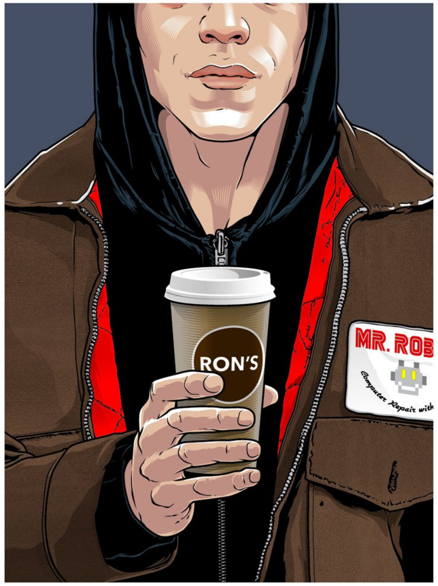
Домашняя работа №2

Илья Дуров

1 марта 2017 г.


RON'S
MR. ROB

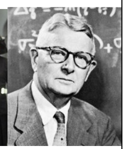

BURGER


Bantha Milk
DAILY FRESH
(1.89L)


DHARMA
DI 9FFTR731
DHARMA INITIATIVE SALAD DRESSING
RANCH COMPO


Foree Elec
SALES ADVISER
SHAUN

Методы наименьших квадратов

Метод	Оценка	Год	Автор	Фото автора	Описание
Метод наименьших квадратов (OLS)	$(X^T X)^{-1} X^T y$	1795	Carl Friedrich Gauss; Adrien-Marie Legendre	 	Метод оценивания параметров эконометрической модели, состоящий в минимизации суммы квадратов расхождений между наблюдаемыми значениями зависимой переменной и значениями этой переменной, вычисленными для наблюдаемых значений независимых переменных по оценённой модели связи.
Обобщённый метод наименьших квадратов (GLS)	$(X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$	1934	Alexander Aitken		Теоретическая процедура оценивания коэффициентов линенйиной модели регрессии в ситуации, когда случайные ошибки имеют разные дисперсии и коррелированы между собой, при этом предполагается, что ковариационная матрица вектора ошибок невырождена и все ее элементы известны.
Взвешенный метод наименьших квадратов (WLS)	$(X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$ при этом $\Omega = \mathrm{diag}(\sigma_1, \sigma_2, \dots \sigma_n)$	тот же	он же	он же	Процедура, состоящая в минимизации определённым образом взвешенной суммы квадратов отклонений наблюдаемых значений зависиммой переменной от значений, вычисляемых по подбираемой модели связи.
Доступный обобщённый метод наименьших квадратов (FGLS)	$(X^T \hat{\Omega}^{-1} X)^{-1} X^T \hat{\Omega}^{-1} y$	тот же	он же	он же	Практически реализуемая процедура оценивания коэффициентов линейной модели регрессии в ситуации, когда случайные ошибки имеют разные дисперсии и коррелированы между собой, повторяющая процедуру обобщенного метода наиеньших квадратов, но импользующая оцененную ковариационную матрицу вектора ошибок.
Косвенный метод наименьших квадратов (ILS)	Те же, что и в двухшаговом МНК, если все уравнения системы являются точно идентифицированными	1928	Philip Wright; Sewall Wright (отец и сын)	 	Метод получения оценок параметров i–го стохастического уравнения структурной формы через оценки наименьших квадратов коэффициентов уравнений приведенной формы. Метод применим в случае точной идентифицируемости i–го структурного уравнения.
Двухшаговый метод наименьших квадратов (2SLS)	$(X^T Z(Z^T Z)^{-1} Z^T X)^{-1} X^T Z(Z^T Z)^{-1} Z^T y$	1953 1957	Henri Theil; Robert Basmann	 	Метод оценивания коэффициентов уравнения структурной формы, состоящий в предварительной очистке стохастической объясняющей переменой от коррелированности с ошибкой в этом уравнении с использованием инструментальных переменных и в последующем оценивании уравнения, в котором исходная объясняющая переменная заменяется ее очищенным вариантом.
Трёхшаговый метод наименьших квадратов (3SLS)	$(\hat{Z}^T (\hat{\Lambda}^{-1} \otimes I_g) \hat{Z})^{-1} \hat{Z}^T (\hat{\Lambda}^{-1} \otimes I_g) y$	они же	они же	они же	Доступный обобщённый метод наименьших квадратов, применённый к системе одновременных уравнений. Принимает во внимание наличие коррелирванности между ошибками в разных структурных уравнениях.

