# Исследование коронавирусной инфекции

Автор: Золотарев Даниил Александрович, факультет МКН СПбГУ.
email: danyzolotarev@gmail.com

Данные: https://ourworldindata.org/coronavirus
Исходный код: https://github.com/DanzillaPepe/covid-19_analysis

Цель: для нахождения закономерностей изучить влияние предоставленных в базе данных параметров на заразность и опасность коронавирусной инфекции.

## Выбор методов

Для достижения поставленной цели будет использоваться метод линейной регрессии, метод k-ближайших соседей и алгоритм имитации отжига для нахождения локального экстремума функции. Язык программирования – Python 3.9, используемые модули: *pandas*, *numpy*, *matplotlib*, *sklearn*.

## Выбор параметров

В качестве параметров будем использованы основные статичные показатели состояния стран, находящиеся в предоставленной базе данных. Они не меняются со временем. Их список будет приведён ниже. Для измерения же динамических показателей, сильно меняющихся с течением времени, введём несколько новых параметров, посчитанных из формулы оптимального (с точки зрения среднеквадратичного отклонения) коэффициента линейной регрессии.

- Средний прирост прививок при распространении заболевания:

$$k_{vaccinations/cases} = k_{v/c} = \frac{cov(X, Y)}{var(X)},$$

где

$$X = total\_cases,$$
$$Y = total\_vaccinations,$$
$$cov(X, Y) - \text{ковариация } X \text{ и } Y,$$
$$var(X) - \text{вариация } X \text{ (дисперсия)}.$$

- Средний прирост смертей при распространении заболевания:

$$k_{deaths/cases} = k_{d/c} = \frac{cov(X, Y)}{var(X)},$$

где

$$X = total\_cases,$$
$$Y = total\_deaths.$$

Для обоснования ввода этих параметров нужно количественно оценить, насколько данные коррелируют. Будем использовать коэффициент корреляции Пирсона, рассчитываемой по формуле:

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y},$$

где

$\sigma_X$ и $\sigma_Y -$ среднеквадратичные отклонения $X$ и $Y$ соотв.

Среднее значение коэффициента Пирсона, посчитанного в 244 странах, составило $0.88$ и $0.94$ для $k_{v/c}$ и $k_{d/c}$ соответственно, что позволяет делать выводы о взаимосвязи исследуемых параметров относительно друг друга.

*Замечание.* Используются именно коэффициенты линейной регрессии, а не обычные частные при делении соответствующих параметров, т.к. мы считаем, что линейная регрессия более устойчива к пробелам в данных и имеет бо́льшую предсказательную силу. На деле оказывается, что относительная разность этих коэффициентов и простых частных равна порядка 30% и 43% для $k_{v/c}$ и $k_{d/c}$ соответственно (под относительной разностью имеется в виду модуль разности значений, делённый на одно из них).

# Корреляция нововведенных параметров с известными

Параметры $k_{v/c}$ и $k_{d/c}$ были посчитаны для каждой страны, основываясь на последних возможных данных. Можно посмотреть на зависимость среднего числа заболевших (`total_cases_per_million`) от нового коэффициента $k_{v/c}$ во всех странах (График 1).
Распределение похоже на экспоненциальное. Поэтому прологарифмируем теперь ось Y и попробуем построить линейную регрессию (График 2).
Коэффициент Пирсона $r = -0.92$, что является показателем наличия довольно сильной антикорреляции.

