Поиск факторов, влияющих на смертность от COVID-19 на разных континентах

Дорушенкова Ольга

2022

Цели проекта:

∗ Найти факторы, влияющие на показатель смертности от ковида на разных материках

∗ Проверить предположение о влияние курения на смертность

∗ Проанализировать полученные данные

∗ Сравнить течение ковида в разных странах

Используемые инструменты для выполнения проекта:

∗ Метод многофакторной линейной регрессии

Используемые параметры и исследуемые континенты:

∗ Под понятием смертность в данной работе подразумевается доля людей, умерших из за COVID-19, от всего населения страны. Такое решение было принято, потому что процент смертности от заболевания, то есть процент умерших от количества зараженных, в целом одинаковый для большинства стран. Это обьясняется тем, что распространение вируса и процент умерших от населения страны зависят от примерно одних и тех же параметров. Поэтому для каждого определенного вируса есть фиксированный процент умирающих заболевших, и зависит он только от тяжести вируса для человеческого организма.

∗ Независимые параметры по каждой стране: индекс человеческого развития; процент: живущих за чертой бедности, больных диабетом, людей страше 70, курящих мужчин и женщин, людей, которые имеют возможность мыть руки; количество мест в больнице на тысячу населения, ожидаемоя длительность жизни при рождении, ВВП на душу населения, общий процент курящих людей.

Параметры, которые были использованы для каждого континента, подбирались отдельно для того, чтобы выбрать наилучшую модель для кадого материка.

∗ Исследуемые континенты: Северная Америка, Азия, Европа и Африка. В Австралии и Южной Африке слишком мало данных для исследования по этим параметрам.

1 Европа

В статистике для Европы использовалась информация о 40 странах.

Перед использованием множественной линейной регрессии была использована простая регрессия относительно каждого из параметров,перечисленных ранее.Прямая зависимость была обнаружена от общего процента курящих людей в стране и индекса человеческого развития. Но хорошими такие модели точно нельзя назвать, так как коэффицент линейной регресии $R^2 = 0,39$ для индекса развития человека и $R^2 = 0,44$ для общего процента курильщиков.

Рис. 1

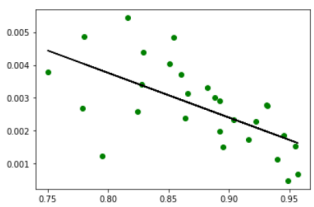
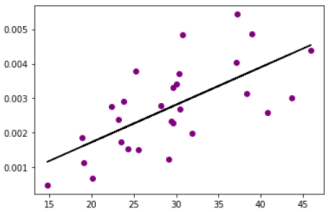


Рис. 2



На первом (зеленом) графике зависимости смертности от индекса человеческого развития четко видно, что зависимость обратная. Действительно, чем больше в стране индекс человеческого развития, тем выше уровень жизни в стране, тем меньше процент смертности.

На втором (фиолетовом) графике зависимости смертности от процента курящих видно, что зависимость прямая. Так как при курении легкие уже поражены до болезни, вирусу проще убить человека.

Теперь перейдем к множественной линейной регресии. Для этого были взяты 9 парметров, их можно увидеть в таблице(3) ниже.

	Coefficient
aged_70_older	0.951497
human_development_index	-2.966266
extreme_poverty	0.034306
diabetes_prevalence	-1.896530
male_smokers	-0.807705
female_smokers	1.229210
hospital_beds_per_thousand	-0.015963
life_expectancy	-0.460566
gdp_per_capita	-0.000046

	Actual	Predicted
4	27.785577	20.005775
31	25.830454	25.997814
21	33.054014	29.825477
24	37.987840	37.301078
7	40.269597	48.284843
18	15.441612	11.464230

.

Коэффицент в этой таблице обозначает величину, на которую изменится смертность (в $0,0001$) при увеличении определенного парметра на 1 процент.

В таблице явно видно,что смертность в основном зависит от процента курящих женщин, индекса человеческого развития и процента болеющих диабетом в стране. Наиболее трудно обьяснить зависимость, причем обратную, от процента больных диабетом. Мое предположение состоит в том, что в странах, где больше процент болеющих диабетом, более развитая медицина и в обществе более развита культура внимательного отношения к здоровью. Так как многие люди живут с диабетом, не зная о своем диагнозе, больший процент в статистике говорит, о том что заболевания просто чаще выявляют, так как люди чаще проверяются. Таким образом больший процент диабетиков является следствием хорошей медицины, наличие которой уменьшает смертность.

