Э П С И Л О Н

E P S I L O N

журнал об эконометрике
и не только о ней

Корреляция: простая, частная и условная

Борис Демешев*

5 августа 2015 г.

Аннотация

Корреляция — это способ описать силу линейной зависимости между двумя случайными величинами одним числом. Каков геометрический смысл корреляции? Что такое частная корреляция? Как связаны частная и условная корреляция?

Ключевые слова: корреляция, частная корреляция, условная корреляция, косинус, проекция.

1 Сколько вешать в граммах?

Почему мы измеряем температуру тела с помощью градусника?

1. Измерить температуру очень удобно
2. Это измерение несёт в себе информацию о здоровье

Сложное описание здоровья сводится измерением температуры к одной цифре. Естественно, куча информации теряется в этой цифре и бессмысленно лечить человека, руководствуясь только температурой его тела. Температура 39° говорит, что не так, но что — непонятно. А температура 36.6° ещё не говорит о том, что у человека идеальное здоровье. Однако процедура очень проста и в некоторых ситуациях (например, при обыкновенной простуде) её достаточно для принятия решения о приёме жаропонижающего.

Если бы для измерения температуры нужно было специальное устройство размером с половину комнаты, никто бы дома её не мерял. Простота измерения очень важна!

Подобном образом дела обстоят и с описанием зависимости между случайными величинами. Зависимость между случайными величинами полностью описывается их совместной функцией распределения, $F(x, y) = ¶(X \leq x, Y \leq y)$. Вместо сложной функции распределения мы хотим получить одно число. Некую «силу зависимости». Назовём это мифическое число $Dep(X, Y)$.

Что мы требуем от этого числа?

1. Посчитать это число очень удобно
2. Это число несёт в себе информацию о зависимости

*НИУ ВШЭ, Москва.

В каком смысле «удобно» считать?

Очень часто возникают суммы случайных величин, поэтому было бы здорово, чтобы у суммы легко считалась наша характеристика $Dep(X, Y)$. Проще всего было бы, если бы:

$$Dep(X + Z, Y) = Dep(X, Y) + Dep(Z, Y)$$

И, конечно, мы ждём, что у независимых случайных величин нулевая сила зависимости, $Dep(X, Y) = 0$, а ненулевая сила зависимости, $Dep(X, Y) \neq 0$, была бы возможна только у зависимых случайных величин.

Этим двум требуемым свойствам (простота подсчёта и информация о зависимости) отвечает ковариация.

Определение 1. *Ковариация величин X и Y измеряет ...*

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

здесь рассказать про прямоугольники и площадь с плюсом/минусом?

2 Корреляция по-русски

Обычно в учебниках даётся такое определение корреляции

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}. \tag{2.1}$$

Естественно, возникает вопрос: «С какого перепугу? Почему это мы делим ковариацию на что-то там?»

Мы дадим определение корреляции словами:

Определение 2. *Корреляция между случайными величинами X и Y показывает на сколько своих стандартных отклонений в среднем растёт случайная величина Y при росте случайной величины X на одно своё стандартное отклонение.*

А теперь из этого словесного определения мы получим формулу 1.1. Разложим величину Y на два слагаемых. Первое слагаемое вбирает в себя всю ту часть Y, которая линейно зависит от X, а второе — всё оставшееся:

$$\frac{Y}{\sigma_Y} = \rho \cdot \frac{X}{\sigma_X} + \varepsilon$$

В этой формуле видно, что с ростом X на одно стандартное отклонений σ_X правая часть изменится в среднем на ρ, и, следовательно, величина Y в среднем изменится на $\rho \cdot \sigma_Y$.

Естественно, мы хотим, чтобы с ростом X величина ε в среднем не менялась, то есть хотим нулевую «силу зависимости» между ними, $\mathrm{Cov}(X, \varepsilon) = 0$.

$$\mathrm{Cov}\left(X, \frac{Y}{\sigma_Y} - \rho \cdot \frac{X}{\sigma_X}\right) = 0$$

По свойствам ковариации получаем

$$\mathrm{Cov}(X,Y)/\sigma_Y - \rho\,\mathrm{Cov}(X,X)/\sigma_X = 0$$

И, тадам, выражаем корреляцию, ρ:

$$\rho = \frac{\mathrm{Cov}(X,Y)/\sigma_Y}{\mathrm{Cov}(X,X)/\sigma_X} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

Несмотря на асимметричность исходного разложения (эпсилон прибавляется в правой части уравнения к величине X), результирующая формула для корреляции получается симметричной. Из этого следует, что ровно такой же результат получится, если начать с разложения:

$$\frac{X}{\sigma_X} = \rho \cdot \frac{Y}{\sigma_Y} + \varepsilon$$

Из определения неочевидно, что корреляция лежит в пределах от -1 до 1. ...

Стоит обратить внимание на немного контр-интуитивный факт. Если бы зависимость между X и Y была бы жесткой детерминистической, и с ростом X на единицу величина Y росла бы на Δ, то с ростом Y на единицу величина X росла бы на $1/\delta$. Для случайных величин обращения не происходит. Если с ростом X на одно своё стандартное отклонение величина Y в среднем растёт на ρ своих стандартных отклонений, то и с ростом Y на одно своё стандартное отклонение величина X в среднем растёт на ρ своих стандартных отклонений.

? парадокс возвращения к среднему ?

3 Геометрический смысл корреляции

Давайте рисовать случаные величины векторами-стрелочками! Не в том смысле, что у стрелочки случайное направление или длина, а в том смысле, что направление и длина стрелочки описывают характеристики этой случайной величины.

Любую геометрию можно задать, задав скалярное произведение. Действительно, если мы умеем считать скалярное произведение двух любых векторов, $<\vec{a},\vec{b}>$, то длина вектора считается ровно как в 9-м классе:

$$|\vec{a}| = \sqrt{<\vec{a},\vec{a}>}$$

И также любой девятикласник помнит, что косинус угла между векторами считается как

$$\cos(\vec{a},\vec{b}) = \frac{<\vec{a},\vec{b}>}{|\vec{a}||\vec{b}|}$$

Мы определим скалярное произведение двух случайных величин как их ковариацию:

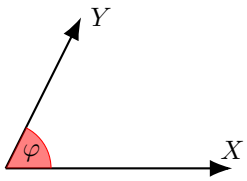
$$<X,Y>=\mathrm{Cov}(X,Y)$$

При таком подходе длиной случайной величины окажется стандартное отклонение:

$$\sqrt{\mathrm{Cov}(X,X)} = \sqrt{\mathrm{Var}(X)} = \sigma_X$$

А корреляция окажется косинусом угла между случайными величинами:

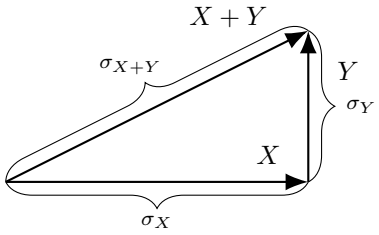
$$\cos\varphi = \cos(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \mathrm{Corr}(X,Y)$$



Значит в нашей геометрии длина стрелочки — стандартное отклонение случайной величины, а косинус угла между двумя стрелочками — это корреляция двух случайных величин. Дисперсия, следовательно, это квадрат длины случайной величины. Перпендикулярными случайными величинами будут те, косинус угла между которыми равен нулю, то есть некоррелированные.

