Лекция 5

Линейная регрессия с несколькими переменными

Автор: д.т.н Джомартова Ш.А.

Множественная регрессия

Построение множественной регрессионной модели

- **Множественная регрессия** является расширением простой линейной регрессии. С помощью простой регрессии оценивалась степень влияния одной независимой переменной (предиктора) на зависимую переменную (критерий). В отличие от простой регрессии (Y=B*X+A), множественная регрессия исследует влияние двух и более предикторов на критерий (Y=B1*X1+B2*X2+B3*X3+…+A).

- Применение множественной регрессии позволяет исследователю ответить на вопрос, **насколько хорошо оцененное уравнение аппроксимирует данные**, есть ли значимая линейная связь, а также каковы оцененные значения коэффициентов для уравнения наилучшего предсказания. Кроме того, может быть определена относительная важность независимых переменных в предсказании зависимой переменной.

- Уравнение множественной регрессии

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon$$

- Основная **цель** множественной регрессии

  – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

Условия включения факторов при построении множественной регрессии.

1. факторы должны быть количественно измеримы.
2. Факторы не должны быть интеркоррелированы.

- Если между факторами существует высокая корреляция, то параметры уравнения регрессии оказываются ***неинтерпретируемыми***

Пример.

Рассмотрим регрессию себестоимости: единицы продукции (тен., *y*) от заработной платы работника (тен., *x*) и производительности его труда (единиц в час, *z*):

$$y = 22600 - 5 \cdot x - 10 \cdot z$$

$$= 0{,}95$$

$$r_{xz}$$

Линейная модель множественной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

y_i — значения зависимой переменной на *i*-ом объекте,

x_{ik} — значения независимых переменных (регрессоров),

k — число коэффициентов (переменных) в модели,

ε_i — случайные ошибки

Методы оценки коэффициентов

Идея метода наименьших квадратов: подбор параметров таким образом чтобы сумма квадратов отклонений

$$e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

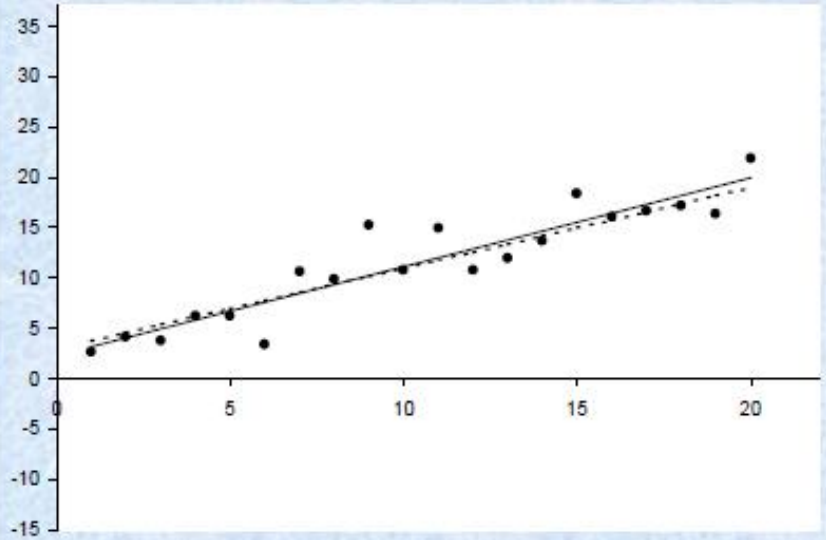
была наименьшей

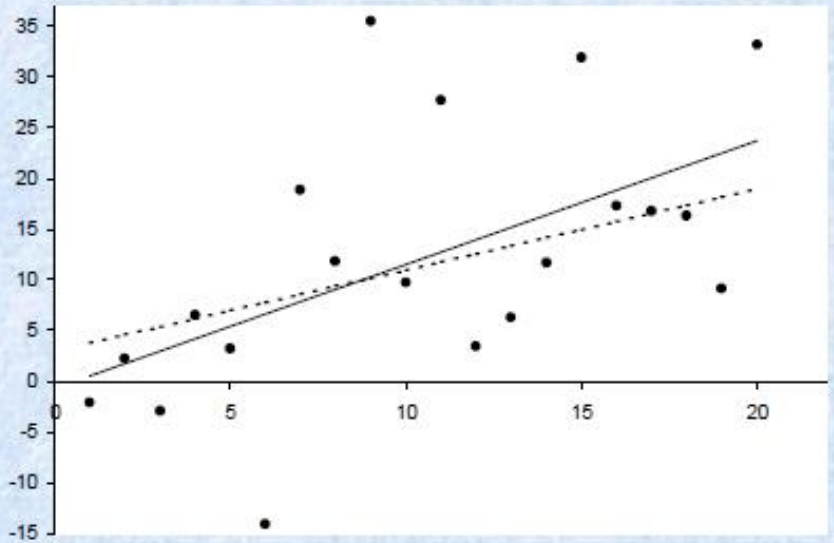
СВОЙСТВА ОЦЕНОК ПАРАМЕТРОВ

влияние параметров выборки на дисперсию оценок

(на примере парной регрессии)

