Лекция 2.
Регрессионный анализ

1. Линейная регрессия
2. Множественная линейная регрессия

Регрессионный анализ – количественное представление связи или зависимости между зависимой переменной (***откликом***) и независимой / независимыми переменными (***предикторами***).

Регрессионный анализ используется по двум причинам:

1) описание зависимости между переменными помогает установить наличие *возможной* причинной связи;
2) для установления предиктора для зависимой переменной, так как уравнение регрессии позволяет предсказывать значения зависимой переменной по значениям независимых переменных: **выявление закономерности, выраженной в виде уравнения регрессии**.

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

Эстонский исследователь Я. Микк, изучая трудности понимания текста, установил «формулу читаемости», которая представляет собой множественную линейную регрессию:

$$y = 0,01x_1 + 0,27x_2 + 0,54x_3 - 2,51$$

— оценка трудности понимания текста, где

x1 - длина самостоятельных предложений в количестве печатных знаков,

x2 - процент различных незнакомых слов,

x3 - абстрактность повторяющихся понятий, выраженных существительными.

Линейную регрессию можно отразить уравнением прямой линии:

$Y = b_1 \cdot X + c$, где:

Y – значения признака по линии регрессии, т. е. теоретические значения,

b_1 – угловой коэффициент регрессии,

X – значения признака-фактора (предиктора),

c – свободный член, константа.

Если независимая переменная одна, то регрессия называется парной.

Простейшая парная регрессионная модель – линейная.