Например, сформулируем в данной геометрии теорему Пифагора. Если случайные величины X и Y перпендикулярны (корреляция или ковариация равны нулю), то дисперсия их суммы (квадрат длины гипотенузы) равен сумме их дисперсий (сумму квадратов длин катетов):

$$\mathrm{Var}(X+Y) = \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y = \mathrm{Var}(X) + \mathrm{Var}(Y)$$



Введение геометрии позволяет говорить о проекции. Например, можно спроецировать случайную величину Y на множество случайных величин пропорциональных величине X, $\{cX | c \in \mathbb{R}\}$. Если на обычной плоскости спроецировать вектор $\vec{a}$ на прямую, порожденную вектором $\vec{b}$, то получится $\cos(\vec{a},\vec{b}) \cdot \vec{b}$. По аналогии, если спроецировать случайную величину Y на множество $\{cX | c \in \mathbb{R}\}$, то получится $\mathrm{Corr}(X,Y) \cdot X$. Другими словами, среди случайных величин пропорциональных X величина $\hat{Y} = \mathrm{Corr}(X,Y) \cdot X$ — самая похожая на величину Y.

Понятие проекции позволяет интепретировать квадрат корреляции. Квадрат косинуса равен отношению квадрата длины прилежащего катета $\mathrm{Var}(\hat{Y})$ к квадрату гипотенузы $\mathrm{Var}(Y)$.

(картинка)

Следовательно, $\mathrm{Corr}(X,Y)^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)}$, то есть квадрат корреляции показывает долю дисперсии Y, которую можно объяснить с помощью величин пропорциональных X.

4 Корреляция и независимость

Теорема 1. *Случайные величины X и Y независимы тогда и только тогда, когда некоррелированы любые функции $f(X)$ и $g(Y)$.*

Другими словами для независимости X и Y необходима некоррелированность пар X и Y, X^2 и $\cos(Y)$, $\exp(X)$ и $1/Y$, и так далее. Из этого следует, что некоррелированность X и Y является необходимым, но недостаточным условием для независимости.

Можно выделить три «степени» независимости случайных величин X и Y:

Некоррелированность Y и X	$\mathrm{Corr}(X,Y)=0$
$\mathbb{E}(Y\|X)=\mathbb{E}(Y)$	$\mathrm{Corr}(f(X),Y)=0$ для всех $f()$
Независимость Y и X	$\mathrm{Corr}(f(X),g(Y))=0$ для всех $f()$ и $g()$

Многие ошибочно считают, что если величина X имеет нормальное распределение $N(\mu_X,\sigma_X^2)$ и величина Y имеет нормальное распределение $N(\mu_Y,\sigma_Y^2)$, и X и Y некоррелированы, то они независимы. Это неверно.

Контрпример. Случайная величина X имеет стандартное нормальное распределение $N(0;1)$, случайная величина Z независима от X и равновероятно принимает значения -1 и $+1$. Определим величину Y как их произведение, $Y=XZ$.

В этом примере величины X и Y зависимы, так как $|X|=|Y|$. Однако Y распределена нормально стандартно и $\mathrm{Corr}(X,Y)=0$.

Правильная теорема звучит так:

Теорема 2. *Если некоррелированные случайные величины X и Y имеют совместное нормальное распределение, то X и Y независимы.*

Попутно упомянем ещё одно неожиданное свойство предъявленного контрпримера. Если случайные величины нормальны по отдельности, то вполне возможно, что их сумма ненормальна. Для пары величин, имеющих совместное нормальное распределение, это невозможно.

5 Частная корреляция

Определение 3. *Частная корреляция между величинами X и Y при фиксированной величине Z показывает на сколько своих стандартных отклонений σ_Y в среднем вырастет Y при росте величины X на одно своё стандартное отклонение σ_X и постоянном значении величины Z.*

Для нахождения частной корреляции используется разложение

$$\frac{Y}{\sigma_Y}=\rho_{XY|Z}\cdot\frac{X}{\sigma_X}+\rho_{YZ|X}\frac{Z}{\sigma_Z}+\varepsilon$$

Альтернативный подход к подсчёту частной корреляции следующий:

1. Спроецируем X на множество величин, некоррелированных с Z. Получим $\tilde{X}$.
2. Спроецируем Y на множество величин, некоррелированных с Z. Получим $\tilde{Y}$.
3. Частная корреляция между X и Y при фиксированной Z — это обычная корреляция между $\tilde{X}$ и $\tilde{Y}$.

(картинка ...)

Два подхода к определению частной корреляции эквивалентны в силу теоремы Фриша-Ву-Ловелла (Frisch–Waugh–Lovell). Обычно эта теорема формулируется применительно к регрессии, а здесь мы приведём её вариат для случайных величин.

Теорема 3. *Если имеют место разложения:*

$$Y = a_1 Z_1 + a_2 Z_2 + \ldots + a_n Z_n + \tilde{Y},\ \text{где}\ \ \tilde{Y} \perp Z_1, Z_2, \ldots, Z_n$$

и

$$X = b_1 Z_1 + b_2 Z_2 + \ldots + b_n Z_n + \tilde{X},\ \text{где}\ \ \tilde{X} \perp Z_1, Z_2, \ldots, Z_n$$

То в разложениях

$$\tilde{Y} = d\tilde{X} + \varepsilon,\ \text{где}\ \ \varepsilon \perp \tilde{X}$$

и

$$Y = c_1 Z_1 + c_2 Z_2 + \ldots + c_n Z_n + dX + u,\ \text{где}\ \ u \perp Z_1, Z_2, \ldots, Z_n, X$$

коэффициенты при $\tilde{X}$ и X совпадают.

6 Условная корреляция

Определение 4. *Условная корреляция между величинами X и Y при известном значении величины Z показывает на сколько своих стандартных отклонений σ_Y в среднем вырастет Y при росте величины X на одно своё стандартное отклонение σ_X при заданном значении величины Z.*

Следует подчеркнуть одно существенное отличие условной корреляции от обычной и частной. Обычная и частная корреляция являеются константами. Условная корреляция $\mathrm{Corr}(X, Y|Z)$ является функцией от Z. Величина Z является случайной, поэтому и условная корреляция $\mathrm{Corr}(X, Y|Z)$ является случайной величиной.

Здесь регрессионное определение????

Чуть более формальное определение:

Определение 5.

$$\mathrm{Corr}(X, Y|Z) = \frac{\mathrm{Cov}(X, Y|Z)}{\sqrt{\mathrm{Var}(X|Z)\,\mathrm{Var}(Y|Z)}},$$

где $\mathrm{Cov}(X, Y|Z) = \mathbb{E}(XY|Z) - \mathbb{E}(X|Z)\mathbb{E}(Y|Z)$ и $\mathrm{Var}(X|Z) = \mathbb{E}(X^2|Z) - (\mathbb{E}(X|Z))^2$

Пример подсчета частной и условной корреляций.

Пример 1.

Закон распределения случайных величин X_1, X_2, X_3 задан двумя таблицами:

$X_3 = 0$

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.06	0.08
$X_1 = 1$	0.24	0.32

$X_3 = 1$

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.1	0
$X_1 = 1$	0	0.2

Найдите условную корреляцию $\mathrm{Corr}(X_1, X_2|X_3)$ и частную корреляцию $\mathrm{pCorr}_{X_3}(X_1, X_2)$.

Решение.

Эти две таблички на самом деле реализуют простую мысль: при $X_3 = 1$ величины X_1 и X_2 связаны детерминистически линейно, а при $X_3 = 0$ величины X_1 и X_2 независимы.