Чем больше дисперсия оценок- тем менее точны оценки коэффициентов

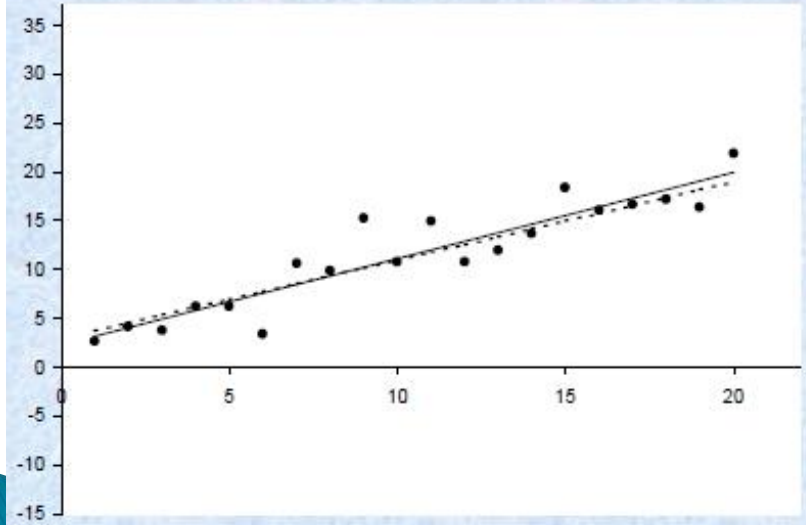


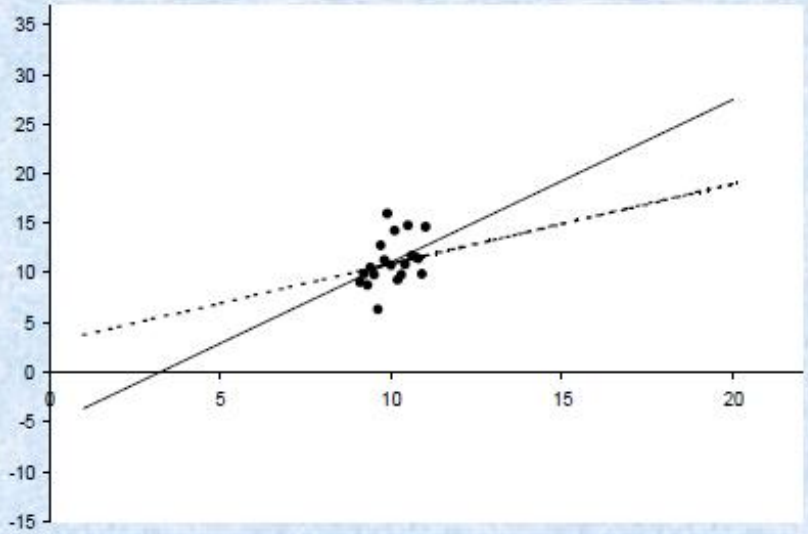


СВОЙСТВА ОЦЕНОК ПАРАМЕТРОВ

влияние параметров выборки на дисперсию оценок (на примере парной регрессии)

Оценки тем точнее, чем разнообразнее выборка по значениям регрессоров





Качество подгонки модели

1. Стандартная ошибка регрессии
2. Коэффициент детерминации R^2
3. Скорректированный (нормированный) коэффициент детерминации R^2

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 1990 – 2017 годах, и его зависимости от цены *P* и личного располагаемого дохода *YD* .

