Статистика 2

Семерикова Елена Вячеславовна, к.э.н.

- **1. Корреляция**
- **2. Линейная регрессия**
- 2.1. Зачем: интерпретация и предсказание
- 2.2. Качество регрессии (R^2)
- 2.3. Связь: правда или ложь?

Дисперсия

- (лат. Рассеяние)
- В физике – разложение белого на различные цвета

- Variance

- Вариация

- КОвариация?

Коэффициенты корреляции

Качественные (порядковые) переменные:

- **Порядковые**
  - ранговый коэффициент корреляции Спирмена

Количественные переменные

- Коэффициент корреляции Пирсона

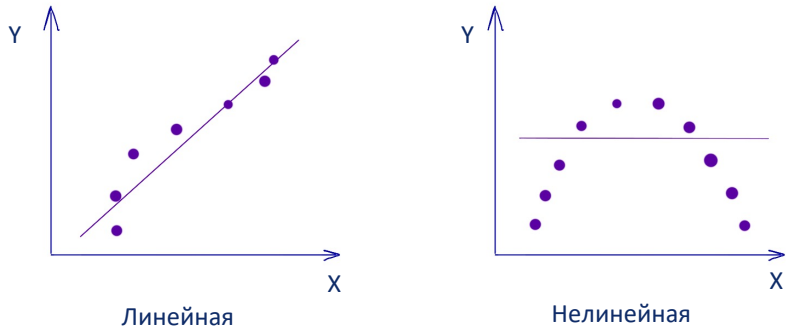
Коэффициент корреляции (Пирсона)

- Часто переменные бывают связаны друг с другом. Что их связывает? Как измерить данную связь?
- **Коэффициент корреляции** – мера взаимосвязи между количественными переменными X и Y
- Данный коэффициент обозначается как $\mathbf{r_{XY}}$ или **r** и получается при нормализации ковариации
- Есть два вида взаимосвязи переменных:
- **Положительная**: при возрастании X растет Y
- **Отрицательная**: при возрастании X уменьшается Y
- Доступный **диапазон значений: [-1; 1]**
- r = 1 идеальная положительная корреляция (точки лежат на прямой с положительным наклоном)
- 0 < r ≤ 1 положительная корреляция (точки лежат вокруг прямой с положительным наклоном)
- r = 0 отсутствие корреляции
- -1 ≤r < 0 отрицательная корреляция (точки лежат вокруг прямой с отрицательным наклоном)
- r = -1 идеальная отрицательная корреляция (точки лежат на прямой с отрицательным наклоном)
- Вычисляется по формуле: $r_{XY} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}}$

$$r_{XY} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} * \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Коэффициент корреляции

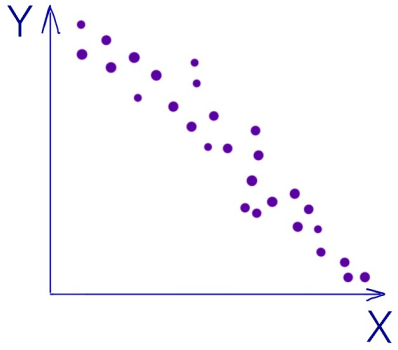
- Коэффициент корреляции является мерой **линейной зависимости**
- Линейная = прямая линия
- **Диаграмма рассеяния** может быть использована для проверки того, является ли зависимость (примерно) линейной. Если она не является таковой, коэффициент корреляции использовать нельзя



- Следует остерегаться **выбросов** – точек, которые находятся далеко от других точек, они могут сильно влиять на r_{XY}