Считаем две вспомогательные условные корреляции, $\mathrm{Corr}(X_1, X_2|X_3 = 0) = 0$, $\mathrm{Corr}(X_1, X_2|X_3 = 1) = 1$.

Отсюда получаем, что $\mathrm{Corr}(X_1, X_2|X_3) = X_3$. Для дискретных случайных величин запись условного ожидания не однозначна, и, например, ответ $\mathrm{Corr}(X_1, X_2|X_3) = X_3^2$ также будет верным.

Пример 2.

Величины X_1, X_2, X_3 имеют совместное нормальное распредлеение с математическим ожиданием $\mathbb{E}(X) = (1, 2, -3)'$ и ковариационной матрицей

$$\mathrm{Var}(X) = \begin{pmatrix} 9 & 2 & -1 \\ 2 & 16 & 1 \\ -1 & 1 & 4 \end{pmatrix}$$

Найдите условную корреляцию $\mathrm{Corr}(X_1, X_2|X_3)$ и частную корреляцию $\mathrm{pCorr}_{X_3}(X_1, X_2)$.

Решение.

...

В данном примере частная и условная корреляция совпали. Это одно из приятных свойств многомерного нормального распределения:

Теорема 4. *Если величины X, Y и Z имеют совместное нормальное распределение, то частная и условная корреляции совпадают.*

Пример 3. AR(1) процесс

7 Выборочные характеристики

В теории обычную корреляцию и частную корреляцию можно посчитать, если известен закон распределения случайных величин. На практике закон распределения не известен, однако доступны наблюдения. Как по имеющимся наблюдениям оценить неизвестные корреляции?

Несколько способов оценки корреляции

Здесь пара картинок: википедийная с корреляциями и два ряда случайного блуждания/тренда

Несколько способов оценки частной корреляции

Доказательство от принцессы

Борис Демешев*

20 сентября 2015 г.

Аннотация

Доказательство от принцессы — частный случай доказательства от противного

Ключевые слова: доказательство от противного, принцесса, доказательство.

1 Доказательство от противного

Как устроено классическое доказательство от противного? Берём кого-нибудь противного, и пусть он доказывает. Допустим нам нужно доказать, что утверждение A верно. Мы, наоборот, предполагаем, что A неверно. Далее каким-нибудь образом приходим к противоречию и, таким образом, получаем вывод, что наше допущение A было ложно.

Довольно часто доказательство от противного используется для того, чтобы доказать, что какой-нибудь объект X не существует. В этом случае очень удобно использовать предлагаемое доказательство от принцессы. Мы представляем себе принцессу, которая замуж не хочет, а по традиции должна объявить конкурс для претендентов руку и сердце. И она объявляет: «Тот, кто принесёт мне X, сможет на мне жениться!». А дальше остаётся объяснить, как она будет аргументированно отказывать каждому претенденту.

2 Пара примеров

Классический пример доказательства от принцессы — доказательство того, что максимальное простое число не существует. Принцесса объявляет: «Тот, кто принесёт мне самое большое просто число во Вселенной, получит меня в жёны!». И к примеру приходит принц и приносит ей p_n. А она ему в ответ: «Не пойду я за тебя замуж, ведь простое число $p_1 \cdot p_2 \cdot p_3 \cdot \ldots \cdot p_n + 1$ больше чем ты принёс!» Так принцесса отказывает всем ухажёрам, а, следовательно, наибольшего простого числа не существует.

Идея доказательства от принцессы возникла так. Я иногда веду вводный курс стохастического анализа для экономистов. Если требуется и позволяет время, то рассказываю про мощности множеств и, в частности, про то, что множество последовательностей из 0 и 1 несчётное. И в нём есть один тонкий момент. Если

*НИУ ВШЭ, Москва.

проводить доказательство в общем виде с произвольными буквами, то оно слишком тяжеловесно. Если проводить на конкретном примере, то возникает вопрос, а почему это доказательство. И принцесса замечательно решает проблему доказательства на частном примере!

Принцесса объявляет: «Тот, кто занумерует натуральными числами все бесконечные последовательности из 0 и 1, получит меня в жёны». И, к примеру, приходит принц датский и говорит: «Я занумеровал!» И предъявляет листочек, на котором все последовательности занумерованы:

1. 000000000...
2. 011001010...
3. 101000000...
4. 010011010...

...

Как принцессе отказать принцу датскому? Она выбирает диагональные элементы этих последовательностей 0110... Затем меняет 0 на 1, а 1 на 0, получая 1001... И спрашивает принца датского: «А последовательность 1001... у Вас под каким номером?» И принц датский начинает перебирать. Под первым номером не может идти, так как первой цифрой отличается, под вторым номером не может идти, так как вторым номером отличается... И принц датский трагично вынужден признать, что эту последовательность он забыл занумеровать. И подобным образом принцесса сможет отказать всем претендентам, а значит множество последовательностей несчётно.

ШАД и линал

Артём Филатов*

23 сентября 2015 г.

Аннотация

Ключевые слова: ШАД, линейная алгебра.

1 Кратко про шад

В 2007 году компания Яндекс основала в своих стенах Школу Анализа Данных. Школа была создана для подготовки специалистов в области анализа больших данных, машинного обучения и других смежных дисциплин. Ежегодно в апреле начинаются экзамены, которые проходят в три этапа: онлайн – тест, письменный экзамен и собеседование. Письменный экзамен включает в себя задачи по теории вероятности, алгоритмам, линейной алгебре, математическому анализу и комбинаторике. Предлагаю вам разбор нескольких занимательных задач по линейной алгебре из письменных экзаменов прошлых лет.

2 Задачи по линейной алгебре из шАДовских экзаменов

Задача №1

Дана матрица A размера $n \times n$, где $a_{i,j} = (i-j)^2, i, j = 1, \ldots, n$. Найдите ранг матрицы A.

Решение:

Посмортим, как выглядит наша матрица.

$$A = \begin{pmatrix} 0 & 1 & 4 & \cdots & (n-2)^2 & (n-1)^2 \\ 1 & 0 & 1 & 4 & \cdots & (n-2)^2 \\ 4 & 1 & 0 & 1 & \cdots & (n-3)^2 \\ 9 & 4 & 1 & 0 & \cdots & (n-4)^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (n-1)^2 & (n-2)^2 & (n-3)^2 & (n-4)^2 & \cdots & 0 \end{pmatrix}$$

*НИУ ВШЭ, Москва.

Из условия каждый элемент матрицы A равен $(i-j)^2 = i^2 - 2ij + j^2$. Но у матрицы из элементов i^2 ранг 1, у матрицы из элементов j^2 ранг тоже единица. Посмотрим на матрицу, образованную ij:

$$\begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ 2 & 4 & 6 & \cdots & 2n \\ 3 & 6 & 9 & \cdots & 3n \\ \cdots & \cdots & \cdots & \cdots & \end{pmatrix}$$

Ее ранг также не превосходит 1. Нам известно, что $\operatorname{rank}(A+B) \leq \operatorname{rank}(A) + \operatorname{rank}(B)$, следовательно $\operatorname{rank}(A) \leq 3$. Но можно показать, что у нас есть ненулевые миноры 3 порядка, следовательно $\operatorname{rank}(A) = 3$.

Задача №2

Дана матрица из нулей и единиц, причем для каждой строки матрицы верно следующее: если в строке есть единицы, то они все идут подряд. Докажите, что определитель такой матрицы равен 0 или ± 1.

Решение:

Посмотрим на то, как выглядит одна из наших матриц:

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Переставим строки так, чтобы образовать некое подобие ступенчатой матрицы.