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)***

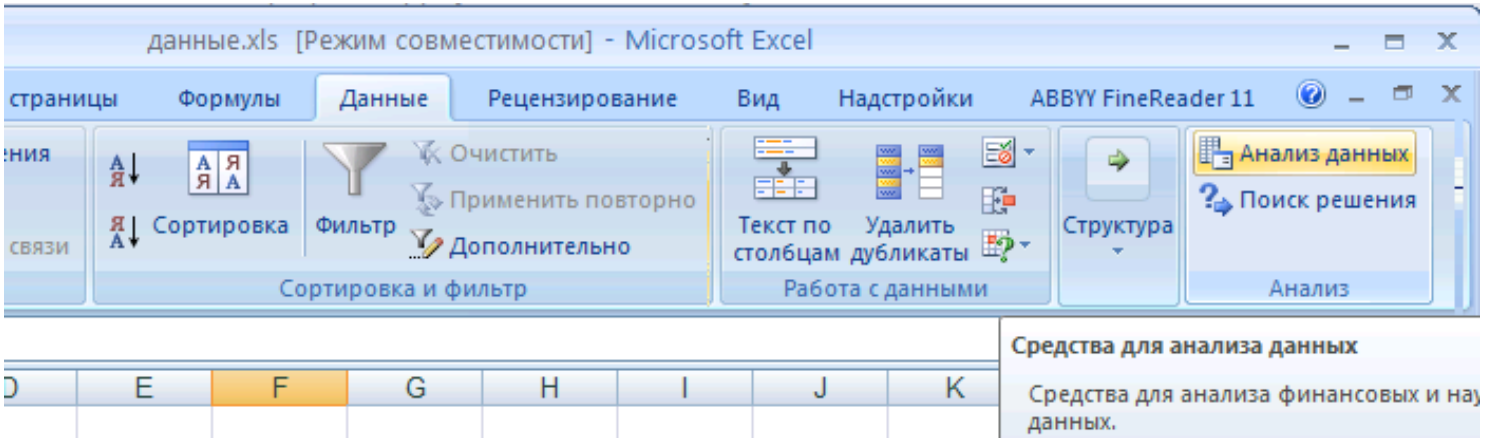
	A	B	C
1	B	P	YD
2	85,1	20,4	6,036
3	87,8	20,2	6,113
4	88,9	21,3	6,271
5	94,5	19,9	6,378
6	99,9	18	6,727
7	99,5	19,9	7,027
8	104,2	22,2	7,28
9	106,5	22,3	7,513
10	109,7	23,4	7,728
11	110,8	26,2	7,891
12	113,7	27,1	8,134
13	113	29	8,322
14	116	33,5	8,562

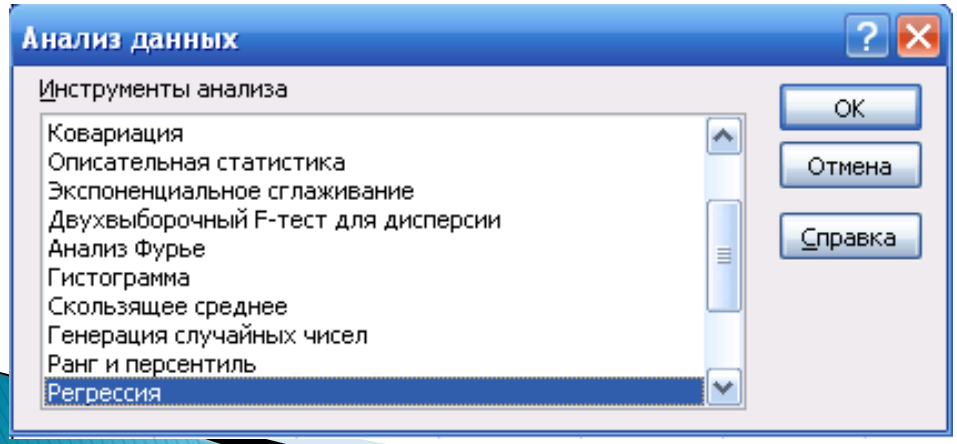
По этим данным нужно подобрать коэффициенты модели a,b,c «наилучшим образом». Наилучшим означает так, чтобы отклонение прогнозируемого по модели спроса у отличалось от наблюдаемого спроса как можно меньше.

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

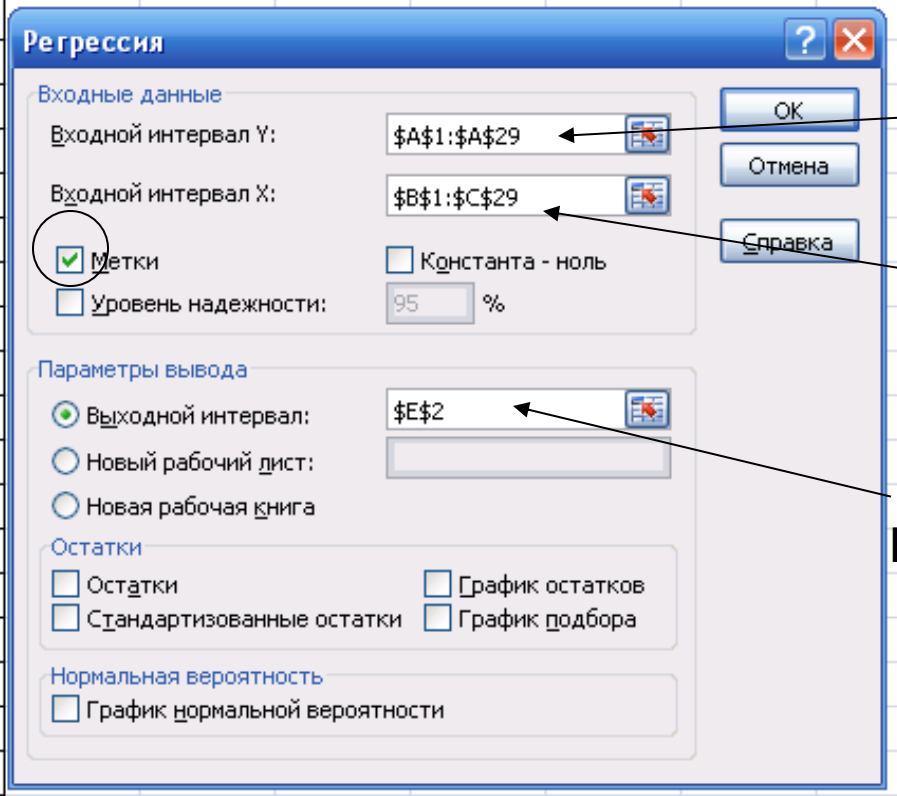
Используя специальные математические алгоритмы, Excel Подбирает наилучшие параметры автоматически.

