 36.38$$

Модели множественной нелинейной регрессии

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересеч	3,594	0,141	25,439	0,000
ln(P)	-0,344	0,062	-5,536	0,000
ln(YD)	1,071	0,148	7,216	0,000

$$a = 1.071$$

$$b = -0.344$$

$$c = 36.58$$

$$B = c \cdot YD^{a} \cdot P^{b}$$

$$B = 36.38 \cdot YD^{1.071} \cdot P^{-0.344}$$ степенная модель множественной регрессии

Коэффициенты - эластичности

Модели множественной нелинейной регрессии

$$B = 36.38 \cdot YD^{1.071} \cdot P^{-0.344}$$ степенная модель множественной регрессии

Показатели степени в степенной модели являются эластичностями
Интерпретация эластичностей:
С ростом цены на мясо на 1% спрос уменьшается на 0,344%
(при неизменном доходе)
С ростом располагаемого дохода на 1% спрос увеличивается на 0,344%
(при неизменной цене)

Модели множественной нелинейной регрессии

Вычислим среднюю ошибку аппроксимации

$$B = 36.38 \cdot YD^{1.071} \cdot P^{-0.344}$$

↓

P	YD	B	ln(P)	ln(YD)	ln(B)	прогноз	e	Относительная ошибка
20,4	6,036	85,1	3,016	1,798	4,444	88,37	-3,27	3,84%
20,2	6,113	87,8	3,006	1,810	4,475	89,88	-2,08	2,37%
21,3	6,271	88,9	3,059	1,836	4,488	90,70	-1,80	2,03%
19,9	6,378	94,5	2,991	1,853	4,549	94,55	-0,05	0,05%
18	6,727	99,9	2,890	1,906	4,604	103,62	-3,72	3,73%
19,9	7,027	99,5	2,991	1,950	4,600	104,90	-5,40	5,42%
22,2	7,28	104,2	3,100	1,985	4,646	104,92	-0,72	0,69%
22,3	7,513	106,5	3,105	2,017	4,668	108,35	-1,85	1,74%
23,4	7,728	109,7	3,153	2,045	4,698	109,84	-0,14	0,13%
26,2	7,891	110,8	3,266	2,066	4,708	108,04	2,76	2,49%
52,6	10,905	107,3	3,963	2,389	4,676	120,18	-12,88	12,01%
61,1	10,97	103,3	4,113	2,395	4,638	114,87	-11,57	11,20%
				Средняя ошибка аппроксимации				4,05%

Считаем как в линейной модели