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Что произойдет с определителем? Он либо не изменился, либо изменил знак, так как перестановка строк меняет знак определителя на противоположный. Теперь сделаем следующее: если позиции первых единиц у строк совпали, то вычтем из той в которой больше единиц, ту в которой меньше единиц. На определитель данное преобразование никак не влияет.

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Переставляя строки и повторяя данную процедуру, мы получим ступенчатую матрицу, которая будет либо вырождена, либо иметь единицы на диагонали. А так

как в такой матрице $\det(A) = \prod_{i=1} a_{i,i} = 1$, то детерминант исходной матрицы равен либо 0, либо ± 1.

Задача №3

Опишите все невырожденные вещественные матрицы A, для которых все элементы матриц A и A^{-1} неотрицательны.

Решение:

Пусть исходная невырожденная матрица A заполнена некотрыми элементами $a_{i,j}$, а обратная к ней A^{-1} элементами $b_{i,j}$. Как известно, $AA^{-1} = E$. Значит, произведение первой строки на первый столбец должно дать 1:

$$a_{1,1} \cdot b_{1,1} + a_{1,2} \cdot b_{2,1} + \cdots + a_{1,n} \cdot b_{n,1} = 1$$

Но произведение первой строки на все остальные столбцы должно дать 0, также нам известно, что все элементы матриц неотрицательны, значит если $a_{1,i} \neq 0$, то $b_{i,j} = 0, j = 2 \dots n$. Это должно быть выполнено для всех $a_{i,j}$. Формально:

$$a_{i,j} \neq 0 \Rightarrow b_{j,z} = 0, z = 1 \dots (i-1), (i+1) \dots n$$

Докажем, что нет такой матрицы A с двумя и более положительными элементами в одном столбце:

Зафиксируем столбец j. Предположим мы встретили первый ненулевой элемент, тогда все кроме одного элементы в j строке матрицы A^{-1} равны 0. Предположим, что мы встретили второй положительны элемент, тогда он занулит все элементы кроме одного, включая тот, который сы не занулили в первый раз. Следовательно, мы получили, что $b_{j,z} = 0$ для всех $z = 1 \dots n$. Но это невозможно, так как это означало бы, что все алгебраические дополнения в некой строке матрицы A равны 0 ($b_{i,j} = \frac{Alg_{i,j}}{\det(A)}$), а следовательно и определитель.

Из всего сказанного следует, что единственно законной матрицей A будет такая матрица, в столбцах которых по одному положительному элементу. Элементарными преобразованиями такая матрица приводится к диагональному виду. Мы показали, что все элементы обратной матрицы зануляться, кроме тех, которые образуют 1 в произведении с ненулевыми элементами матрицы A, следовательно обратная матрица будет иметь аналогичный вид.

Задача №4

Имеется некоторый ненулевой вектор – столбец v. Найти все собственные значения матрицы $v \cdot v^T$.

Решение:

Первым делом необходимо понять сколько собственных значений нужно найти. Нам известно, что $\operatorname{rank}(A) \cdot \operatorname{rank}(B) \leq \min \operatorname{rank}(A), \operatorname{rank}(B)$. Следовательно итоговая матрица будет иметь ранг равный единице. Ранг при замене базиса не изменяется, тогда мы можем перейти к диагональному виду матрицы с базисом из собственных векторов, где на диагонали будет лишь одно собственное значение. Собственные значения также не изменяются при замене базиса. Осталось его найти!

Для того, чтобы найти собственное значение (можно догадаться чему оно равно) воспользуемся ещё одним интересным свойством. Оказывается, что след матрицы (сумма диагональных элементов) равен сумме собственных значений матрицы с

учётом кратности. Легко увидеть, что сумма диагональных элементов это скалярное произведение вектора на самого себя.

Следовательно, мы имеем собственное значение $v^T \cdot v$, и нулевое собственное значение кратности $n - 1$.

Измерение и наглядное представление практической значимости регрессионных связей

Кирилл Фурманов*

12 января 2016 г.

Аннотация

Ключевые слова:

.

«Over time we learn about and use fancier and more abstract regression models... The utility of these fancier models diminishes if we have greater difficulty interpreting and visualizing the results»[1]

Среди экономических исследований значительную долю составляют работы, направленные на определение детерминант какой-либо интересующей исследователя величины (заработной платы, продолжительности периода безработицы и т.п.). Основным инструментом анализа в таких случаях оказывается, как правило, модель множественной регрессии. Интерпретация результатов регрессионного анализа сводится, по большей части, к двум аспектам: истолкованию оценок коэффициентов регрессии и их статистической значимости. Во многих моделях (особенно нелинейных) оценки с трудом поддаются интерпретации, что побуждает обращаться к дополнительным средствам: расчёту предельных эффектов, сравнению прогнозных значений объясняемой переменной при различных значениях детерминант, построению графиков функции отклика. Однако функции отклика — характеристика исключительно динамических моделей, а предельные эффекты и различия в прогнозах зависят от того, при каких уровнях объясняющих переменных их рассчитывать. Результаты таких расчётов сложно систематизировать, свести в одну ясную картину связи переменной отклика с регрессорами, так что основная цель

*Кафедра математической экономики и эконометрики, НИУ ВШЭ, Москва.

[1] Michael N. Mitchell. Interpreting and Visualizing Regression Models Using Stata. Stata Press, 2012.

моделирования — сведение большого объёма информации к небольшому числу интерпретируемых параметров — не достигается. Что касается значимости оценок, то значительное число недоразумений и вольных истолкований (среди них — подмена статистической значимостью практической важности) позволяют утверждать, что неверная интерпретация статистической значимости — одна из наиболее распространённых ошибок в статистике (Good, Hardin, 2012). В действительности, применение аппарата проверки статистических гипотез даёт весьма скудную информацию о практической важности связи по многим причинам:

- «однобокость» вывода (можно обнаружить связь, но не её отсутствие),
- невозможность ранжировки объясняющих переменных по значимости их вклада,
- возможность высокой статистической значимости совершенно незначительных с практической точки зрения коэффициентов,
- сомнительная применимость вероятностных методов к не экспериментальным данным.

В этой статье я попытаюсь показать, как можно дополнить традиционные способы представления оценок регрессионных моделей и приблизить исследователя к оцениванию практической значимости статистических связей. Конечно, практическую значимость невозможно формализовать — в каждой реальной задаче есть свои особенности, определяющие важность модельных коэффициентов. Можно лишь снабдить исследователя набором способов представления данных и сведения их к небольшому количеству осмысленных параметров, которые помогли бы ему самому оценить значительность выявленных связей. Здесь я ограничусь измерением вклада объясняющих переменных в разброс отклика, то есть рассмотрю способы ответить на вопрос: в какой мере различия между наблюдениями в величине отклика могут быть связаны со значениями некоторого регрессора.

Прежде чем перейти к сути, рассмотрим ещё одну опасность в интерпретации статистических оценок — подмену практической важности связи её теснотой. Рисунок **??** даёт пример связи тесной, при которой одна из величин почти не изменяется при изменении другой (чёрные кружки, линия регрессии со слабым наклоном), и пример связи куда менее тесной, при которой, однако, средний уровень одной из величин заметно зависит от значений другой (квадраты с белой заливкой, линия регрессии с большим наклоном).