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-Значение*	*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:

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

$$e = B - B$$

	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	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%

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

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 $$B = aYD + bP + c$$

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.

Задача:

- Предприниматель желает сдать в аренду на один год принадлежащий ему отель (80 комнат), расположенный в престижной курортной зоне, обладающий собственным пляжем, общая площадь территории отеля составляет 3,42 кв. км. Для установления величины арендной платы, предприниматель решил изучить объявления владельцев подобных отелей и смоделировать величину арендной платы.

1. Ввод исходных данных

	A	B	C	D	E
1	Величина годовой платы за аренду отеля, тыс. руб. (Y)	Число комнат в отеле (X1)	Престиж-ность района, в котором расположен отель (X2)	Наличие у отеля собствен-ного пляжа (X3)	Общая площадь территории, принадлежащей отелю, кв. км.(X4)
2	123	25	1	1	1
3	115	25	0	0	0,8
4	126	30	1	1	1,2
5	130	30	0	1	1,5
6	125	30	1	0	1,4
7	134	45	0	0	2
8	142	45	1	0	2,5
9	140	45	0	1	2,2
10	143	45	0	0	2,7
11	150	60	0	1	2,8
12	153	60	0	1	3
13	158	60	1	0	3,6
14	160	75	1	1	3,5
15	163	75	0	0	3,8
16	164	75	1	1	3,6
17	168	75	0	0	3,75
18	172	100	0	1	4,1
19	180	100	1	1	4,7
20	177	120	1	1	4,25
21	182	120	0	1	4,65

2. Нахождение матрицы линейных коэффициентов парной корреляции

	A	B	C	D	E	F
1		*Величина годовой платы за аренду отеля, тыс. руб. (Y)*	*Число комнат в отеле(X1)*	*Престижность района, в котором расположен отель(X2)*	*Наличие у отеля собственного пляжа(X3)*	*Общая площадь территории, принадлежащей отелю, кв. км.(X4)*
2	Величина годовой платы за аренду отеля, тыс. руб. (Y)	1				
3	Число комнат в отеле (X1)	0,959208926	1			
4	Престижность района, в котором расположен отель (X2)	0,013800523	0,006840396	1		
5	Наличие у отеля собственного пляжа (X3)	0,275193758	0,333429826	0,123091491	1	
6	Общая площадь территории, принадлежащей отелю, кв. км. (X4)	0,99233351	0,938654049	0,006464841	0,192293213	1

3. Расчет оценок коэффициентов регрессии

	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	*Регрессионная статистика*						
4	Множественный R	0,9961					
5	R-квадрат	0,9921					
6	Нормированный R-квадр	0,9912					
7	Стандартная ошибка	1,9285		$Y=101,84+3,58*X_3+16,21*X_4+e$			
8	Наблюдения	20,0000					
9				F(0,05;2;17)	3,5915		
10	Дисперсионный анализ						
11		*df*	*SS*	*MS*	*F*	*Значимость F*	
12	Регрессия	2,0000	7958,5245	3979,2623	1069,9398	0,0000	
13	Остаток	17,0000	63,2255	3,7191			
14	Итого	19,0000	8021,7500				
15							
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
17	Y-пересечение	101,8480	1,1587	87,8968	0,0000	99,4033	104,2927
18	Наличие у отеля собственного пляжа X3	3,5817	0,8970	3,9931	0,0009	1,6892	5,4742
19	Общая площадь территории, принадлежащей отелю, кв. км. X4	16,2149	0,3647	44,4582	0,0000	15,4454	16,9844
20							
21			t(0,05/2;17)	2,10981556			

Таким образом, построенная модель может быть записана в виде

$$Y=101,84+3,58*X_3+16,21*X_4+e\ .$$

Коэффициент b_3 этой модели показывает, что наличие у отеля собственного пляжа приводит к удорожанию арендной платы в среднем на **3,58** тыс. руб. при неизменной общей территории, принадлежащей отелю.