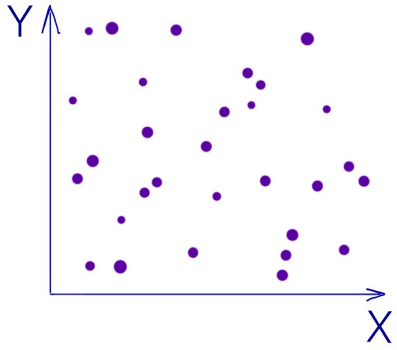
Диаграммы рассеяния

Коэффициент корреляции отрицательный



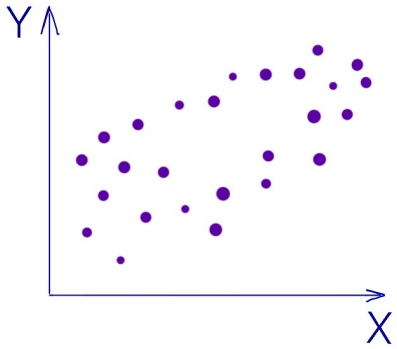
r ≈ -1

Корреляция отсутствует

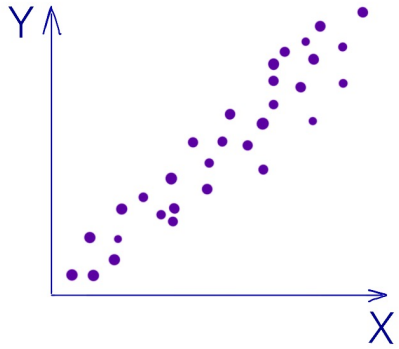


r = 0

Коэффициент корреляции положительный – связь есть

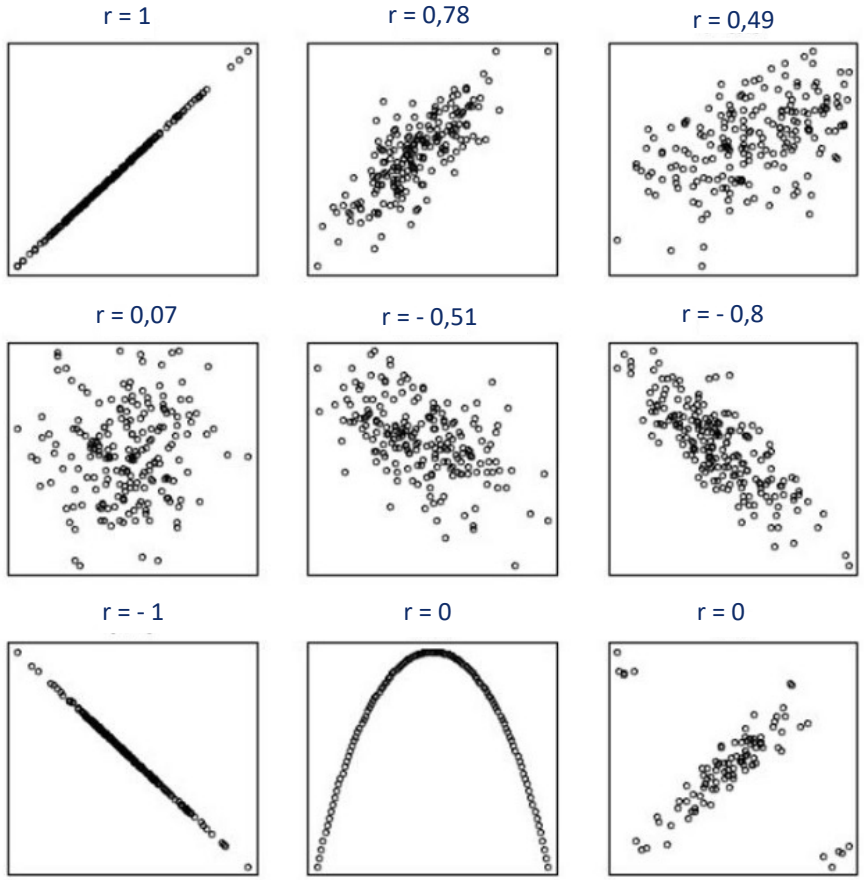


r ≈ 0,5



r ≈ 1

Примеры диаграмм рассеяния для различных коэффициентов корреляции

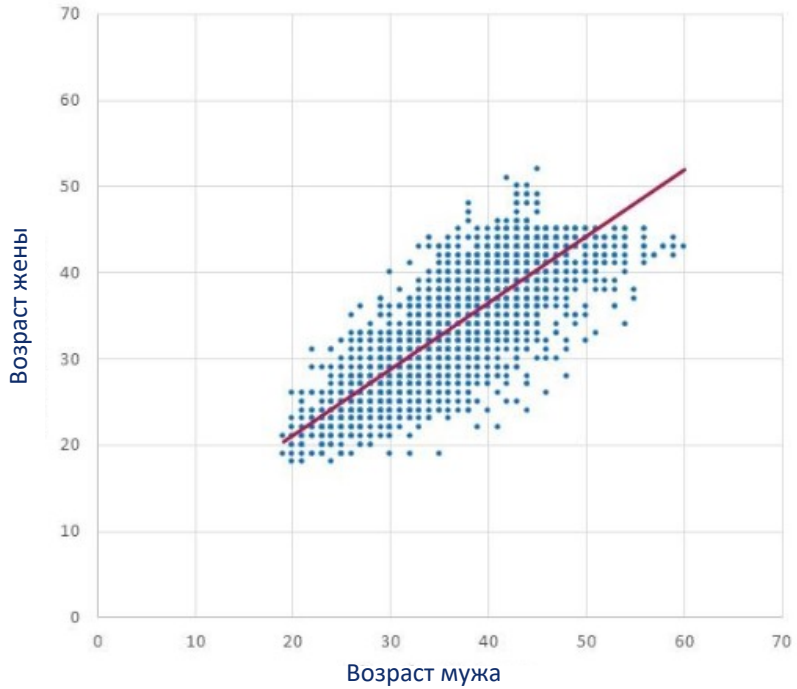


Пример: билеты в кино и фаст-фуд

- Нам известны цены обеда в фаст-фуд кафе и 2 билетов в кино в различных городах мира. Какова корреляция между ними?
- $r_{XY} = \frac{\sum(x-\bar{x})(y-\bar{y})/(n-1)}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2/(n-1)}} = \frac{cov(X,Y)}{S_X S_Y} = \frac{6,976}{(1,292)(6,337)} \approx 0,835$
- Цены на фаст-фуд обед и билеты в кино положительно коррелируют. Это значит, что в городах с наиболее дорогими фаст-фуд обедами обычно самые дорогие билеты в кино, и наоборот. Однако, из этого не следует, что дорогой фаст-фуд является причиной или следствием из дорогих билетов в кино.

	A	B	C	D	E	F	G
1	**Город**	**Цена обеда**	**Билеты в ки**	**(X-Xcp)^2**	**(Y-Ycp)^2**	**(X-Xcp)(Y-Ycp)**	
2	Лос-Анджел	5,99	32,66	1,022121	157,176369	12,674907	
3	Цюрих	7,62	28,41	6,974881	68,674369	21,885967	
4	Нью-Йорк	5,75	20	0,594441	0,015129	0,094833	
5	Мадрид	4,45	20,71	0,279841	0,344569	0,310523	
6	Токио	4,99	18	0,000121	4,507129	0,023353	
7	Париж	5,29	19,5	0,096721	0,388129	0,193753	
8	Берлин	4,39	18	0,346921	4,507129	1,250447	
9	Москва	3,7	16	1,635841	16,999129	5,273317	
10	Рим	4,62	18,05	0,128881	4,297329	0,744207	
11	Варшава	2,99	9,9	3,956121	104,509729	20,333547	
12			**Сумма**	15,03589	361,41901	62,784854	
13							
14							
15				Вычисления			
16				Xcp	4,979	=СРЗНАЧ(B2:B11)	
17				Ycp	20,123	=СРЗНАЧ(C2:C11)	
18				n-1	9	=СЧЁТ(B2:B11)-1	
19				Ковариация	6,97609489	=F12/E18	
20				Sx	1,29253799	=КОРЕНЬ(D12/E18)	
21				Sy	6,33700779	=КОРЕНЬ(E12/E18)	
22				r	0,83480863	=КОРРЕЛ(B2:B11;C2:C11)	
23							