После рассмотрения этого рисунка может возникнут мысль, что практическую значимость отражает величина коэффициента регрессии, однако обратим внимание на два факта:

- Коэффициент регрессии зависит от единиц измерения регрессора. Всегда можно подобрать такие единицы измерения, чтобы какой-либо регрессор имел наибольший коэффициент и, таким образом, представлялся самым

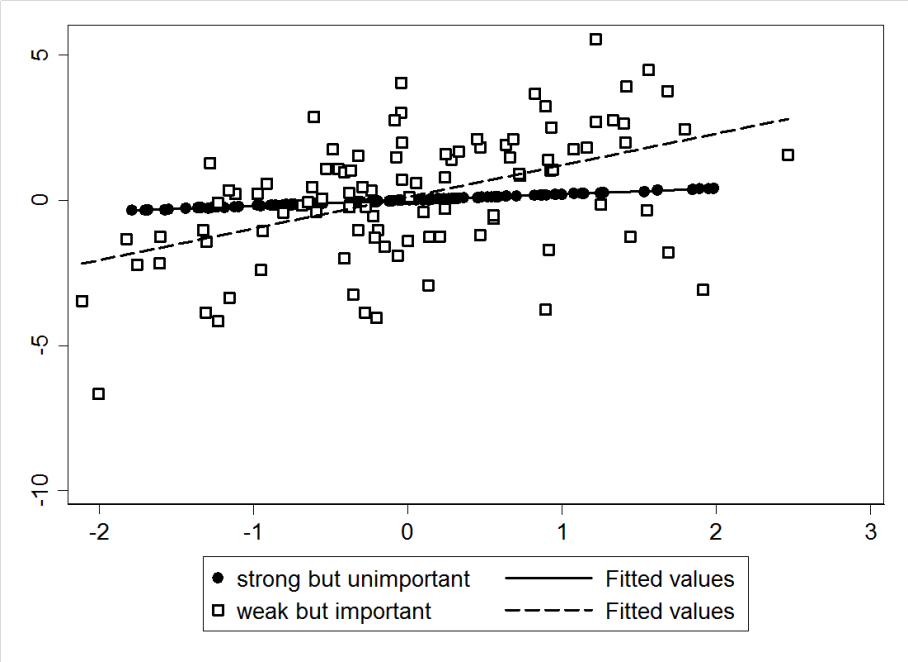


Рис. 1. Название рисунка

важным. Единицы измерения отклика тоже играют роль, но не мешают сравнению вклада объясняющих переменных. Хотя, пожалуй, стоит сделать уточнение: сравнивая два облака на рисунке, мы предполагаем, что речь идёт о связи одних и тех же статистических признаков, единицы измерения которых для обоих случаев совпадают.

- Большой коэффициент может стоять при почти не варьирующемся признаке, так что вклад этого признака в разброс объясняемой величины будет невелик.

Решение этих проблем известно: чтобы ранжировать регрессоры по уровню важности их вклада в разброс объясняемой переменной можно рассчитать стандартизованные коэффициенты регрессии[2]. Стандартизованный коэффициент (иногда используется термин «бета-коэффициент») определяется так:

$$\beta_j^* = \beta_j \frac{\sigma_{x_j}}{\sigma_y}$$

Умножение на стандартное отклонение соответствующего регрессора и деление на стандартное отклонение отклика приводят к тому, что стандартизо-

[2] Существует множество других мер относительной важности, которые изложены в обзорах (Kruskal, 1984), (Nathans et al., 2012), (Johnson, LeBreton, 2004), (Soofi, 2000), но они, по большей части, основаны скорее на измерении тесноты связи, способности объясняющей переменной снизить необъяснённую часть разброса отклика.

ванный коэффициент не зависит от единиц измерения переменных модели. Кроме того, если коэффициенты при двух объясняющих переменных совпадают, стандартизованный коэффициент окажется больше у регрессора с большей дисперсией.

Интересный факт: стандартизованные коэффициенты почти игнорируются экономистами. О частоте использования каждый может получить представление сам, зайдя, например, на сайт repec.org и запустив поиск словосочетания «standardized coefficients» или «standardized regression coefficients». Современные учебники по эконометрике Грина, Вербика, Баума, Камерона и Триведи либо не уделяют внимания стандартизованным коэффициентам, либо упоминают о них вкратце как о способе получить оценки, не зависящие от единиц измерения. Это замечание не относится к книгам по неэкономическим приложениям статистических методов, а также к учебнику (Johnston, DiNardo, 1997).

Хотя стандартизация полезна при изучении практической значимости регрессоров, потому что позволяет сравнивать их вклад в разброс объясняемого признака, стандартизованные коэффициенты всё же имеют существенные недостатки:

- Их неудобно интерпретировать. Мы можем сказать, что увеличение регрессора x_j на одно стандартное отклонение сопряжено с ожидаемым увеличением объясняемой величины на β_j^* стандартных отклонений при прочих равных условиях. Это толкование вряд ли удовлетворительно, потому что стандартное отклонение — неудобная для интерпретации характеристика.
- Нормировка на σ_i имеет нежелательный эффект: большой стандартизованный коэффициент может наблюдаться в случае, когда объясняемая переменная почти не варьируется при изменении регрессора. С практической точки зрения нам скорее важно оценивать изменение отклика в натуральных единицах.

Второй недостаток легко решаем: достаточно перейти к полустандартизованным (semistandardized) коэффициентам $\beta_j^{**} = \beta_j \sigma_{x_j}$, однако проблема неинтерпретируемости остаётся. Так как проблема эта вызвана недостатком стандартного отклонения как меры разброса, разумным решением будет использование другой меры. Отдадим предпочтение мерам, основанным на квантилях — квантильному размаху (разности двух квантилей случайной величины) и квантильному коэффициенту (отношению двух квантилей). Переход к квантильным мерам — не единственный способ превратить полустандартизованный коэффициент в нечто интерпретируемое, но именно этот способ будет удобен нам для графического представления.

Измерение вклада объясняющей переменной в случае линейной зависимости

Рассмотрим линейное уравнение $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon$. Обозначим за $\mathcal{Q}_j(p)$ функцию квантилей[3] регрессора x_j. Квантильный размах порядка γ равен $\mathcal{Q}_j\left(\frac{1+\gamma}{2}\right) - \mathcal{Q}_j\left(\frac{1-\gamma}{2}\right)$ — это длина отрезка, включающего долю γ всех наблюдений за величиной x_j. Домножив эту величину на коэффициент β_j, получаем удобную характеристику важности объясняющей переменной:

$$CS_j = \beta_j \left(\mathcal{Q}_j\left(\frac{1+\gamma}{2}\right) - \mathcal{Q}_j\left(\frac{1-\gamma}{2}\right) \right)$$

Буквы CS — аббревиатура для contribution spread (размах вклада). Интерпретация: если бы наблюдения в нашей выборке отличались только значениями регрессора x_j, а остальные объясняющие переменные и случайная ошибка не менялись бы от наблюдения к наблюдению, то, согласно нашей модели, квантильный размах значений объясняемой величины составил бы CS_j единиц. Иначе говоря, в средних $\gamma \times 100\%$ наблюдений наибольшее различие между величиной отклика составило бы CS_j.

Наиболее ясной такая интерпретация выглядит при использовании полного размаха — разности между наибольшим и наименьшим значением признака, но 100% размах чувствителен к выбросам. Кроме того, использование 100% размаха создаёт проблемы при распространении результатов на генеральную совокупность: наиболее часто применяемые вероятностные распределения имеют бесконечный размах. Поэтому разумнее сосредоточиться на центральных 90% или 95% наблюдений или любой другой доле по желанию исследователя.