Рассчитаем коэффициент детерминации $R^2$ по формуле:

$$R^2 = 1 - \frac{var(\varepsilon)}{var(Y)},$$

где

$\varepsilon$ – распределение остатков:

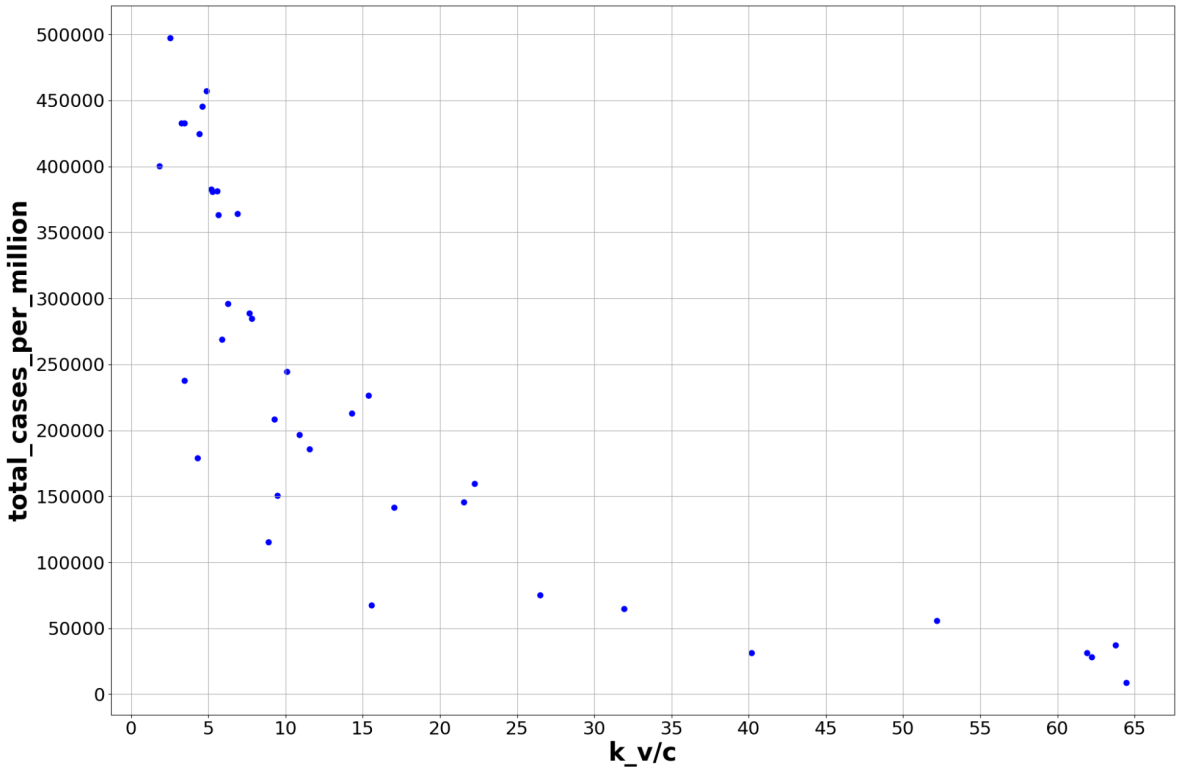
$$\varepsilon = Y - (k * X + b),$$

$$Y = total\_cases\_per\_million,$$

$k$ и $b$ – параметры полученной регрессии $Y = kx + b$.

В нашем случае коэффициент оказывается равен $0.84$. То есть около 84% дисперсии доли заболевших в стране можно объяснить темпом, с которым жители делают прививки.
Посмотрим теперь, как плотность населения (`population_density`) коррелирует с коэффициентом смертности $k_{d/c}$ (График 3). Здесь также неплохой коэффициент корреляции -$0.71$ и коэффициент детерминации $R^2 = 0.51$, значит примерно половина дисперсии роста смертности объясняется плотностью населения.