Линейная регрессия



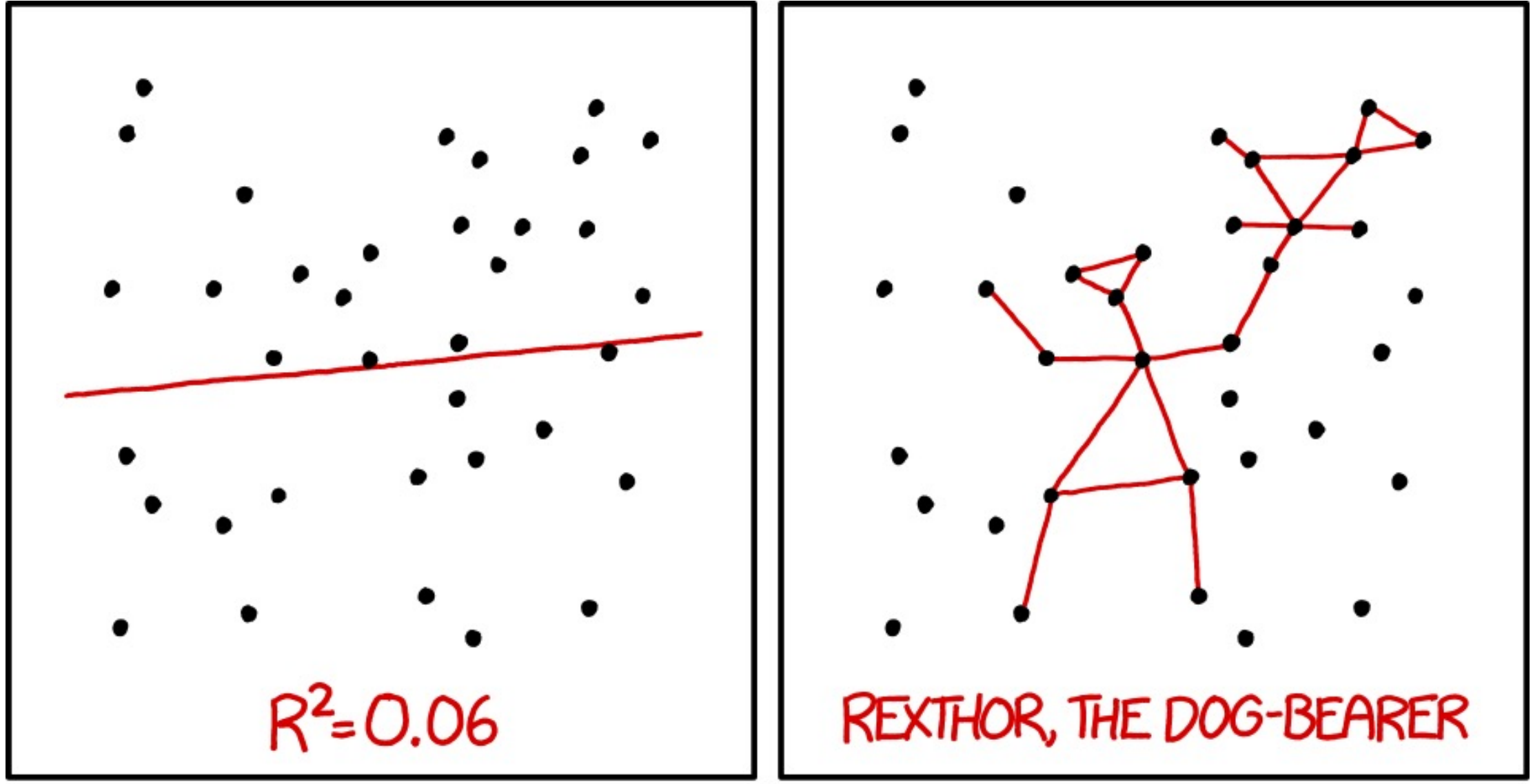
- Возраст супругов
- По горизонтали - возраст мужа, по вертикали – жены.
- Прямая линия может быть описана линейной функцией:
  - $\boldsymbol{Y = \alpha + \beta X}$
  - где α – точка пересечения с осью Y, β – наклон по оси Y при увеличении X на одну единицу
    - В нашем случае зависимость имеет вид: $Y = \alpha + 0{,}8X$
- В данном случае мы можем сделать вывод, что при возрасте мужа (X) выше на 1 год, возраст жены (Y) будет выше на 0,8 лет.
- **β характеризует, насколько больше будет значение Y при X выше на 1 единицу**

R-квадрат

- Можно заметить, что в первом примере облако точек лежало ближе к линии регрессии, чем во втором. *О чем это говорит?*
- Так происходит из-за того, что в первом случае взаимосвязь выше, чем в другом
- Данный факт можно проследить при помощи **R-квадрата**. Данная мера показывает, какую **долю дисперсии переменной Y объясняет переменная X**
- R-квадрат принимает значения в **диапазоне [0;1]**
- 0 – отсутствие взаимосвязи
- 1 – тесная взаимосвязь
- *Очень уж похоже на коэффициент корреляции…*

R-квадрат – та же корреляция?

- Возьмем для примера модель зависимости количества отдыхающих на пляже от количества солнечных дней.
- R^2 в ней равняется 0,865
- Коэффициент корреляции между X (количество солнечных дней лета) и Y (число отдыхающих) r равен 0,93
- $R^2 = 0,93*0,93 = 0,865$
- Вывод: **простая линейная регрессия связана с коэффициентом корреляции**. Оба коэффициента показывают взаимосвязь между количественными переменными (**r_{XY} в квадрате = R^2**)



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.


РАНЬШЕ Я ДУМАЛ, ЧТО КОРРЕЛЯЦИЯ ОБУСЛОВЛЕНА ПРИЧИННОСТЬЮ
ПОТОМ Я ПОСЕТИЛ УРОК СТАТИСТИКИ И ПОНЯЛ, ЧТО ЭТО НЕ ТАК
ПОХОЖЕ, ЧТО УРОК СТАТИСТИКИ ПОМОГ
ХМ, ВОЗМОЖНО

Корреляция и причинность

- Если понаблюдать за маленькими детьми, то можно заметить, что навыки чтения и размер обуви сильно коррелируют
- *Значит ли это, что увеличение размера ноги влечет за собой умение читать?*
- *Правда ли, что навыки чтения зависят от размера обуви?*
  - **Конечно же, это не так!**
  - Есть общая **связующая переменная** – возраст
- В данном случае идет речь о **мнимой (spurious) корреляции**

Мнимая корреляция

Когда ребенок взрослеет → Он начинает лучше читать

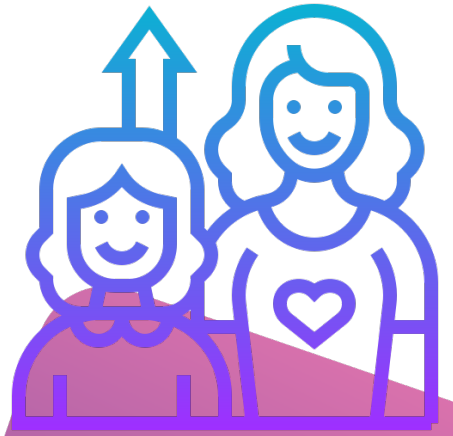
Когда ребенок взрослеет → Его размер обуви увеличивается

Он начинает лучше читать ↕ Его размер обуви увеличивается

- Корреляция r не защищает от наличия общих нарушающих факторов
  - r измеряет силу и направление линейной зависимости X и Y