Пример 1. Модель участия женщин в рабочей силе

По данным о 50 штатах США[4] оценивалось уравнение:

$$LFP_i = \beta_1 + \beta_2 Income_i + \beta_3 Educ_i + \beta_4 UR_i + \epsilon_i,$$

где LFP_i — уровень участия женщин в рабочей силе в штате i (%),
$Income_i$ — медианный доход домохозяйства (тыс. долл.),
$Educ_i$ — средняя продолжительность обучения среди женщин (годы),
UR_i — уровень безработицы среди женщин (%).

Таблица ниже содержит оценки коэффициентов и важности вклада каждого из объясняющих признаков:

[3] Точнее, выборочную функцию квантилей. В дальнейшем везде речь идёт именно о выборочных характеристиках.

[4] Взяты из книги (Newbold, 2007)

[5] Размах, конечно, не может быть отрицательным. Знак «минус» добавлен, чтобы отражать направление связи.

признак	*коэффициент*	*станд. коэфф.*	*размах вклада*	*90% размах вклада*
Доход	0.406	0.257	5.090	3.915
Образование	4.842	0.209	3.389	3.123
Безработица	-1.554	-0.510	-10.101[5]	-8.454

Сравнив стандартизированные коэффициенты регрессии, мы придём к выводу, что в рамках нашей модели наибольший вклад в различия между штатами по уровню участия женщин в рабочей силе даёт уровень безработицы. К тому же заключению приведут нас и следующие столбцы таблицы, однако приведённые в них значения позволят описать и величину этого вклада. Так, если бы все штаты различались только уровнем безработицы, то наибольшее различие между штатами в уровне участия женщин в рабочей силе составило бы 10.1%, а при отбрасывании 5% штатов с самим высоким уровнем безработицы и 5% штатов с низкой безработицей этот разрыв сократился бы до 8.5%. Различия в продолжительности получения образования при прочих равных условиях соответствовали бы размаху уровня участия женщин в рабочей силе в 3.4% для всех штатов и 3.1% для «средних» 90% штатов.

Конечно, возникает вопрос, как именно выбирать используемые квантили. Проблема легко решается графически: можно построить график, в котором по горизонтальной оси откладывались бы порядки квантилей p, а по вертикальной оси — величина $\hat{\mathcal{Q}}_j(p) - \hat{\mathcal{Q}}_j(0.5)$. Сдвиг квантильной функции на выборочную медиану $\hat{\mathcal{Q}}_j(0.5)$ удобен при сопоставлении вкладов разных признаков и делает положение графика нечувствительным к выбросам. Можно отразить и направление связи, если для признаков, коэффициент перед которыми отрицателен, откладывать на графике величину $\hat{\mathcal{Q}}_j(0.5) - \hat{\mathcal{Q}}_j(p)$, чтобы соответствующая линия имела отрицательный наклон. Приведём такой график для оценённого уравнения участия в рабочей силе:

Квантильный размах вклада объясняющего признака отражён на этом графике как величина прироста или спада соответствующей линии между нужными квантилями. Такой график позволяет более полно представить себе важность объясняющих переменных. Например, из него видно, что доход и длительность обучения имеют схожий по величине вклад в переменную отклика. Вклад дохода имеет больший размах только за счёт нескольких штатов с высоким уровнем благосостояния — об этом свидетельствует близость линий «educ» и «income» на графике и ускоренный рост линии «income» в правой части (около девятой децили и правее).

Отметим, что линии на графике — результат сдвига и пропорционального растяжения квантильных функций объясняющих признаков, потому они несут в себе всю информацию о частном распределении регрессоров, так как частное распределение однозначно задаётся функцией квантилей. Поэтому графики такого рода дают наглядную поддержку не только регрессионным оценкам,

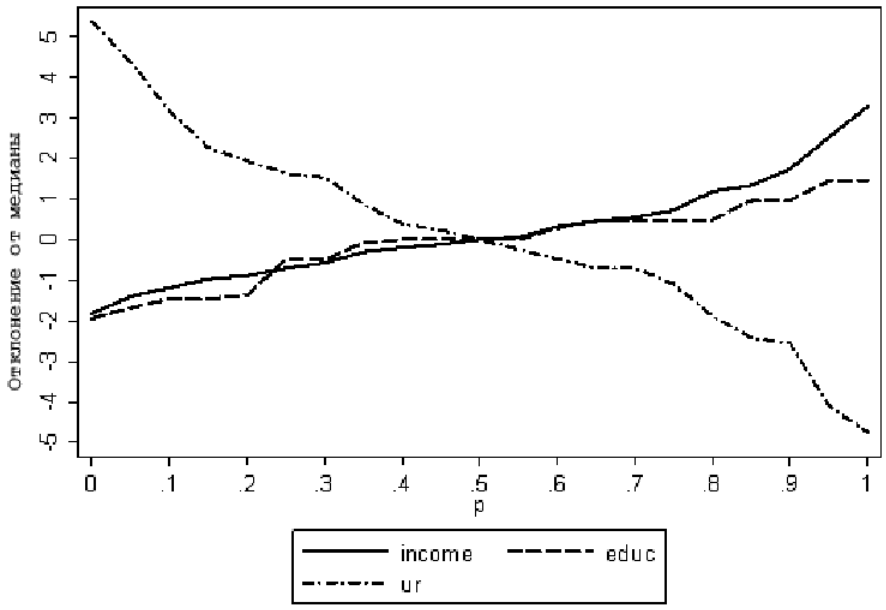


Рис. 2. Название рисунка

но и описательной статистике.

Общий случай. Рассмотренный подход к измерению и наглядному представлению вклада может быть применён и для нелинейных зависимостей. Пусть объясняемая переменная связана с набором регрессоров и случайной составляющей следующим образом:

$$y = f(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k, \epsilon) + g(x_j), \tag{0.1}$$

где $f(), g()$ — функции, на которые мы не накладываем ограничений[6]. Тогда мы можем оценить важность признака x_j по квантильному размаху значений оценённой функции $g(x_j)$ в имеющихся наблюдениях. При этом теряется возможность отразить направление связи — это, однако, не проблема способа измерения важности объясняющего признака, а просто следствие того, что связь может иметь разные направления на разных участках значений объясняющей переменной. При этом в уравнение регрессии признак может быть включён с помощью нескольких переменных, например, линейным и квадратичным членами, либо набором двоичных величин. Не возникает трудностей и при анализе логарифмических зависимостей вида:

$$\ln y = f(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k, \epsilon) + g(x_j). \tag{0.2}$$

[6] Здесь можно порассуждать на тему «а что если эти функции не могут быть случайными величинами», но ценность таких рассуждений сомнительна. В любом случае, выборочные квантили будут существовать, даже если нет теоретических.

Отличие от предыдущего случая заключается в том, что здесь можно измерить относительный вклад вместо абсолютного — оценить, *во сколько раз* изменяется величина отклика при изменениях признака x_j. Показателем важности признака будет квантильный коэффициент $\mathcal{Q}_G(1+\gamma/2)/\mathcal{Q}_G(1-\gamma/2)$ для величины $G = \exp(g(x_j))$. В силу инвариантности квантилей к монотонным преобразованиям этот коэффициент равен потенцированному квантильному размаху значений функции $g(x_j)$ — вклада признака x_j в логарифм переменной отклика. Рассмотрим этот случай на ещё одном примере из области статистических исследований рынка труда.