Еще раз дадите подгонять размеры картинок на глазок, скажу все второкурсникам, что нельзя идти на этот факультатив

перечитавший кучу мануалов невыспавшийся человек

1. Бесконечные множества

Возникает естественный вопрос, все ли бесконечные множества равномощны?

1.1. Первая занимательная теорема

Пусть S множество всех бесконечных вправо последовательностей из 0 и 1. Например, одним из элементов S является последовательность 1010101010…

Множество S бесконечно, но не равномощно множеству N.

Допустим противоположное, что S и N равномощны. Тогда существует взаимно-однозначное соответствие между натуральными числами и последовательностями. К примеру оно могло бы выглядеть так:

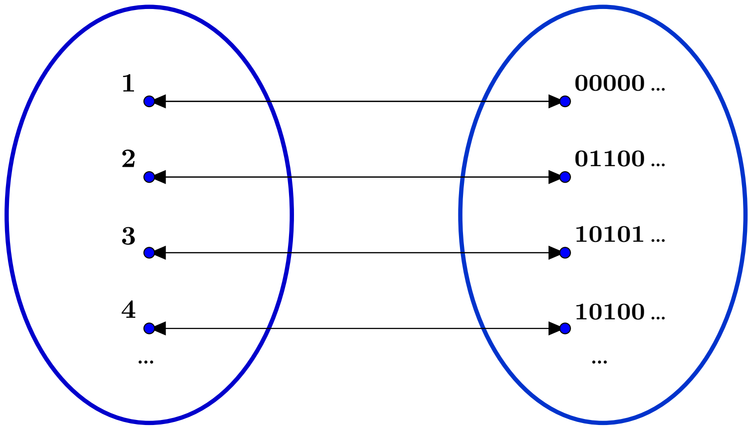


Рис. 1: Попытка построить соответствие

Оказывается какое бы соответствие ни было создано, всегда существует последовательность, которой не сопоставлено ни одно число!

Создадим последовательность a по следующему принципу: возьмем первую цифру из первой последовательности, затем вторую из второй, затем третью из третьей и т.д.

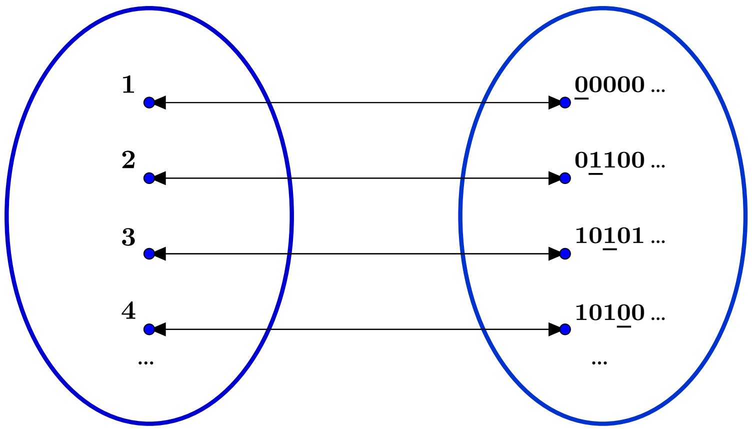


Рис. 2: Нарушение соответствия

Получаем последовательность $a = 0110\ldots$ Затем построим последовательность b заменив единицы на нули, а нули на единицы в последовательности a. В нашем примере $b = 1001\ldots$

Вне зависимости от того, какое соответствие мы взяли последовательность b не может идти в нём ни под каким номером! Она не может идти под номером 1, так как отличается от первой последовательности первой цифрой. Она не может идти под номером 2, так как отличается от второй последовательности второй цифрой и т.д.