≠

- причинно-следственная связь

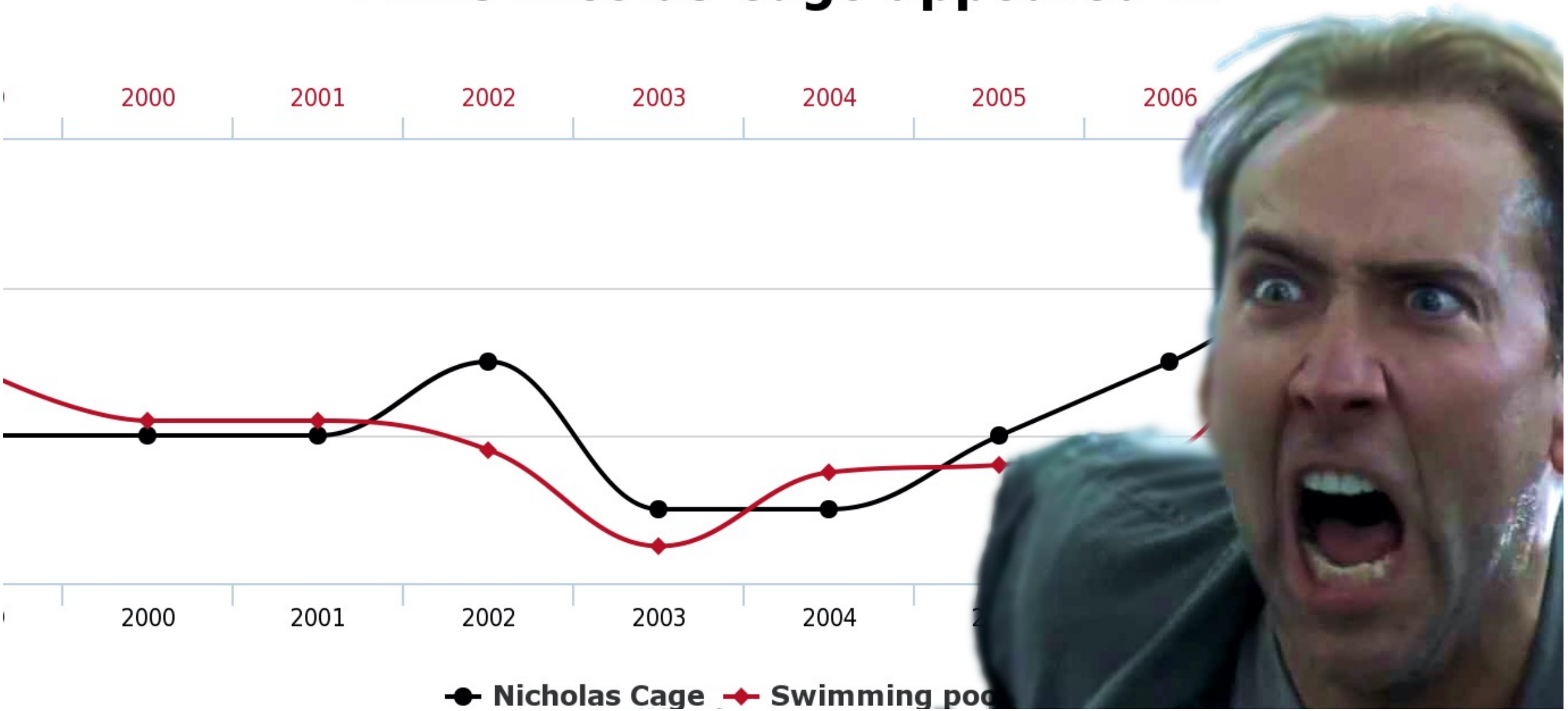


https://rationalnumbers.ru/all/bezumnye-korrelyacii/?ysclid=l6l336pq4q13281671

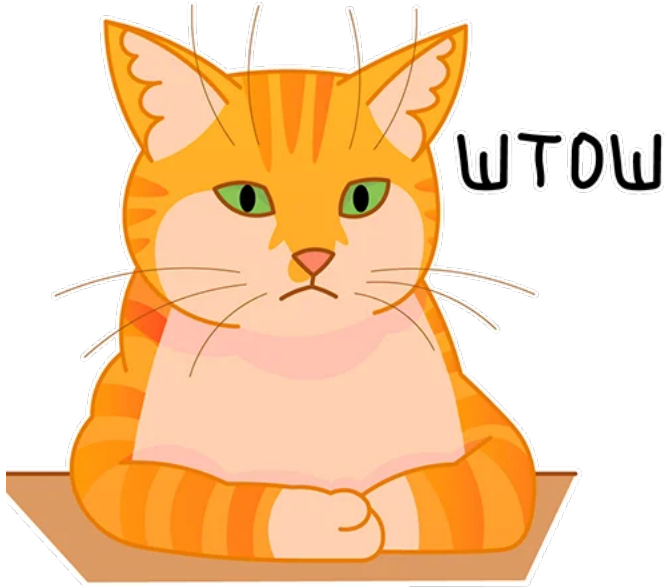
Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



Грустное резюме



Две количественные переменные – корреляция Пирсона
Порядковые переменные – коэффициент ранговой корреляции Спирмена

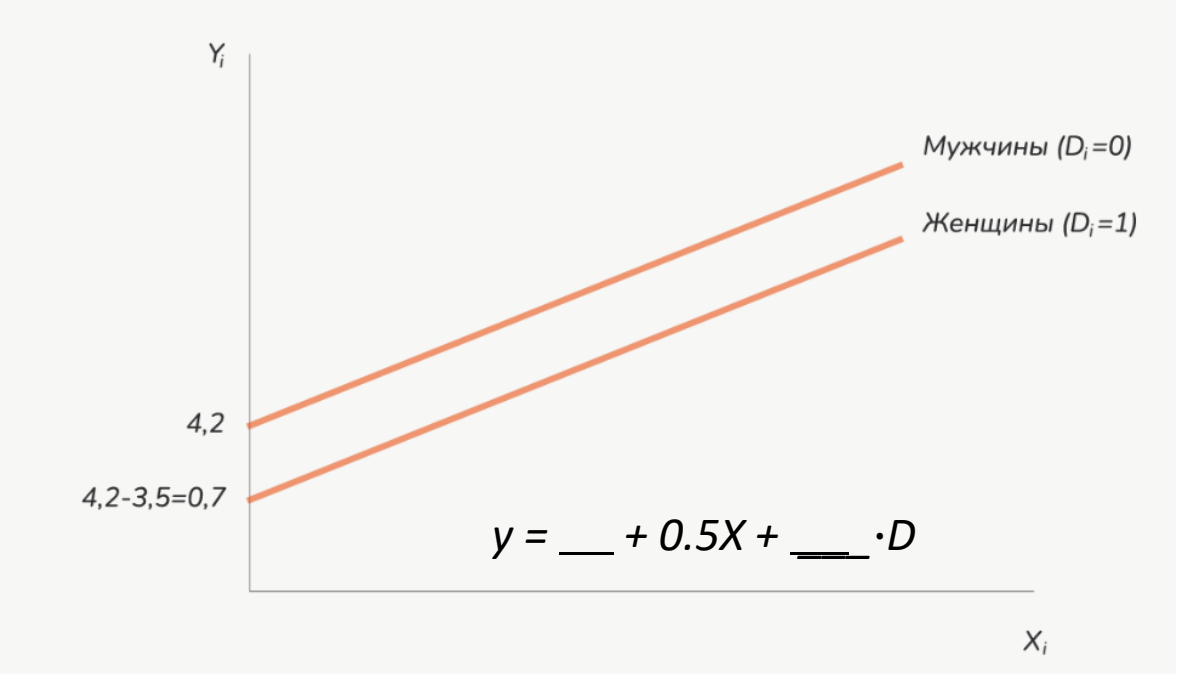
- Корреляция ничего не скажет….
- И линейная регрессия тоже ☹
- Причинно-следственную взаимосвязь еще попробуй докажи…