Коэффициент b_4 этой модели показывает, что увеличение общей территории на 1 кв.км приводит к удорожанию арендной платы на **16,2**1 тыс.руб.

4.Множественный коэффициент корреляции

2		
3	*Регрессионная статистика*	
4	Множественный R	0,9961
5	R-квадрат	0,9921
6	Нормированный R-квадр	0,9912
7	Стандартная ошибка	1,9285
8	Наблюдения	20,0000

Множественный коэффициент корреляции достаточно высокий, что свидетельствует о существенной зависимости величины арендной платы от включенных в модель факторов.

5. Скорректированный коэффициент детерминации

	A	B
1	ВЫВОД ИТОГОВ	
2		
3	*Регрессионная статистика*	
4	Множественный R	0,9961
5	R-квадрат	0,9921
6	Нормированный R-квадр	0,9912
7	Стандартная ошибка	1,9285
8	Наблюдения	20,0000

Скорректированный коэффициент детерминации свидетельствует, что 99,12% дисперсии арендной платы формируется под влиянием факторов, включенных в модель.

6. Расчет бетта-коэффициентов

$$\beta_3 = \frac{0,2752 - 0,9923 \cdot 0,1923}{1 - 0,1923^2} = 0,0876;$$

$$\beta_4 = \frac{0,9923 - 0,2752 \cdot 0,1923}{1 - 0,1923^2} = 0,9755.$$

Полученные значения бетта-коэффициентов позволяют проранжировать факторы по степени их влияния на арендную плату следующим образом:
1) общая площадь территории, принадлежащей отелю;
2) наличие собственного пляжа.

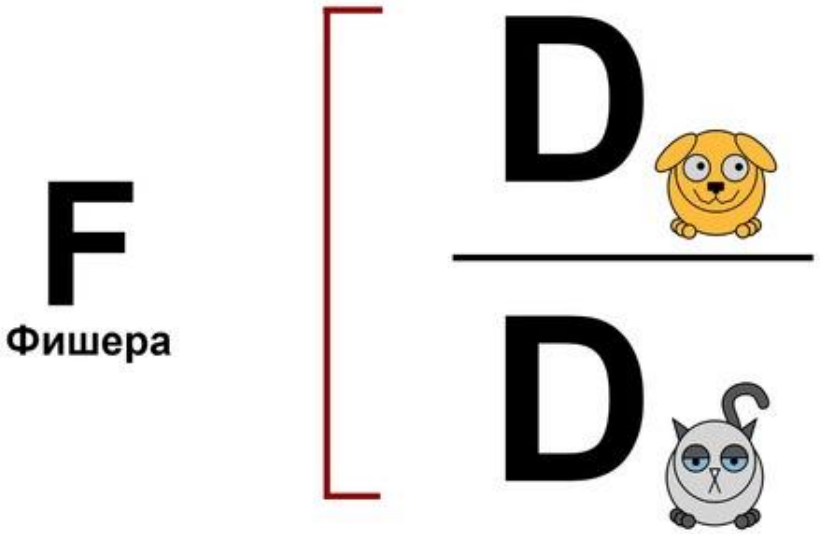
7. F-тест

F(0,05;2;17)	3,5915	

MS	*F*	*Значимость F*
3979,2623	1069,9398	0,0000
3,7191		

Сравнение расчетного значения F-критерия с табличным (1069,9398 > 3,5915) с вероятностью 95% позволяет сделать вывод об адекватности и статистической значимости модели в целом.

- Мы можем посмотреть, являются ли песики более разнообразными по размеру, чем котики, или же нет. Для этого мы можем воспользоваться *F-критерием равенства дисперсий Фишера,* который укажет нам, насколько различаются между собой эти показатели.



Необходимо заметить, что в этой формуле сверху всегда должна стоять большая дисперсия, а снизу — меньшая.

8. t-тест

Стандартная ошибка	t-статистика
1,1587	87,8968
0,8970	3,9931
0,3647	44,4582
t(0,05/2;17)	2,10981556

Сравнение модулей полученных t- статистик с табличным значением (2,1098) подтверждает значимость коэффициентов регрессии b_0, b_3, b_4.





Чтобы понять, насколько они отличаются друг от друга, необходимы так называемые *меры различий для несвязанных выборок*. Большая часть таких мер показывает, насколько типичный песик отличается от типичного котика. Например, самая популярная из них — *t-критерий Стьюдента для несвязанных выборок* — оценивает, насколько различаются их средние размеры.

Если разность средних достаточно большая, а стандартная ошибка очень маленькая, то значение t-критерия будет весьма внушительным.
А чем больше t-критерий, тем с большей уверенностью мы можем утверждать, что в среднем песики отличаются от котиков.