# График 1




# График 2

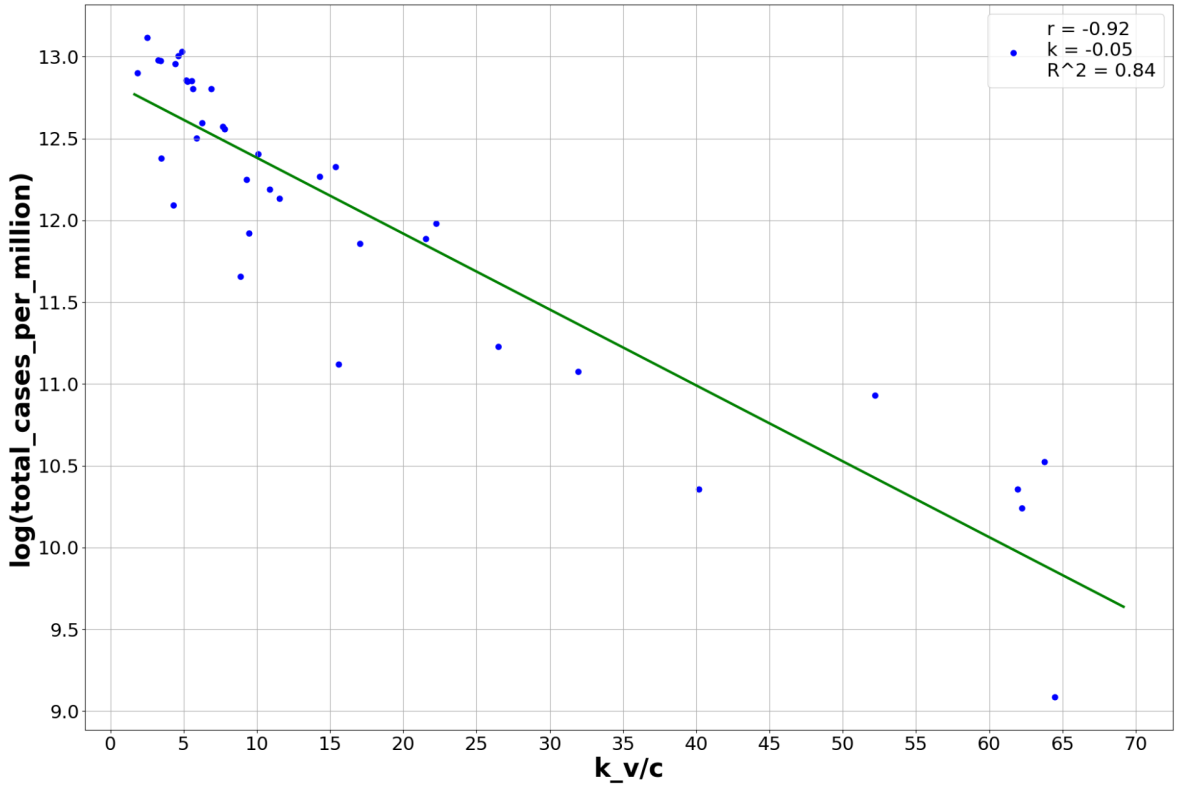
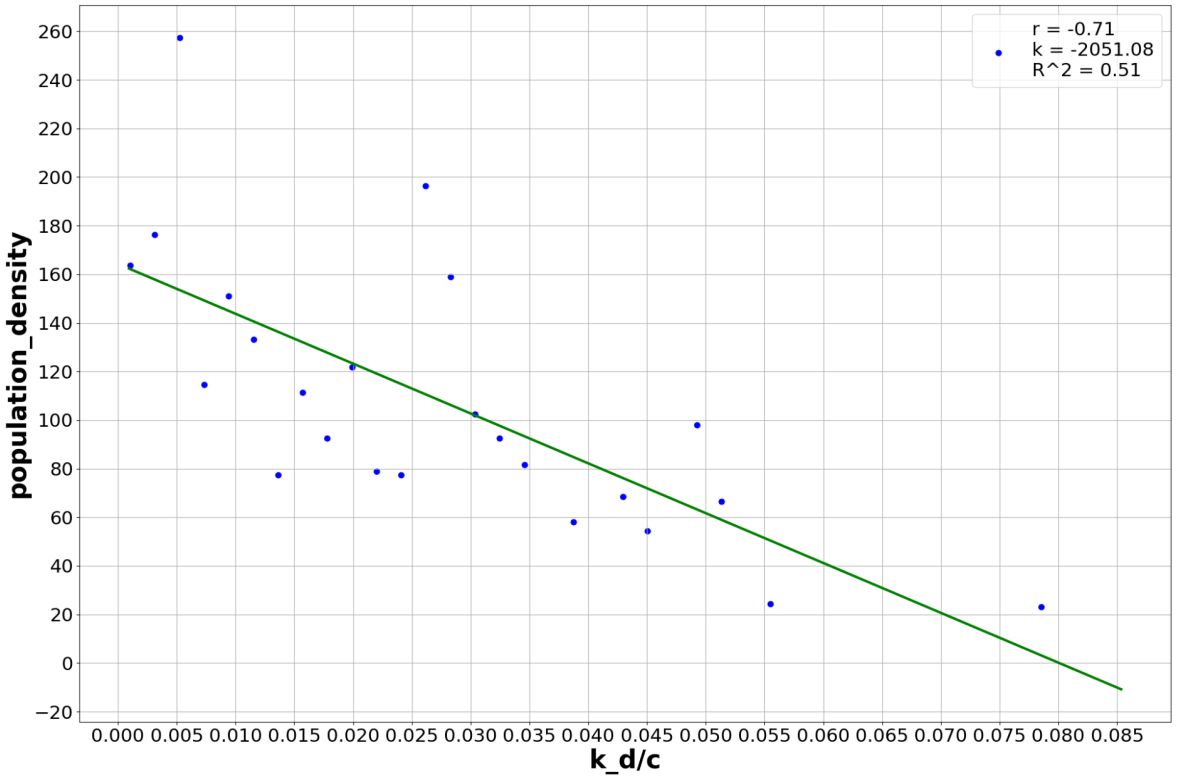