Многофакторная регрессия

$$y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot X_2 + \ldots . \beta_n X_n$$

- При добавлении новой переменной в регрессию, коэффициент перед «старой» переменной снижается

Бинарные переменные



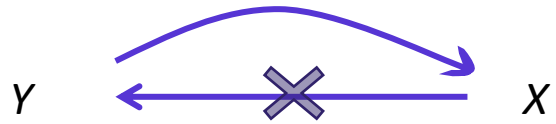
$$y = \beta_1 + \beta_2 X + \beta_3 \cdot D$$

D = 0, если ...

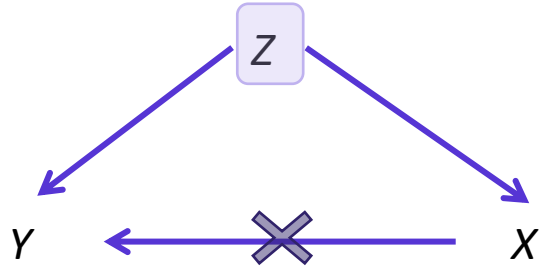
D = 1, если ... (другое)

Ошибки регрессий:

- Смещенная выборка (напр., проблема самоотбора)
- Обратная взаимосвязь

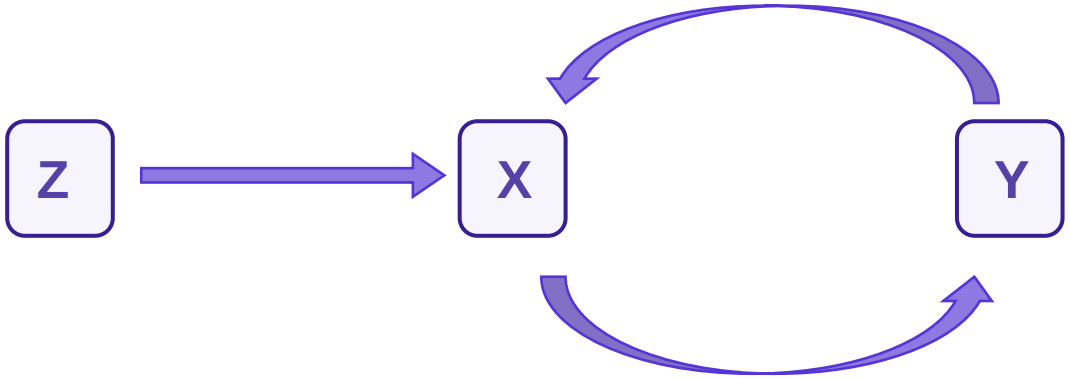


- Неправильная функция (напр., парабола вместо прямой)
- Пропущенная переменная



Инструментальные переменные

- X зависит от Y,
- Y зависит от X



Выход:

- Переменная Z:
  - Влияет на X
  - Не связана (напрямую) с Y

- Можно оценить причинно-следственную связь

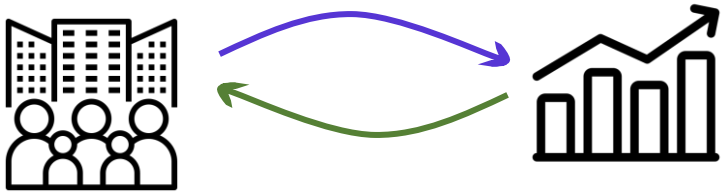
  X **->** Y

Институты* и экономический рост

Эмпирические исследования

В какую сторону причинно-следственная связь?

- Институты влияют на экономический рост
- Но и в богатых странах легче улучшать экономические институты



*Институты - не ВУЗы, а механизмы, по которым действовало общество

Институты* и экономический рост

Эмпирические исследования

В какую сторону причинно-следственная связь?

- Институты влияют на экономический рост
- Но и в богатых странах легче улучшать экономические институты

Инструмент?

- *Асемоглу-Джонсон-Робинсон:* смертность европейских колонистов от местных болезней 500 лет назад
- - В колониях, где европейцы жить не могли (Африка), они не вкладывались в развитие норм. институтов
- -В колониях, где европейцы жить могли (Америка), они развивали такие же институты, как в Европе

Результат: институты влияют на рост

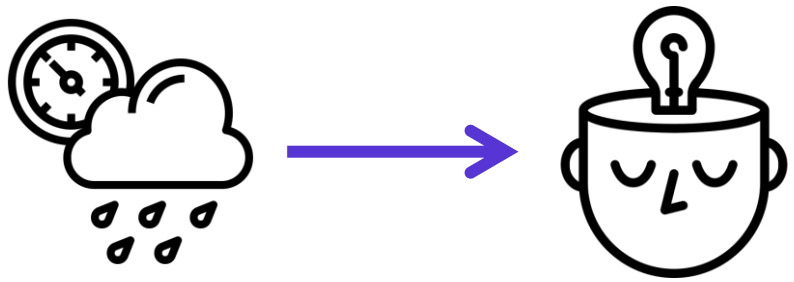
*Институты - не ВУЗы, а механизмы, по которым действовало общество

Атмосферное давление и мыслительные способности

Эмпирические исследования

Как оценить мыслительные способности людей поминутно?