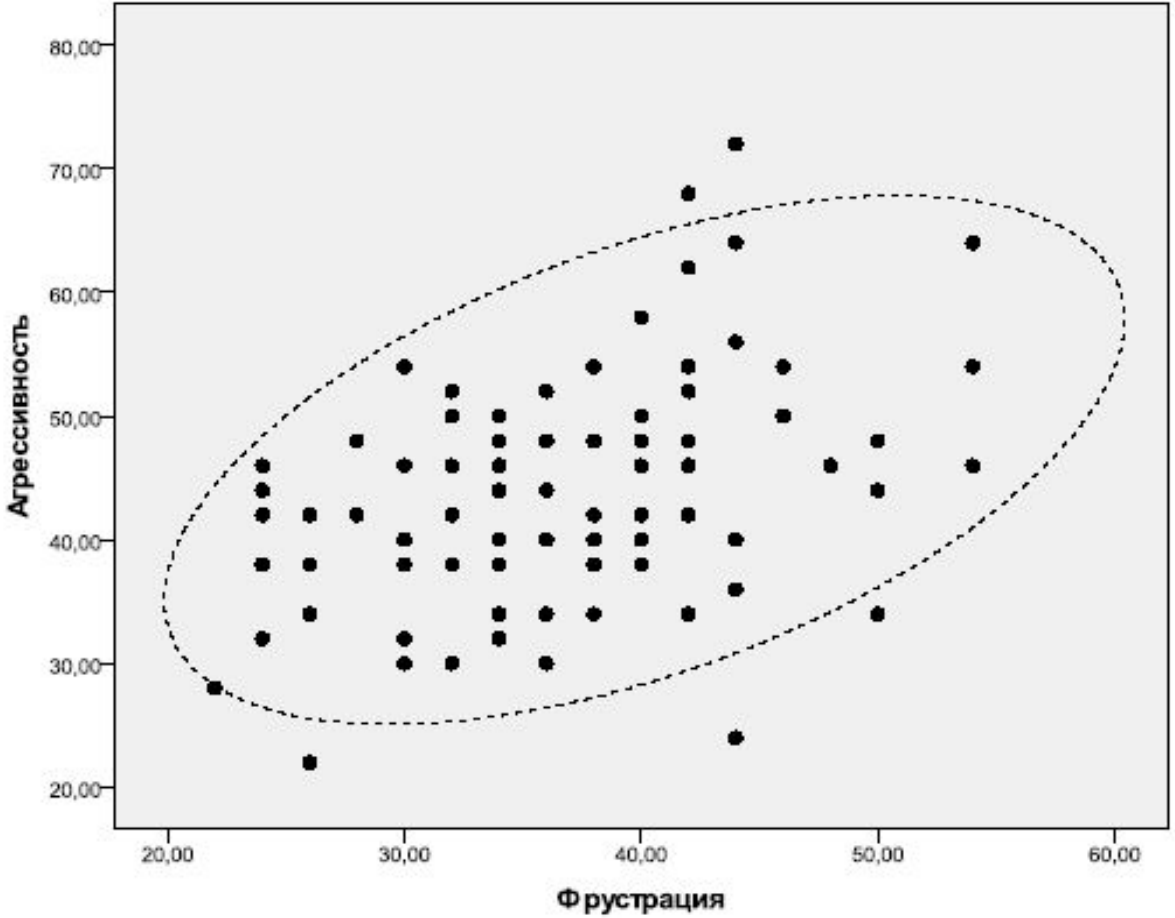


Рис.1.Связь фрустрации и агрессивности

Пример: зависимость агрессивности у спортсменов от фрустрации

Корреляция

		Фрустрация	Агрессивность
Фрустрация	Корреляция Пирсона	1	,418**
	Знч.(2-сторон)		,000
	N	98	98
Агрессивность	Корреляция Пирсона	,418**	1
	Знч.(2-сторон)	,000	
	N	98	98

**. Корреляция значима на уровне 0.01 (2-сторон.).

Сводка для модели

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,418[a]	,174	,166	8,47990

а. Предикторы: (конст) Фрустрация

H – это коэффициент корреляции между зависимой и независимой переменными (r = 0,418),

R-квадрат - коэффициент детерминации (R^2 = 0,174).

R^2 определяет долю вариации одной из переменных, которая объясняется вариацией другой переменной.

В данном случае R^2 =0,174, т.е. доля вариации агрессивности объясняется вариацией фрустрации на 17%, или 17% изменчивости в агрессивности могут быть объяснены различиями во фрустрации среди спортсменов. Остальные 83% объясняются воздействиями других факторов.

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	**24,721**	4,344		5,690	,000
	Фрустрация	**,522**	,117	,418	4,455	,000

а. Зависимая переменная: Агрессивность

$Y = b_1 \cdot X + c$, b_1 – нестандартизированный коэффициент B, c – *константа* ⇒ ***«Агрессивность» = 0,522 · «Фрустрация» + 24,721.***

В уравнение могут быть приняты только те регрессионные коэффициенты, которые статистически значимы (критерий t-Стьюдента). Стандартизированные коэффициенты регрессии (Бета) - показатели вклада каждой переменной в регрессионную модель. В парной регрессии стандартизированный коэффициент - коэффициент корреляции между зависимой и независимой переменными.

Общее назначение **множественной регрессии** (Pearson, 1908) - анализ связи между несколькими независимыми переменными (регрессорами или предикторами) и зависимой переменной (откликом).

Множественная регрессия позволяет исследователю задать вопрос: "что является лучшим предиктором для...". Например, какие индивидуальные качества позволяют лучше предсказать степень социальной адаптации индивида. Термин "множественная" указывает на наличие нескольких предикторов или регрессоров, которые используются в модели:

$$Y = b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3 + \ldots + b_k \cdot X_k + c$$

- При расчетах оценок параметров регрессионной модели применяется метод наименьших квадратов.
- В условиях **нормального распределения ошибок** оценки параметров модели, построенные методом наименьших квадратов, являются оптимальными. Если распределение отличается от нормального, то свойство оптимальности может быть утрачено.

Пример: зависимость агрессивности у спортсменов от фрустрации и тревожности

«Агрессивность» = b_1 ·*«Фрустрация»* + b_2 · *«Тревожность»* + *c,* где:

b_1 – угловой коэффициент регрессии,

b_2 – угловой коэффициент регрессии,

c – свободный член (константа).

Корреляция

		Фрустрация	Агрессивность	Тревожность
Фрустрация	Корреляция Пирсона	1	,418**	,683**
	Знч.(2-сторон)		,000	,000
	N	96	96	96
Агрессивность	Корреляция Пирсона	,418**	1	,432**
	Знч.(2-сторон)	,000		,000
	N	96	96	96
Тревожность	Корреляция Пирсона	,683**	,432**	1
	Знч.(2-сторон)	,000	,000	
	N	96	96	96

**. Корреляция значима на уровне 0.01 (2-сторон.).

Сводка для модели

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,464[a]	,215	,198	8,31284

а. Предикторы: (конст) Фрустрация, Тревожность

H – ***коэффициент множественной корреляции*** между зависимой и набором независимых переменных (0,464), а R-квадрат - коэффициент множественной детерминации (R^2 = 0,215). Он определяет долю вариации одной из переменных, которая объясняется вариацией других переменных, т.е. доля вариации агрессивности объясняется вариацией тревожности и фрустрации на 22%. Остальные 78% объясняются воздействиями других факторов.