График 3




# Статичные параметры

Отберём те параметры, данные по которым представлены в достаточном количестве стран.
Показатели состояния здоровья населения (будут отмечаться зелёным):

- Процент больных диабетом (`diabetes_prevalence`)
- Смертность от сердечно-сосудистых заболеваний (`cardiovasc_death_rate`)
- Средний возраст (`median_age`)
- Доля жителей старше 65 лет (`aged_65_older`)
- Доля жителей старше 70 лет (`aged_70_older`)

Социально-экономические и санитарные показатели (будут отмечаться синим):

- Плотность населения (`population_density`)
- ВВП на душу населения (`gdp_per_capita`)
- Число мест в больнице на каждую 1000 человек (`hospital_beds_per_thousand`)

- Коэффициент прививания $k_{v/c}$ (`k_v/c`)
- Средний уровень мер по сдерживанию вируса (`stringency_index`)

## Метод k-ближайших соседей

Чтобы понять, какие данные больше всего определяют опасность вируса, построим kNN-модель для числа соседей k=5. Профильтровав все страны по наличию этих данных, получаем 148 стран. Из них выбираются ~90% стран – данные для обучения и ~10% оставшихся стран – тестовых данных. В задаче kNN существенность параметра можно соотнести с его “весом”, мультипликатором при пересчёте расстояний для выбора соседей:

$$d(X_1, X_2) = \sqrt{\sum_{i=1}^{n} w_i (X_1^i - X_2^i)^2},$$

где

$X_1, X_2 -$ два набора параметров,

$w_i -$ вес параметра $i$,

$X_1^i, X_2^i -$ $i$ параметр 1 и 2 набора параметров,

$n -$ число параметров.