Пример 2. Уравнение заработной платы

По индивидуальным данным[7] обследования рабочих в Бельгии (1994 г.) оценивалось уравнение:

$$\ln W_i = \beta_1 + \beta_2 Male_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 E2_i + \beta_6 E3_i + \beta_7 E4_i + \beta_8 E5_i + \epsilon_i,$$

где W_i — заработная плата i-го рабочего в выборке (бельгийские франки),
$Male_i$ — пол (1— мужчина, 0 — женщина),
Exp_i — опыт работы (годы),
$E2_i, \ldots, E5_i$ — дамми-переменные для уровня образования ($E5 = 1$ для наивысшего уровня, базовая категория — самый низкий уровень образования).

Признак	Коэффициент	Оценка	Потенцированная оценка	Вклад, относительный размах[8]	Вклад, квантильный коэффициент $Q(0.95)/Q(0.05)$
Пол	β_2	0.115	1.122	1.122	1.122
Опыт работы	β_3	0.034	1.035	1.821	1.700
	β_4	-0.0005	0.9995		
Образование	β_5	0.141	1.152	1.898	1.898
	β_6	0.308	1.362		
	β_7	0.481	1.618		
	β_8	0.641	1.898		

Отметим, что коэффициенты β_3, β_4 практически не поддаются интерпретации (остальные интерпретируемы в потенцированном виде) и что ни один из стандартизированных или полустандартизованных коэффициентов не имеет

[7] взятым отсюда: http://www.wiley.com/legacy/wileychi/verbeek2ed/datasets.html — здесь выложены файлы с данными для примеров из учебника М. Вербика по эконометрике.

[8] Под относительным размахом здесь понимается отношение наибольшего значения к наименьшему.

смысла. Для двоичных переменных бессмысленно рассматривать изменение на одно стандартное отклонение, для переменных Exp и Exp^2 не может идти речь о «прочих равных условиях» — абсурдно рассматривать изменение одной из них при постоянстве другой. Тем не менее, оценки уравнения регрессии могут быть сопровождены осмысленными мерами вкладов каждого из трёх признаков.

Согласно полученным оценкам, наиболее существенный вклад в уровень зарплаты приходится на уровень образования: различия по этому признаку объясняют расхождение заработной платы в 1.898 раз. Схожую по величине роль играет и общий стаж. Если бы наблюдения в выборке отличались только по числу отработанных лет, то модельные значения заработной платы отличались бы не более чем в 1.821 раз во всей выборке и не более чем в 1.7 раза в средних 90% наблюдений.

Наглядное изображение важности вклада признаков даёт график отношения квантилей вкладов к медианному вкладу, на котором по горизонтальной оси откладывается порядок квантили p, а по вертикальной — величина $\mathcal{Q}_G(p)/\mathcal{Q}_G(0.5)$:

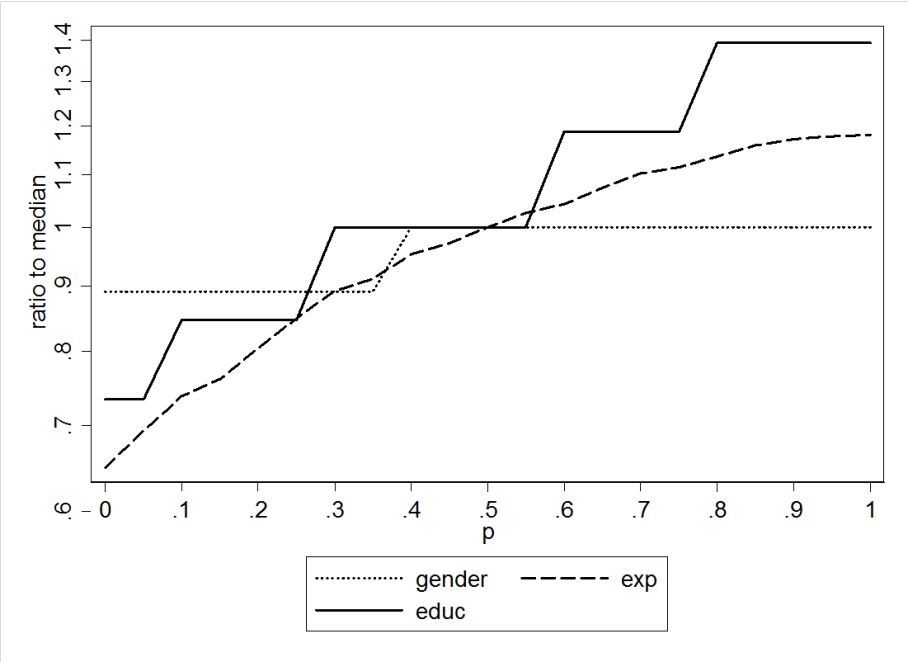


Рис. 3. Название рисунка

Так как по вертикальной оси откладывается отношение, график предпочтительно изображать в логарифмической шкале. Отражая схожий во величине вклад образования и опыта работы, график подчёркивает асимметричность: при прочих равных условиях у людей без опыта работы заработная плата

отклоняется от медианной сильнее, чем у наиболее опытных.

Критика, ответ и опять критика. В отношении предложенного способа измерения практической важности вклада можно выразить ту же критику, что неоднократно высказывалась против стандартизованных коэффициентов: оговорка «при прочих равных» выглядит слишком отдаляющей от действительности. Для экономических приложений естественна связь между объясняющими переменными: изменению одной из них должны сопутствовать изменения других. Само по себе такое замечание неоспоримо, однако стоит иметь в виду, что цель множественного регрессионного анализа — именно выделение связей, изолированных от постороннего влияния. То есть, критика относится скорее к самому подходу, при котором основой анализа становится одна модель множественной регрессии без попыток изучения опосредованных связей (mediation). При желании «освободить» какой-либо из регрессоров можно просто исключить его из модели и сравнить результаты, полученные до и после исключения. Это только увеличит объём возможно полезных сведений в копилке исследователя. Правда, чем больше этих сведений, тем сложнее их свести к однозначному выводу.

Возможные расширения применимости. Формулы (??) и (??) не стоит рассматривать как строгие границы, вне которых для измерения важности объясняющих переменных придётся искать существенно иные подходы. Заметим, что в качестве объясняемой величины y не обязательно должен фигурировать именно статистический признак — это может быть какая-либо интерпретируемая характеристика распределения этого признака, например шансы (odds) какого-либо события при моделировании бинарного выбора или функция риска (hazard function) при моделировании времени жизни.

Дополнительные возможности для иллюстрации предлагаемого подхода даёт использование панельных данных: разбиение вклада признаков на межгрупповой (between) и внутригрупповой (within) разброс, сравнение вклада наблюдаемой разнородности (отдельных переменных или всей совокупности регрессоров) с ненаблюдаемой (индивидуальным эффектом).

Пара абзацев в конце. Ещё раз отмечу: в статье речь идёт о таком аспекте практической значимости, как вклад разброса объясняющего признака в разброс регрессанта при прочих равных условиях («dispersion importance», говоря словами К.Эйкена – см. (Achen, 1982, стр. 73-77)). Исследование практической значимости, конечно, не сводится только к этому аспекту. Это вообще не то, что можно было бы полностью формализовать и свести к математическим мерам. Моя цель скромнее — предложить такой способ численно и графически описать связь между признаками, который был бы полезным подспорьем при выяснении практической значимости.

Квантили — не единственный способ получить интерпретируемые меры разброса. Можно опираться на среднее абсолютное отклонение величины от своего среднего или медианы, хотя я не знаю, как на их основании построить

столь же информативные графики. Предлагаю читателю самому обдумать соответствующие характеристики вклада объясняющих переменных.