9. Расчет средней прогнозируемой величины арендной платы

$Y=101,84+3,58*X_3+16,21*X_4+e$; $X_3=1$; $X_4=3,42$ кв.км

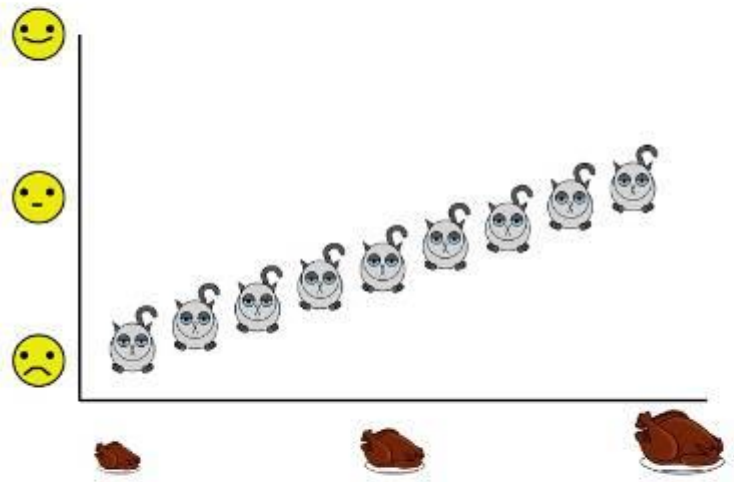
$Yx=101,84+3,58*1+16,21*3,42=160,8846\approx161$ тыс.руб.

Задачи оценки взаимосвязи между переменными или прогноза

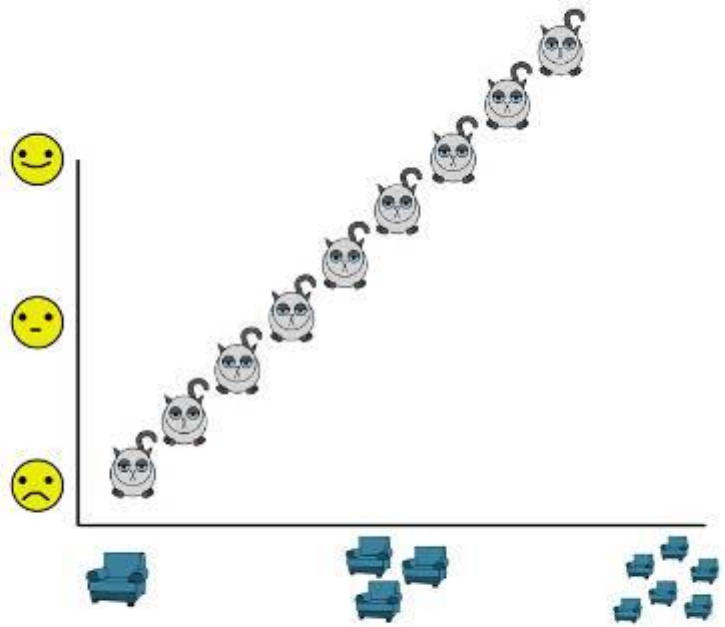
Задача	Количественные нормально распределенные переменные	Количественные ненормально распределенные переменные или ранги	Биноминальные данные (два возможных результата)
Оценить взаимосвязь между двумя переменными	Коэффициент парной корреляции Пирсона	Коэффициенты ранговых корреляций (Спирмена, Кендалла)	Коэффициенты связи
Предсказать изменение одной переменной, если была измерена другая переменная	Простая линейная регрессия или нелинейная регрессия	Непараметрическая (ранговая) регрессия	Простая логистическая регрессия
Предсказать значение, базируясь на нескольких переменных	Множественная линейная (нелинейная) регрессия	Множественная линейная ранговая медианная) регрессия	Множественная логистическая регрессия

Формула счастья котиков

Очевидно, что каждый подранный диван делает котиков гораздо счастливее, чем очередное увеличение пайков. Эта разница математически описывается с помощью коэффициента b1.