Предсказываемыми параметрами будут общее число заболевших на миллион человек (`total_cases_per_million`) и число смертей на миллион человек (`total_deaths_per_million`)

## Использование имитации отжига

Для подбора оптимальных весов используется алгоритм имитации отжига. На каждом шаге совершается локальное изменение весов. Это прибавление к случайно выбранному из них случайного числа в диапазоне $[-1, 1]$. Далее пересчитывается значение коэффициента детерминации $R^2$ для новых весов. Если результат улучшился, он сразу принимается и становится текущим состоянием. Условие принятия при ухудшенном значении выглядит так:

$$rnd \leq e^{\frac{new\_score - old\_score}{t}},$$

где

$rnd$ – случайное число в диапазоне $[0,\ 1]$,

$new\_score,\ old\_score$ – старое и новое значение $R^2(weights)$,

$t$ – текущая температура.

Данное условие проверяется, если $new\_score$ оказывается не больше $old\_score$, и если оно оказывается выполнено, то всё равно старому значению $old\_score$ присваивается новое значение $new\_score$ и поиск продолжается. С уменьшением температуры условие срабатывает всё реже. Начальная температура $tMax$ имеет значение среднего изменения показателя $R^2$, выведенного эмпирическим способом и оказавшееся равным $0.05$. Меняется температура каждый раз домножением на коэффициент $tMult$, выбираемые в зависимости от желаемого времени работы. В нашем случае он составлял $1 - 4 * 10^{-7}$ или $1 - 1 * 10^{-7}$. Минимальная температура $tMin = 0.1 * tMax$. Алгоритм продолжается, пока текущая температура не упадёт ниже значения минимальной температуры $tMin$.

# Результаты

Для заболеваемости (Таблица 1) наиболее значимыми оказались возраст, количество больных диабетом, плотность населения и лишь затем темп прививания.
Для смертельности (Таблица 2) же наоборот, возраст имеет куда меньшую значимость, тогда как на первое место встаёт прививание, ВВП, число мест в больницах на 1000 человек, плотность населения и меры сдерживания. Сердечно-сосудистые заболевания в обоих случаях играют незначительную роль в 7% и 4%.

Таблица 1

| **Целевой параметр** | `total_cases_per_million` |
|---|---|
| *Значение* $R^2$ | 0.90 |

| № | Имя параметра | Вес | Доля |
|---|---|---|---|
| 1 | aged_65_older | 621.26 | 24% |
| 2 | aged_70_older | 442.50 | 17% |
| 3 | diabetes_prevalence | 386.34 | 15% |
| 4 | median_age | 273.72 | 11% |
| 5 | population_density | 207.91 | 8% |
| 6 | k_v/c | 204.10 | 8% |
| 7 | cardiovasc_death_rate | 164.89 | 7% |
| 8 | stringency_index | 151.77 | 6% |
| 9 | hospital_beds_per_thousand | 87.93 | 3% |
| 10 | gdp_per_capita | 10.85 | 1% |

Таблица 2

| **Целевой параметр** | total_deaths_per_million |
|---|---|
| *Значение* $R^2$ | 0.85 |

| № | Имя параметра | Вес | Доля |
|---|---|---|---|
| 1 | k_v/c | 1694.61 | 24% |
| 2 | gdp_per_capita | 1516.79 | 21% |
| 3 | hospital_beds_per_thousand | 1136.59 | 16% |
| 4 | population_density | 894.36 | 13% |
| 5 | stringency_index | 677.34 | 9% |
| 6 | diabetes_prevalence | 610.48 | 9% |
| 7 | cardiovasc_death_rate | 290.15 | 4% |

| | | | |
|---|---|---|---|
| 8 | aged_70_older | 249.74 | 3% |
| 9 | aged_65_older | 30.31 | 1% |
| 10 | median_age | 1.78 | 0% |

# Вывод

Как мы увидели из обработанных данных, на заболеваемость бо´льшее влияние имеют факторы здоровья: возраст, хронические заболевания.

В то же время смертность зависит скорее от социально-экономических и санитарных условий: прививание, ВВП на душу населения, число мест в больницах на 100 человек, меры по сдерживанию. Естественно предположить, что в обоих случаях не учтены некоторые дополнительные параметры, такие, как климат с одной стороны или экономические издержки с другой.