Multiple R – коэффициент множественной корреляции. Может принимать значения от 0 до 1 и характеризует тесноту линейной связи между зависимой и всеми независимыми переменными.

- Коэффициент детерминации R^2 измеряет долю разброса относительно среднего значения, которую «объясняет» построенная регрессия.

  Значение R^2 является индикатором степени подгонки модели к данным. Чем ближе коэффициент детерминации к 1, тем лучше регрессия «объясняет» зависимость в данных.
- Значение коэффициента детерминации R^2 возрастает с ростом числа переменных в регрессии, что не означает улучшения качества предсказания. Поэтому для оценки качества подгонки регрессионной модели к наблюдаемым значениям вводится **скорректированный (adjusted)** коэффициент детерминации.

  Различные регрессии (с различным набором переменных) можно сравнивать по этому коэффициенту и принять тот вариант регрессии, для которого он максимален.

Дисперсионный анализ[b]

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
1	Регрессия	1760,024	2	880,012	12,735	,000[a]
	Остаток	6426,601	93	69,103		
	Всего	8186,625	95			

a. Предикторы: (конст) Фрустрация, Тревожность

b. Зависимая переменная: Агрессивность

Значение критерия F-Фишера равно 12,735,

его p-уровень значимости – 0,000.

Это означает, что коэффициент множественной корреляции между зависимой и двумя независимыми переменными статистически значим и модель регрессии может быть содержательно интерпретирована.

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	19,432	4,893		3,971	,000
	Тревожность	,320	,146	,276	2,195	,031
	Фрустрация	,287	,157	,229	1,824	,071

а. Зависимая переменная: Агрессивность

В таблице – стандартизированные коэффициенты регрессии (Бета) – 0,276 и 0,229, значения критерия t-Стьюдента (2,195 и 1,824) и уровни значимости (0,031 и 0,071).
Регрессионный коэффициент, показывающий вклад фрустрации в изменчивость агрессивности, статистически не значим ($p = 0,071 > 0,05$) ⇒ может быть исключен из модели. Линейное уравнение принимает вид парной регрессии: «Агрессивность» = 0,320 · «Тревожность» + 19,432.

- Бета-коэффициенты β - это коэффициенты, которые получатся, если предварительно стандартизовать все переменные к среднему 0 и стандартному отклонению 1. Таким образом, величина этих Бета-коэффициентов позволяет сравнивать относительный вклад каждой независимой переменной в предсказание зависимой переменной.

Множественная регрессия: общая структура и предположения

$$Y = a + b1 * X + b2 * X^2 + \ldots + b3 * X^n$$

$$Y = a + b_1 * X_1 + b_2 * X_2 + \ldots + b_p * X_p$$

- Линейности частных зависимостей;
- Нормальности всех переменных;
- **Нескоррелированности независимых переменных;**
- **Не менее (7)10–20 независимых наблюдений на каждый предиктор в модели;**

Частная корреляция

Частная корреляция - анализ взаимосвязи между двумя величинами при фиксированных значениях остальных величин.

- Частная корреляция – корреляция между двумя переменными, когда одна или больше из оставшихся переменных удерживаются на постоянном уровне. Частная корреляция представляет самостоятельный вклад соответствующей независимой переменной в предсказание зависимой переменной.
- В идеальной регрессионной модели независимые переменные вообще не коррелируют друг с другом. *Если две независимые переменные сильно коррелированы с откликом и друг с другом, то достаточно включить в уравнение только одну из них. Обычно включают ту переменную, значения которой легче и дешевле измерять.*

Пример: у группы спортсменов измерили результат в прыжках в длину (X), массу тела (Y) и силу мышц нижних конечностей (Z). Рассчитали коэффициенты линейной корреляции: XY=0,78, XZ=0,89, YZ=0,95.