Мы пришли к противоречию, в S есть «лишняя» незанумерованная последовательность b. Значит S и N неравномощны.

Аналогичным способом можно построить «лишнюю» последовательность для любого другого предложенного соответствия.

1.2. Разные бесконечности

Мы говорим, что множество A имеет мощность **континуум**, если оно равномощно множеству S бесконечных вправо последовательностей из 0 и 1.

Множество A называется **счётным**, если оно конечно или равномощно множеству N натуральных чисел.

Будьте бдительны при чтении других источников: некоторые авторы определяют счётные как равномощные натуральным числам, но таких авторов меньшинство.

2. Отрезок

$$[0;1]$$

Отрезок $[0;1]$ — множество мощности континуум! Покажем, что множество $[0;1]$ равномощно множеству S бесконечных вправо последовательностей из 0 и 1.

Любое число $x \in [0;1]$ можно записать в виде бесконечной двоичной дроби. Первый знак этой дроби равен 1 или 0 в зависимости от того, попадает ли число x в левую или правую половину отрезка. Чтобы выбрать следующий знак, надо снова поделить выбранную половину пополам и посмотреть, куда попадет x, и т.д.

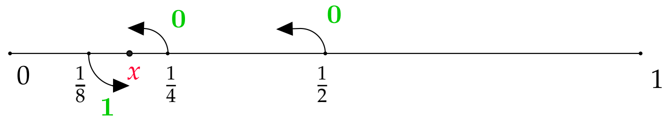


Рис. 3: Взаимно-однозначное соответствие множеств $[0;1]$ и S

Точка $1/2$ на рисунке 2 является серединой отрезка $[0;1]$. Точка x лежит левее $1/2$, то есть попадает в левую половину отрезка. Первый знак в элементе из S, который соответствует x будет 0. Точка $1/4$ является серединой левой половины отрезка $[0;1]$. Точка x снова находится левее $1/4$, значит второй знак в элементе из S также будет 0. Точка $1/8$ является серединой отрезка $[0;1/4]$. Точка x находится правее $1/8$, значит третий знак в элементе из S будет 1. Далее посмотрим в какой части отрезка $[1/8;1/4]$ будет лежать точка x и получим четвертый знак элемента, затем пятый и так далее.

Это же соответствие можно описать в другую сторону: последовательности из нулей и единиц $x_0x_1x_2\ldots$ соответствует число, являющееся суммой ряда

$$\frac{x_0}{2} + \frac{x_1}{2^2} + \frac{x_2}{2^3} + \ldots$$

Например, последовательности $010100\ldots$ соответствует число

$$\frac{0}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{1}{2^4} + \frac{0}{2^5} + \frac{0}{2^6} + \ldots = \frac{1}{4} + \frac{1}{16} = \frac{5}{16}.$$

Описанное соответствие пока что не совсем взаимно-однозначное: дроби вида $m/2^n$ имеют два представления. Например, число $3/8$ можно записать в виде $0011000\ldots$ и в виде $0010111\ldots$ Соответствие станет однозначным, если отбросить последовательности с бесконечным хвостом из единиц, кроме последовательности $01111\ldots$ Таких дробей счётное число и на мощность это никак не повлияет.

3. Очень тонкие вопросы

3.1. Первый вопрос

Какому числу соответствует последовательность $011111\ldots$?

Последовательность $011111\ldots$ соответствует единице.

$$\frac{0}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} + \ldots = \frac{1/2}{1-1/2} = 1.$$

Именно поэтому последовательность $011111\ldots$ нельзя отбросить.