Основа этого текста — доклады на семинаре кафедры математической экономики и эконометрики и семинаре департамента прикладной экономики НИУ ВШЭ, сделанные в 2013 году. Автор благодарит Г.Г. Канторовича, Э.Б. Ершова, Б.Б. Демешева и А.А. Пересецкого за обсуждение.

Случайная перестановка (рабочее название)

Борис Демешев*

18 октября 2015 г.

Аннотация

Случайная перестановка

Ключевые слова: задача, случайная перестановка, киллер.

1 Классные свойства случайных перестановок

здесь про E(), Var() и т.д.
https://terrytao.wordpress.com/2011/11/23/the-number-of-cycles-in-a-random-permutation/
Утверждения
Случайная перестановка.
Случайная величина C_k — количество циклов длины k.
Случайная величина C — количество циклов произвольной длины

1. Пусть $A_k(i)$ — элемент i входит в цикл длины заданной k. ¶$(A_k(i) = 1/n$
2. Какова вероятность того, что a, b и c лежат в одном цикле?
   Начинаем идти по циклу от a. Рано или поздно цикл «вонзится» в множество $A = \{a, b, c\}$. Нам надо пройти сквозь b или c, отсюда $2/3$. Допустим, прошли сквозь b. Идём по циклу дальше, он снова рано или позщно вонзится в A. Нам надо пройти сквозь c, отсюда $1/2$. При следующем прохождении цикла через множество A мы обязательно попадаем в a.

$$\frac{2}{3}\frac{1}{2} = \frac{1}{3}$$

3. Какова вероятность того, что a, b и c лежат в одном цикле длины $m \geq 3$?
4. Назовём цикл «длинным», если его длина больше $n/2$. Какова вероятность того, что существует длинный цикл длины m?
   Доказательство
5. Какова вероятность того, что существует хотя бы один длинный цикл?
   Ответ: $\approx \ln 2 \approx 0.69$
6. Какова ожидаемая длина цикла, в котором лежит элемент i?
   $(N + 1)/2$

*НИУ ВШЭ, Москва.

7. $\mathbb{E}(C_k) = 1/k$
   Доказательство:
   Рассмотрим случайную величину kC_k — это количество элементов, входящих в циклы длины k. Разложим это количество в сумму индикаторов, $kC_k = X_1 + X_2 + \ldots + X_n$. Здесь X_i — входит ли элемент i в цикл длины k. Следовательно, $\mathbb{E}(kC_k) = ¶(X_1 = 1) + \ldots ¶(X_n = 1) = n \cdot \frac{1}{n} = 1$.
   http://math.stackexchange.com/questions/306977/cycles-permutation-random-probability
   — доказательство через производящие функции
8. $\mathbb{E}(C) = 1 + \frac{1}{2} + \ldots + \frac{1}{n}$
   Доказательство $\mathbb{E}(C) = \mathbb{E}(C_1) + \mathbb{E}(C_2) + \ldots + \mathbb{E}(C_n)$
9. $\mathbb{E}(C^j_{C_k}) = 1/k^j j!$
10. обобщение предыдущей
11. Асимтотически C_k имеет пуассоновское распределение с $\lambda = 1/k$
12. Асимптотически количества циклов разных длин независимы
    http://www.ams.org/mathscinet-getitem?mr=1175278
13. $\mathbb{E}(m^C) = C^{n+m-1}_n$

https://en.wikipedia.org/wiki/Random_permutation_statistics
http://www.inference.phy.cam.ac.uk/itila/cycles.pdf

2 Задачки

2.1 Сумасшедшая старушка

В самолете 100 мест и все билеты проданы. Первой в очереди на посадку стоит Сумасшедшая Старушка. Сумасшедшая Старушка очень переживает, что ей не хватит места, врывается в самолёт и несмотря на номер по билету садиться на случайно выбираемое место. Каждый оставшийся пассажир садится на своё место, если оно свободно, и на случайное выбираемое место, если его место уже кем-то занято.

1. Какова вероятность того, что последний пассажир сядет на своё место?
2. Чему примерно равно среднее количество пассажиров севших на свои места?

2.2 Судьба Дон-Жуана

У Дон-Жуана n знакомых девушек, и их всех зовут по-разному. Он пишет им n писем, но по рассеянности раскладывает их в конверты наугад. Случайная величина X обозначает количество девушек, получивших письма, адресованные лично им.

1. Найдите $\mathbb{E}(X)$, $\mathrm{Var}(X)$
2. Какова при большом n вероятность того, что хотя бы одна девушка получит письмо, адресованное ей?

2.3 Киллер

Правила игры «Киллер» просты. Игроки пишут на бумажках, как их зовут, и кладут бумажки в шляпу. Каждый тянет из шляпы имя своей первой жертвы. Если первой жертвой игрока является он сам, то он совершает «самоубийство» и дальше

не играет[1]. Чтобы убить жертву, надо остаться с ней наедине и сказать: «Ты убит!». Убийца забирает себе все бумажки, набранные убитым, и начинается охотиться за тем, за кем охотился убитый. Побеждает тот, кто наберёт больше всех бумажек к концу игры. Заметим, что в «Киллере» каждый игрок оказывается втянут в одну из нескольких цепочек.

В «Киллера» играют 30 человек, из них 20 девушек.

1. Какова вероятность того, что в цепочке, начинающейся с Маши Сидоровой ровно 5 человек?
2. Какова вероятность того, что в цепочке, начинающейся с Маши Сидоровой ровно 5 девушек?
3. Какова вероятность того, что все девушки попадают в одну цепочку убийц и жертв?
4. Какова вероятность того, что все игроки попадают в общую цепочку?
5. Сколько в среднем цепочек в «Киллере»?
6. Сколько в среднем «самоубийц»?

2.4 Ключи и копилки

На столе стоят n свиней-копилок. Достать содержимое копилки можно двумя способами: либо разбить копилку, либо открыть дно специальным ключиком. К каждой копилке подходит единственный ключ. Мы раскладываем ключи по копилкам наугад, один ключ в одну копилку. Затем разбиваем k копилок и получаем хранящиеся в них ключи. Далее мы будем копилки только открывать ключами.

1. Какова вероятность того, что мы сможем достать все ключи?
2. Какая доля ключей в среднем будет найдена?

2.5 Задача о макаронах

В тарелке запутавшись лежат $n >> 0$ макаронин. Я по очереди связываю попарно все торчащие концы макаронин.

1. Какова примерно вероятность того, что я свяжу все макаронины в одно большое кольцо?
2. Сколько в среднем колец образуется?
3. Каково среднее число колец длиной в одну макаронину?

2.6 Задача о 100 заключенных

У ста узников тюрьмы есть последний шанс на спасение. В комнате стоит шкаф в котором сто занумерованных ящичков. Палач кладёт в каждый ящичек бумажку с номером ровно одного из заключенных в случайном порядке и задвигает все ящички. Узники заходят в комнату один за другим. Каждый узник может открыть любые 50 ящичков. После каждого узника все ящички задвигаются в исходное положение. Если каждый узник находит свой номер, то все узники будут помилованы. Если хотя бы один из узников не найдёт свой номер, то все будут казнены.

Узники могут предварительно договориться о стратегии.

[1] В некоторых вариантах правил, если игрок вытянул из шляпы своё имя, то он должен вытянуть другую бумажку.

1. Какова оптимальная стратегия?
2. Какую вероятность выигрыша она обеспечивает?

2.7 Три игрока и три вопроса

2.8 что-то про детерминант?