Представим, что исследователя интересует "чистая" корреляция между результатами в прыжках в длину и массой тела, исключая влияние на эту взаимосвязь силы мышц нижних конечностей испытуемых.

$$r_{xy\cdot(z)} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

$$r = \frac{0{,}78 - 0{,}89 \cdot 0{,}95}{\sqrt{(1 - 0{,}89^2)(1 - 0{,}95^2)}} = \frac{-0{,}07}{0{,}14} = -0{,}50$$

Отрицательное значение частного коэффициента корреляции свидетельствует о том, что при прочих равных условиях (одинаковой силе мышц нижних конечностей) спортсмены с большей массой тела прыгали бы хуже.

Частные коэффициенты на основе **стандартизированных коэффициентов регрессии** (бета-коэффициентов) дают меру тесноты связи каждого предиктора с показателем (результатом) в чистом виде.

Summary Statistics; DV: NEP (ЭкПс-2016-17-для МногомАнализ.sta)

	Value
Multiple R	0,60909
Multiple R?	0,37099
Adjusted R?	0,35002
F(4,120)	17,69402
p	0,00000
Std.Err. of Estimate	0,47615

Regression Summary for Dependent Variable: NEP (ЭкПс-2016-17-для МногомАнализ.sta) R= ,60908975 R?= ,37099032 Adjusted R?= ,35002333 F(4,120)=17,694 p

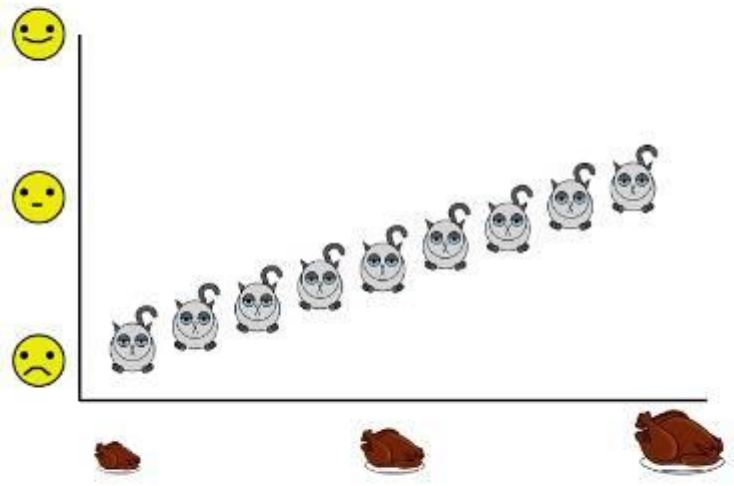
	Beta	Std.Err.	B	Std.Err.	t(120)	p-level
Intercept			-1,86859	0,328474	-5,68869	0,000000
Schultz	0,141544	0,086939	0,05202	0,031952	1,62808	0,106130
идент_прир_nrs	0,112122	0,093408	0,10524	0,087671	1,20034	0,232372
экстрапол_nrs	**0,502318**	**0,075528**	**0,47357**	**0,071206**	**6,65076**	**0,000000**
эмоц оп взаим_nrs	0,079344	0,079295	0,06173	0,061688	1,00062	0,319022

Variables currently in the Equation; DV: NEP (ЭкПс-2016-17-для МногомАнализ.sta)