3.2. Второй вопрос

Обычно в памяти компьютера числа хранятся в двоичной системе счисления.

- *Можно ли данным способом хранить в памяти число 0.15?*
- *А правда ли что с точки зрения компьютера $0.4 - 0.3$ равно 0.1?*

Если попытаться записать 0.15 в виде последовательности из 0 и 1, мы получим периодическую бесконечную последовательность $0,1010101\ldots$, которую компьютер не сможет запомнить, так как объем памяти в нём ограничен.

Если выполнить сравнение $0.4 - 0.3 == 0.1$ в большинстве языков программирования (R, Python, Julia, C++, ...) то результатом будет FALSE.

Не зная этого факта, можно получить довольно много проблем.

3.3. Проблема номер один. Мультиколиниарность.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Как всем известно, МНК-оценку в модели множественной регрессии можно получить по формуле (3.3).
При этом делается предположение, что определитель матрицы $X^T X$ не равен нулю.
Ситуацию когда $\det(X^T X) = 0$ называют мультиколинеарностью. В случае её возникновения матрица X содержит линейно-зависимые столбцы и МНК-оценки не существует.
Именно на этом факте строится знаменитая дамми-ловушка, в которую попадают некоторые студенты.

$$\begin{pmatrix} 1 & x_{21} & \dots & x_{k1} & 0 & 1 \\ 1 & x_{22} & \dots & x_{k2} & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & \dots & x_{kn} & 0 & 1 \end{pmatrix} \tag{1}$$

Например, в случае (1) неправильно введены дамми-переменные. Последние два столбца в сумме дают первый стобец. Это приводит к тому, что МНК-оценки не существует.
Однако, в силу причин, которые были перечислены выше, R (или любая другая программа) может оценить модель за счёт возникновения машинных бесконечно малых. При этом коэффициенты, скорее всего, получатся очень большими по модулю.
Опасайтесь мультиколинеарности и не попадайте в дамми-ловушки!

3.4. Проблема номер два. Несимметричная матрица.

Более того, матрица $X^T X$ вследствие машинных малых может получиться несимметричной. Это приводит к небольшому сдвигу вашей регрессии. Однако если к этой проблеме добавить немножечко зависимости регрессоров друг от друга, то сдвиг станет более ощутимым.

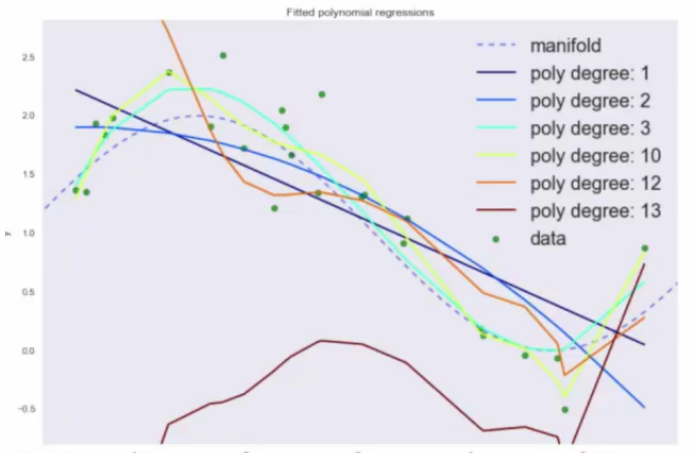


Рис. 4: Съехавший полином

Например, если вы любитель полиномов, то с вами может произойти следующая история. На рисунке 3.4 изображено несколько разных моделей. Каждая новая линия отвечает за новую модель с большим количеством степеней в правой части. При этом на 13 степени выскакивает описанная выше неадекватность. Линия, включающая в себя 13 степеней съехала куда-то вправо очень причудливым образом. Оценки коэффициентов в модели оказались искажены.

3.5. Мораль

Не забывайте думать!