- Атмосферное давление меняется каждую минуту
- Оценить изменение в том, как хорошо человек мыслит в каждую минуту, довольно тяжело

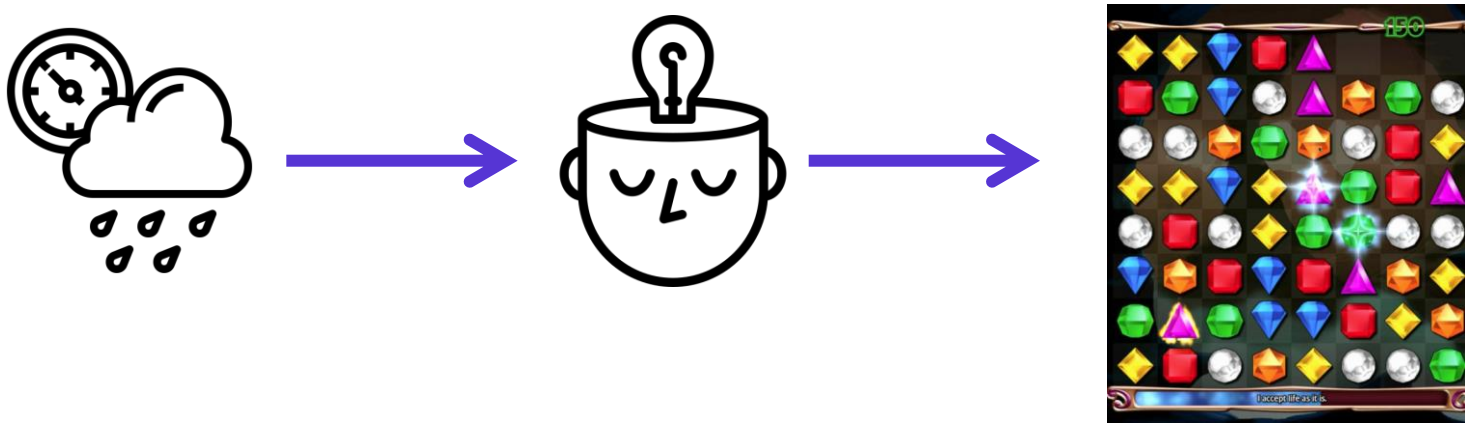


Атмосферное давление и мыслительные способности

Эмпирические исследования

Как оценить мыслительные способности людей поминутно?

- Атмосферное давление меняется каждую минуту
- Оценить изменение в том, как хорошо человек мыслит в каждую минуту, довольно тяжело



Метод?

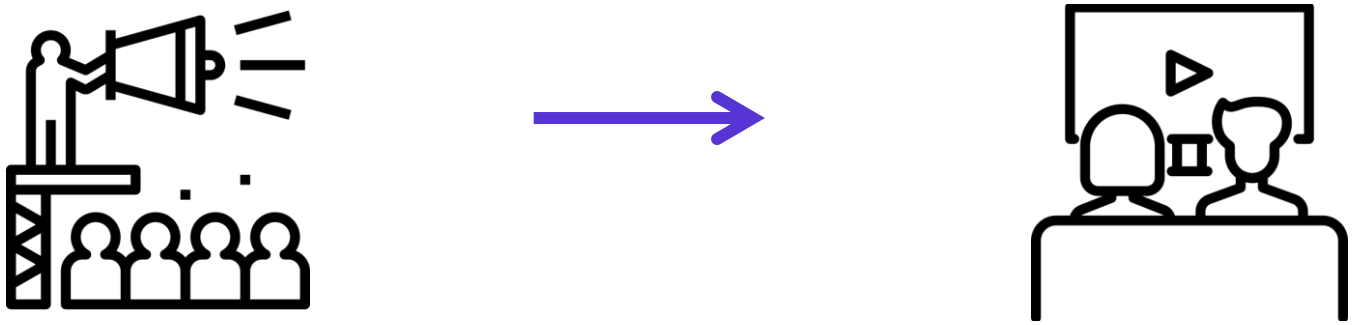
- Авторы взяли данные по результатам прохождения популярной онлайн игры на логику (в каждый момент времени большое количество людей играло в нее)
- Нашли отрицательную корреляцию между правильными ответами и изменением атмосферного давления

Результат: атмосферное давление влияет на когнитивные способности

Влияние пропаганды. Руанда

Эмпирические исследования

- – Как оценить влияние пропаганды без ошибки самоотбора?



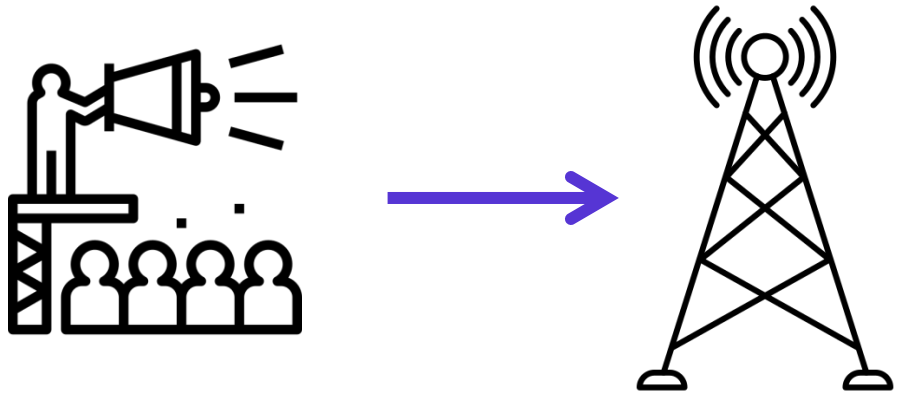
Метод?

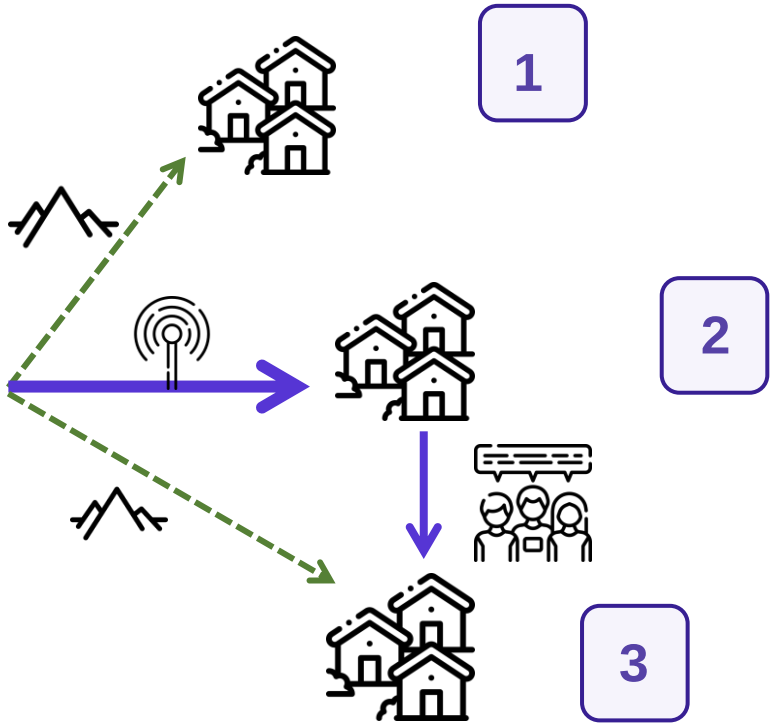
- Руанда 1994: геноцит народа Тутси <- Хуту (0.5млн-1млн человек)
- Пропаганда: «Свободное радио тысячи холмов»
- Страна горная => радиоволны проходили не везде => где-то Хуту слушали пропаганду, где-то – нет

Влияние пропаганды. Руанда

Эмпирические исследования

- – Пропаганда Хуту против народа Тутси





3 группы деревни

- Первая категория – удаленные горные деревни, где радио не ловило.
- Вторая – радиоволны проходили свободно, пропаганда вещала на полную.
- И третья – деревни, где радио не ловило, но расположенные в шаговой доступности от деревень второй категории

Выяснилось, что с особой жестокостью хуту убивали тутси именно в деревнях третьей категории (и, ожидаемо, гораздо меньше убивали во второй, где радио не было).