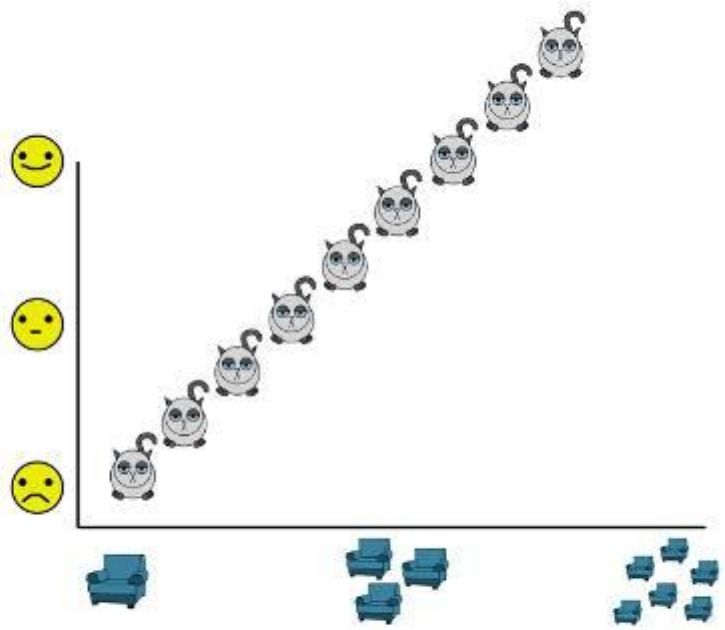
	Beta in	Partial	Semipart	Tolerance	R-square	t(120)	p-level
Schultz	0,14154	0,1470	0,11787	0,69349	0,30650	1,6280	0,1061
идент_прир_nrs	0,11212	0,10892	0,08690	0,60076	0,39923	1,2003	0,2323
экстрапол_nrs	**0,50231**	**0,51896**	**0,48151**	**0,91888**	**0,08111**	**6,6507**	**0,0000**
эмоц оп взаим_nrs	0,07934	0,09096	0,07244	0,83366	0,16633	1,0006	0,3190