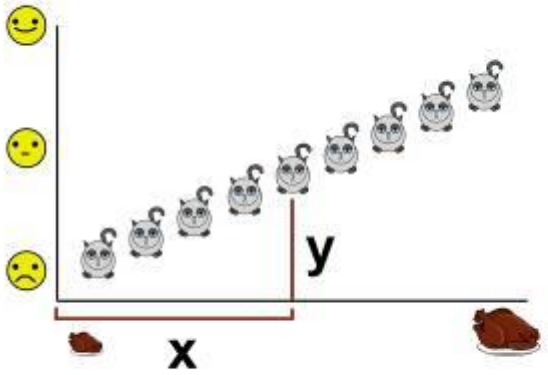


r = 1

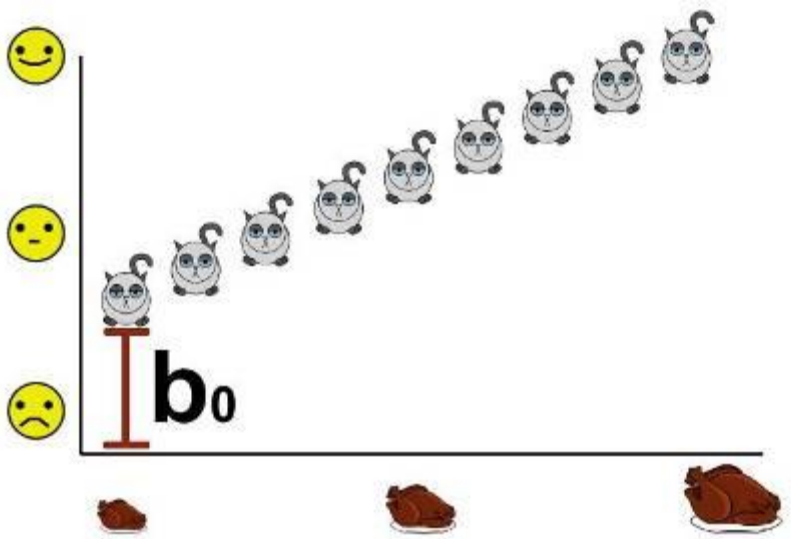


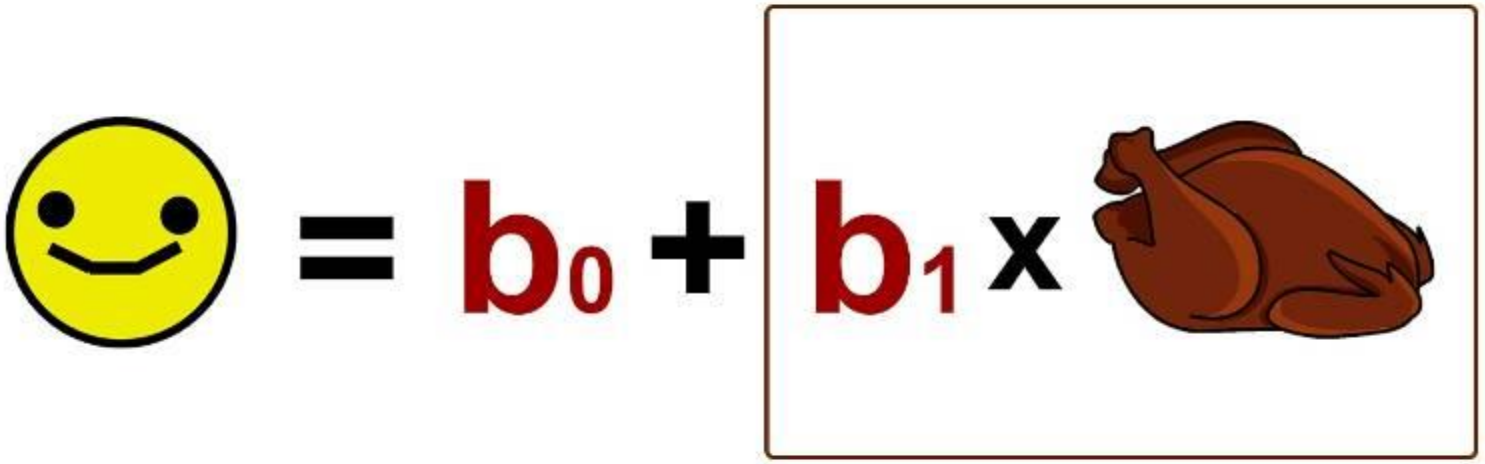
r = 1

- Коэффициент b1 определяется как тангенс угла между линией котиков и оси x. Чем больше этот коэффициент, тем сильнее растет уровень счастья от каждой новой порции.
- Вторая величина, которая может описывать прямую, называется b0. Она показывает насколько счастливы котики, если их совсем не кормить.