Рост как мера качества жизни.

Эмпирические исследования

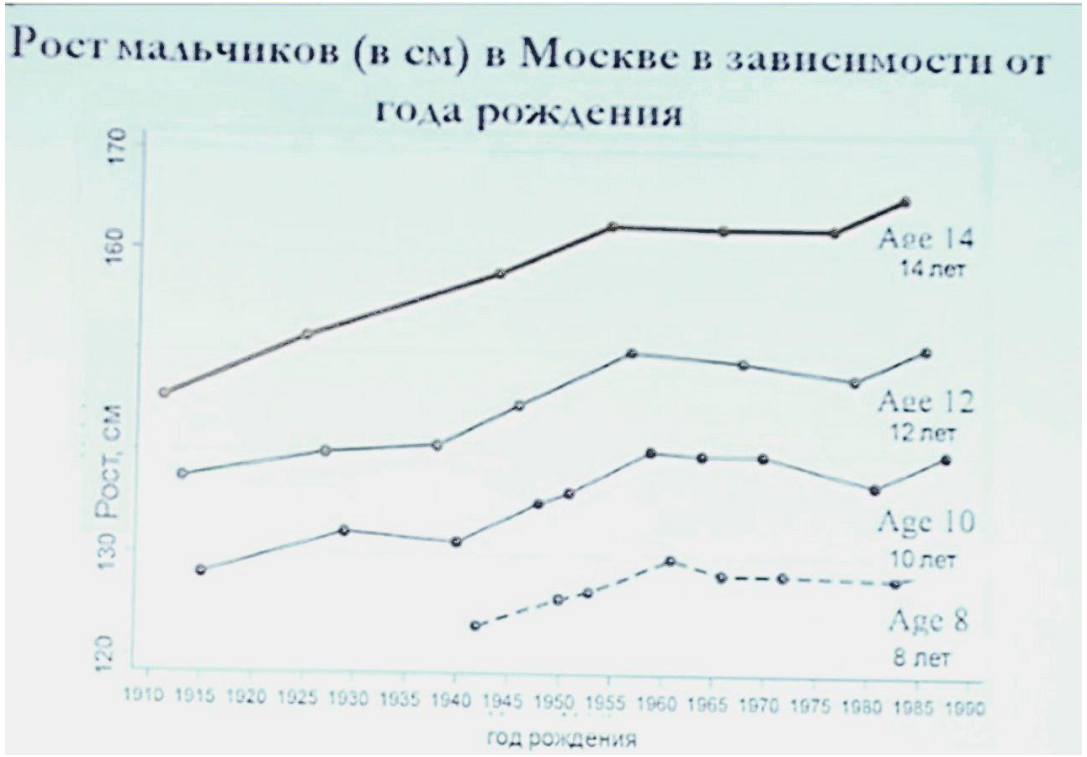
- Как сравнить благосостояние людей в странах (где нет данных о доходах)?

Метод?

- Рост - зеркало уровня развития (как, например, продолжительность жизни или детская смертность)

- Например, южные корейцы на несколько сантиметров (!) выше северных

- Brainerd (2008):
- В СССР качество жизни росло до 1960х гг., потом начало снижаться и даже отставать от США
- Переход к рынку улучшил ситуацию



Политический режим и экономический рост

Эмпирические исследования

• **(Перссон-Тамбеллини, 2008)**

- Демократизации приводят к увеличению темпов роста на

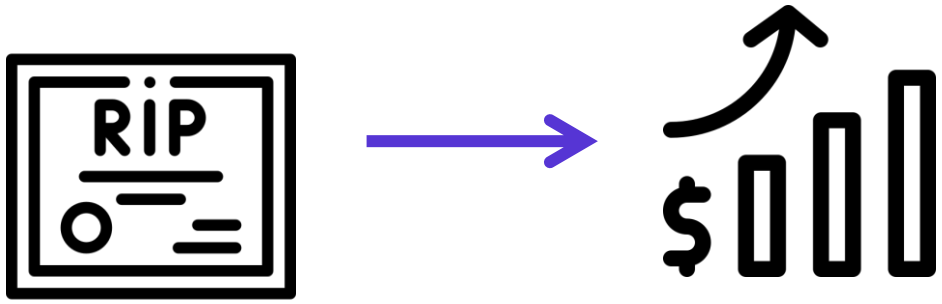
1% по сравнению с похожими странами, которые не демократизируются