Формула счастья котиков

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

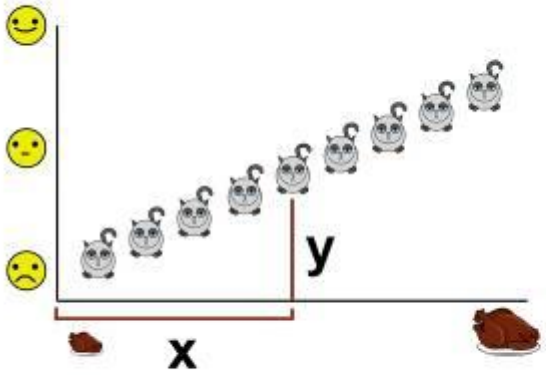


r = 1

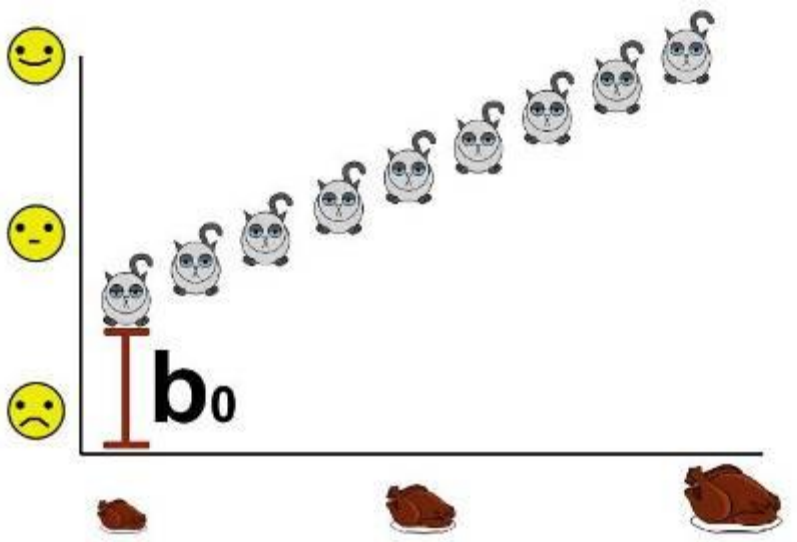


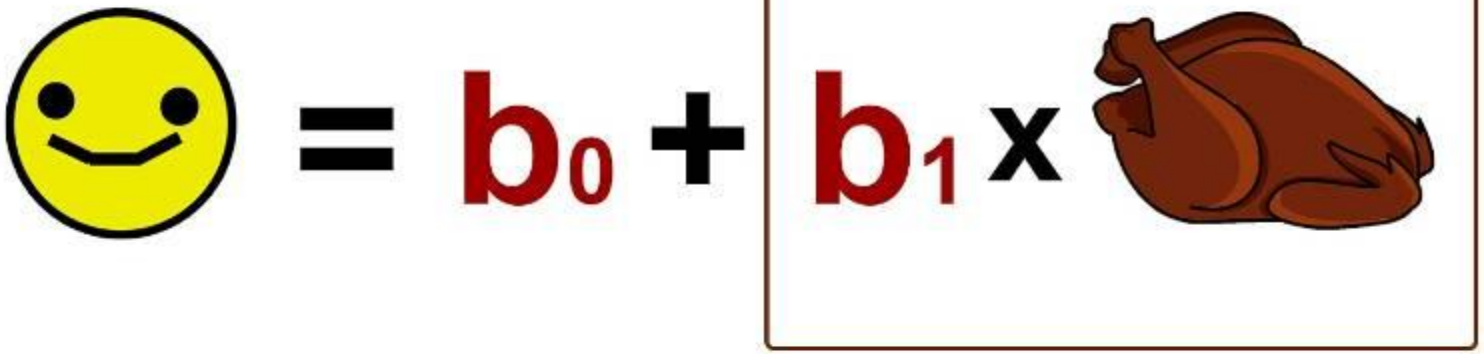
r = 1

- 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.