$$b_1 = \frac{y}{x}$$

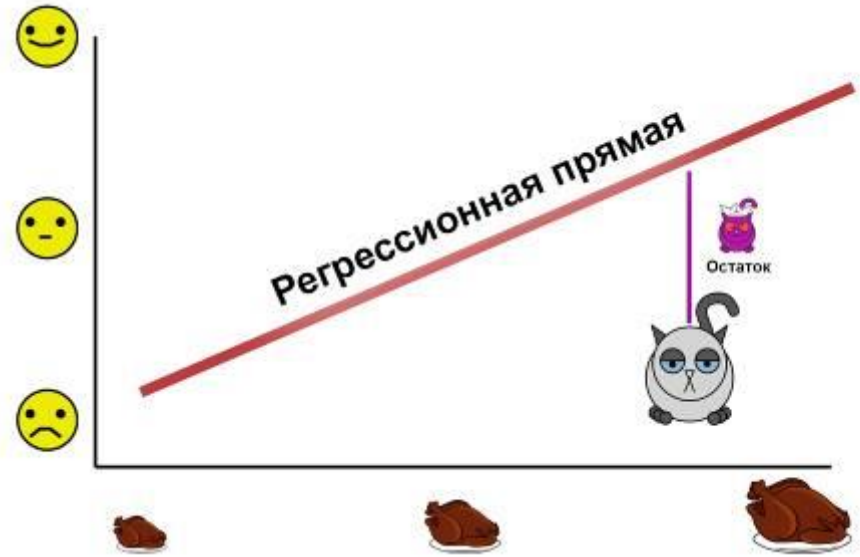


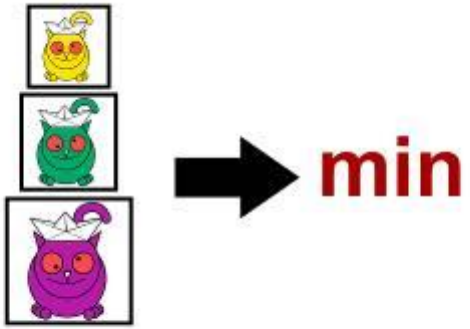


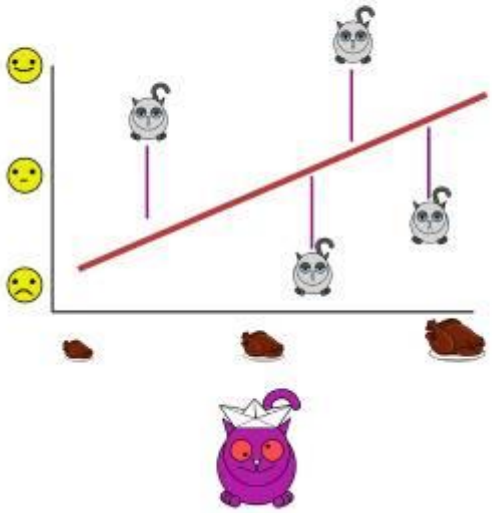
По этой формуле вы сможете выяснить, какие факторы наиболее сильно влияют на котиковое счастье, и предсказывать, насколько будет счастлив тот или иной котик по их значениям.

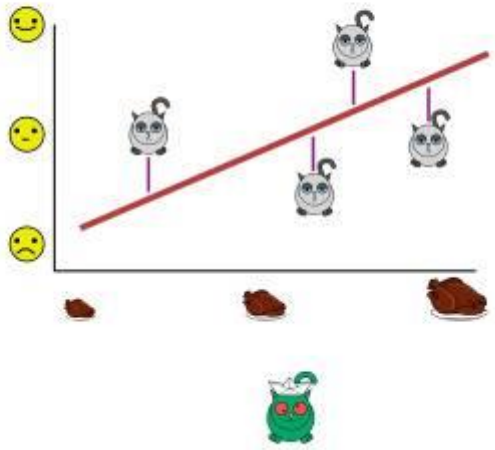
Реальные взаимосвязи мало похожи на прямую линию. Чаще они напоминают собой огурец, а в запущенных случаях – авокадо. Но описывать такие вещи довольно сложно, поэтому статистиками был разработан специальный метод, который позволяет подобрать такую прямую, которая смогла бы заменить этот овощ с минимальными потерями данных. Этот метод называется регрессионным анализом



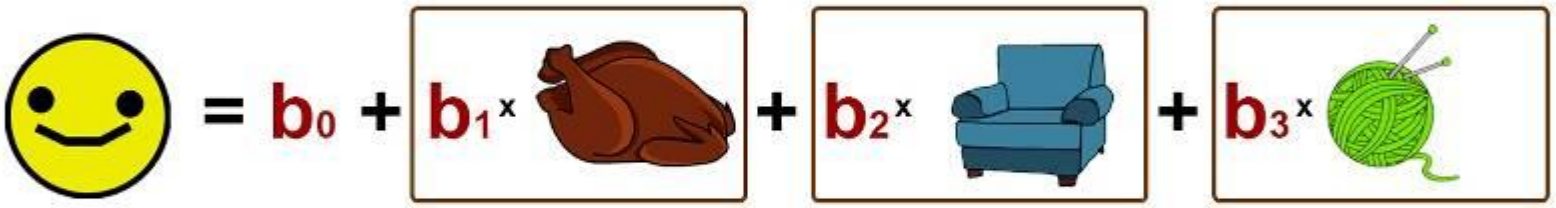








суть регрессионного анализа заключается в том, чтобы подобрать такую прямую, при которой эта сумма была бы минимальной.


= b0 + b1x + b2x + b3x

Спасибо за внимание!