- Переходы от демократии к диктатуре приводят к снижению темпов роста на **2%**

• **Event studies:**

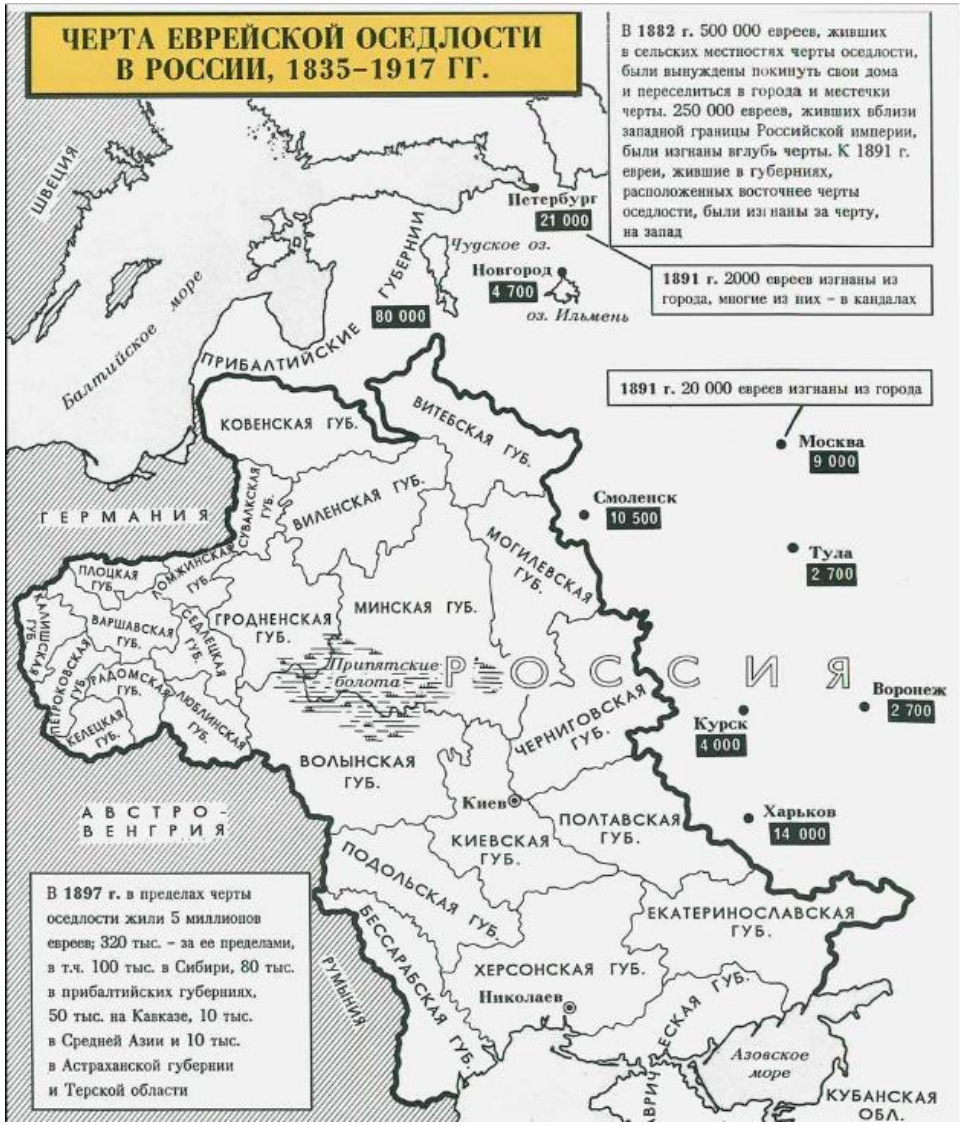
- Неожиданная смерть диктатора приводит к ускорению

экономического роста

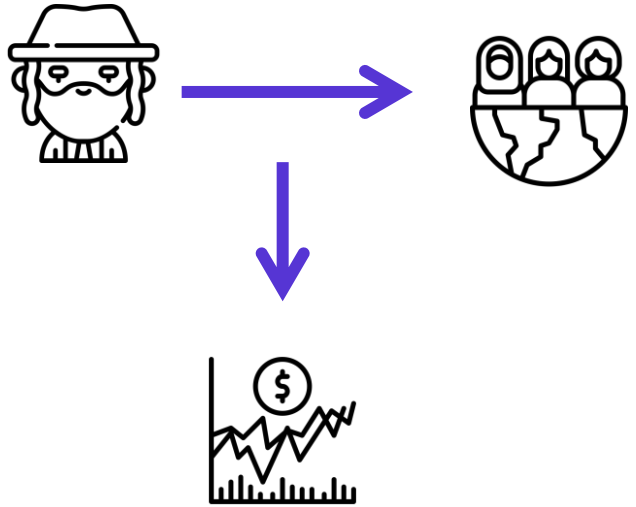


Черта оседлости и отношение к рынку

Эмпирические исследования



Е. Журавская, И. Гросфельд, А. Роднянский

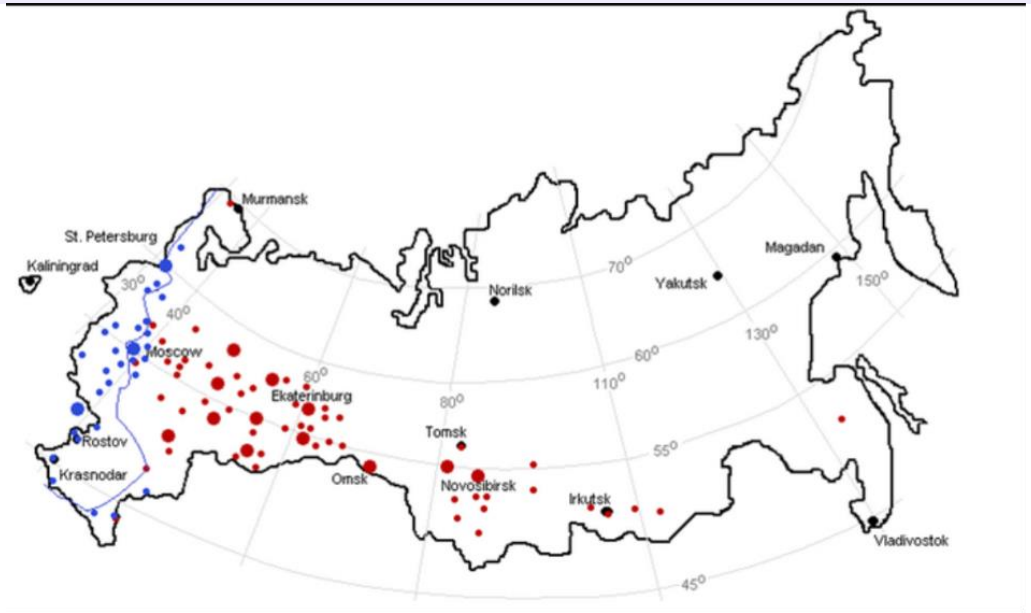


- Влияние черты оседлости (отмененной в 1917 году) на формирование ценностей : отношение к рынку, к « чужим » и т.п.

- Культурные различия, сформированные за годы ее существования, не только сохранялись в течение 40 или 60 лет, но проявляются и сегодня.

Гулаг и рост городов

Эмпирические исследования



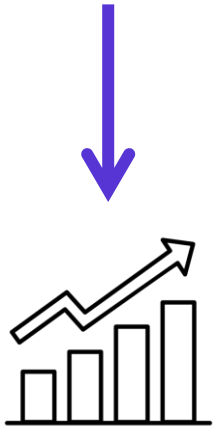
Т. Михайлова

- Влияние расстояния от ГУЛАГа до города на рост населенного пункта

- Результат: влияние ГУЛАГа имеет место до сих пор

ГУЛАГ:

– бесплатная рабочая сила
– «обслуживание городов»
– строительство капитала близ городов



Как измерить популяцию слонов?

Эмпирические исследования

- Ходить по джунглям и считать слонов дорого и неэффективно

Как измерить популяцию слонов?

Эмпирические исследования

- Ходить по джунглям и считать слонов дорого и неэффективно

Метод?

- «Прослушивать» слонов
- С помощью нейронной сети вычислить звуки слонов
- Бюджетный способ посчитать их количество

Лингвистика и экономика

Эмпирические исследования

- Ш. Вебер:
- Расстояние между языками (по «похожести» слов)

- Торговля между странами (народами)