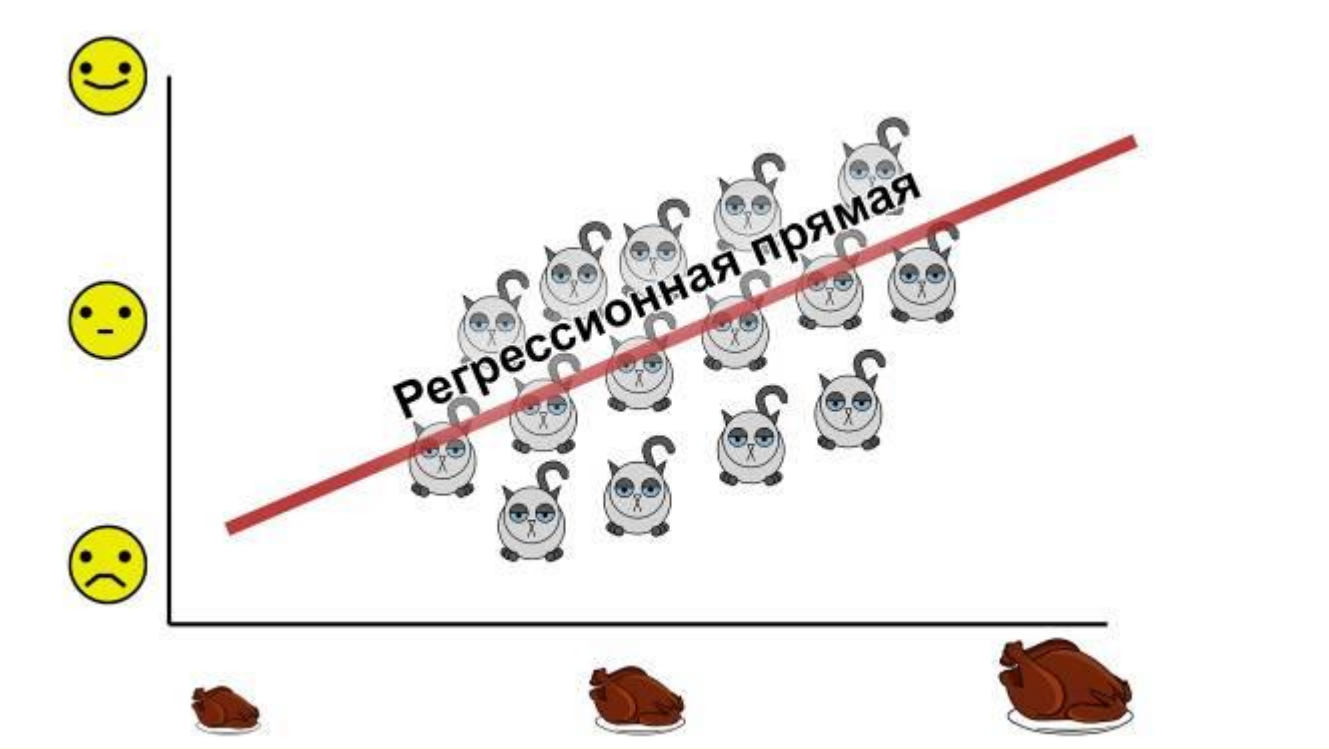


$$b_1 = \frac{y}{x}$$

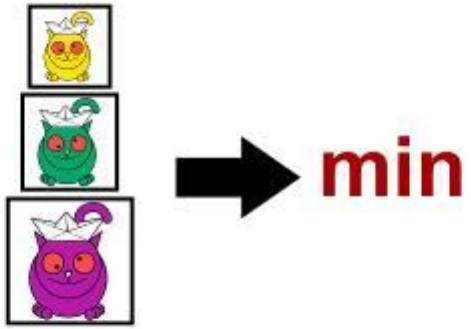


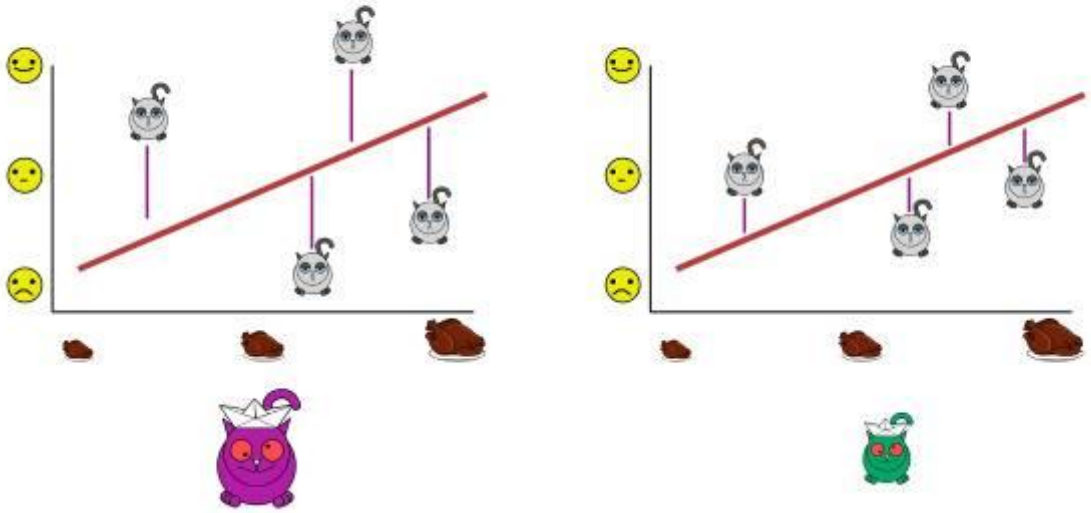

= b0 + b1 x

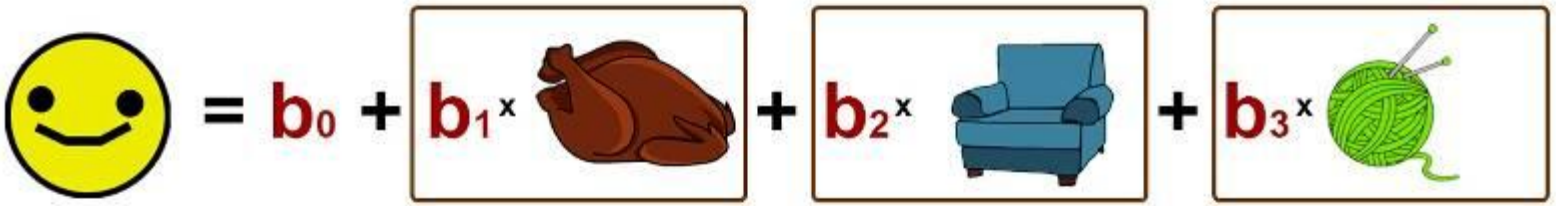
Реальные взаимосвязи мало похожи на прямую линию. Чаще они напоминают собой огурец, а в запущенных случаях – авокадо. Но описывать такие вещи довольно сложно, поэтому статистиками был разработан специальный метод, который позволяет подобрать такую прямую, которая смогла бы заменить этот овощ с минимальными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




Регрессионная прямая
Остаток


min




= b0 + b1x + b2x + b3x

Предположения, ограничения и обсуждение практических вопросов

www.statsoft.ru/home/textbook/modules/stmulreg.html

Предположение линейности. Предполагается, что связь между переменными является линейной. На практике это предположение, в сущности, никогда не может быть подтверждено; к счастью, процедуры множественного регрессионного анализы в незначительной степени подвержены воздействию малых отклонений от этого предположения. Однако всегда имеет смысл посмотреть на двумерные диаграммы рассеяния переменных, представляющих интерес. Если нелинейность связи очевидна, то можно рассмотреть или преобразования переменных или явно допустить включение нелинейных членов.

Предположение нормальности. В множественной регрессии предполагается, что **остатки** (предсказанные значения минус наблюдаемые) распределены нормально (т.е. подчиняются закону нормального распределения). И снова, хотя большинство тестов (в особенности F-тест) довольно робастны (устойчивы) по отношению к отклонениям от этого предположения, всегда, прежде чем сделать окончательные выводы, стоит рассмотреть распределения представляющих интерес переменных. Вы можете построить гистограммы или нормальные вероятностные графики остатков для визуального анализа их распределения.

- **Нормальный вероятностный график остатков** наглядно показывает наличие или отсутствие больших отклонений от высказанных предположений (Стандартный регрессионный анализ в STATISTICA: **http://www.statosphere.ru/blog/115-stat-regress.html**)

Ограничения. Основное концептуальное ограничение всех методов регрессионного анализа состоит в том, что они позволяют обнаружить только числовые зависимости, а не лежащие в их основе причинные (causal) связи.

Например, можно обнаружить сильную положительную связь (корреляцию) между разрушениями, вызванными пожаром, и числом пожарных, участвующих в борьбе с огнем.

Следует ли заключить, что пожарные вызывают разрушения?

Конечно, наиболее вероятное объяснение этой корреляции состоит в том, что размер пожара (внешняя переменная, которую забыли включить в исследование) оказывает влияние, как на масштаб разрушений, так и на привлечение определенного числа пожарных (т.е. чем больше пожар, тем большее количество пожарных вызывается на его тушение).

Хотя этот пример довольно прозрачен, в реальности при исследовании корреляций альтернативные причинные объяснения часто даже не рассматриваются.

Выбор числа переменных. Множественная регрессия предоставляет пользователю "соблазн" включить в качестве предикторов все переменные, какие только можно, в надежде, что некоторые из них окажутся значимыми.

Проблема также возникает, когда и число наблюдений относительно мало. Интуитивно ясно, что едва ли можно делать выводы из анализа вопросника со 100 пунктами на основе ответов 10 респондентов. Большинство авторов советуют использовать, по крайней мере, **от 10 до 20 наблюдений (респондентов) на одну переменную**, в противном случае оценки регрессионной линии будут, вероятно, очень ненадежными и, скорее всего, невоспроизводимыми для желающих повторить это исследование.

- **Принцип парсимонии**: по отношению к регрессорам - чем меньше, тем лучше. Другой регрессор будет позволять объяснить немножко больше, но очень часто это приводит к тому, что наше понимание затуманивается.

Принцип здравого смысла:

регрессор должен иметь логические взаимоотношения с зависимой переменной,

кроме статистических взаимоотношений

Наилучшие регрессионные модели

- Поиск наилучшей регрессионной модели – искусство, у которого нет рецептов. С одной стороны, для получения надёжных прогнозов значений отклика *y* в модель нужно включать как можно больше независимых переменных. С другой стороны, с увеличением их числа возрастает дисперсия прогноза и увеличивается затратность исследования. Некоторые общие требования к регрессионным моделям:
- Регрессионная модель должна объяснять не менее 80 % вариации зависимой переменной, т.е. $R2>0,8$ (что в психологических исследованиях достигается крайне редко)
- Чем меньше сумма квадратов остатков, чем меньше стандартная ошибка оценки и чем больше $R2$, тем лучше уравнение регрессии.
- Коэффициенты уравнения регрессии и его свободный член должны быть значимы по уровню 0,05.
- Остатки от регрессии должны быть без заметной автокорреляции ($r<0,3$), нормально распределены и без систематической составляющей.

Понятие «наилучшая регрессионная модель» является субъективным, так как нет никакой единой статистической процедуры для выбора соответствующего подмножества независимых переменных.

Дополнительные ресурсы

- **http://www.statcats.ru/2016/05/blog-post_10.html**
- http://www.statsoft.ru/home/textbook/modules/stmulreg.html
- Обзорная презентация http://www.myshared.ru/slide/764056/
- http://www.myshared.ru/slide/616696/
- http://pubhealth.spb.ru/SASDIST/MLR.htm
- Для продвинутых ☺: http://forum.disser.ru/index.php?showtopic=2439