На второй таблице видно, как отличаются реальные показатели некоторых стран от предсказанных с помощью данной модели. Коэффицент $R^2 = 0,63$, что в сумме с таблицей с прогнозируемыми данными позволяет сделать вывод,что модель вполне отражает реальный процент смертности в данных странах.

2 Африка

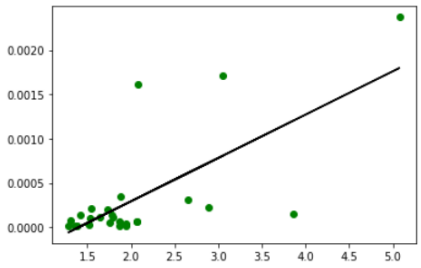
В статистике для Африке использовалась информация о 36 странах.

Отличительной чертой этого континента является бедность и неразвитая медицина. В некоторых странах процветает голод, также из за низкого уровня образования и недоверия к современной медицине жители Африки чаще в первую очередь обращаются к народной медицине. Все это отражается на статистике смертности. Скорее всего данные не совсем корректны из-за невозможности учесть смерть людей, которые так и не обратились к медикам.

Как и ранее, перед использованием множественной линейной регрессии была использована простая регрессия относительно каждого из параметров. Прямая зависимость смертности с коэффицентом $R^2 = 0,49$ была найдена от процента людей старше 70. Но на графике(3) явно видно, что большинство точек расположены в левом углу и это не позволяет назвать модель качественной.

В множественной регресии было использовано 7 параметров. В представленной таблице явно видно, что в основном на смертность влияет процент людей старше 70. Причина заключается в том, что старым людям в африканских странах редко оказывается медицинская помощь, и смертность среди них от любого вируса в эпидемию наибольшая. Коэффицент множественной линейной регрессии $R^2 = 0,71$, в целом модель для Африки хорошая.

Рис. 3

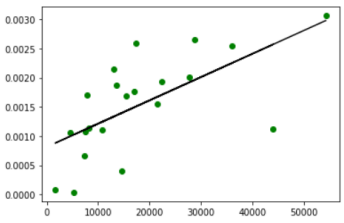


	Coefficient
aged_70_older	5.500176
human_development_index	0.146653
extreme_poverty	-0.091020
diabetes_prevalence	-0.269628
male_smokers	0.061775
handwashing_facilities	-0.133046
female_smokers	0.752483

3 Северная Америка

В статистике для Северной Америки использовалась информация о 35 странах.

Рис. 4



	Coefficient
aged_65_older	-0.038561
life_expectancy	-0.205845
diabetes_prevalence	1.059700
gdp_per_capita	0.000333

	Actual	Predicted
16	6.658044	10.474147
22	11.309569	13.036913
35	30.626240	26.719077
2	25.437411	22.140655
20	0.765228	8.296446

.

В северной Америке расположено три крупных страны Канада, Мексика и США, остальные же - это в основном островные государства или совсем небольшие страны.

При применении простой линейной регресии лучшей зависимостью, которую удалось найти, является зависимость смертности от ВВП на душу населения. $R^2 = 0,42$, следовательно эта модель, как и аналогичные на других континентах, не очень качественная, что можно заметить на самом графике(4).

Теперь переходим к множественной регресии: было использовано 4 параметра, что сильно меньше по сравнению с другими континентами, но при добавлении других параметров качество модели сильно ухудшалось. Интересная особенность заключается в том, что в отличии от Европы, где зависимоть смертности от процента больных диабетом была обратной, в Америке зависимость прямая, и это основной фактор, от которого зависит смертность. Мое предположение состоит в том, что процент больных диабетом настолько большой в данных странах, что эффект, при котором больше людей следит за своим здоровьем, теряется на фоне негативного влияния сахарного диабета на иммунитет и здоровье в целом, что делает человека более уязвимым для вируса. Итоговый коэффицент $R^2 = 0,84$, это лучшаая полученная модель среди всех континентов, но в таблице с реальными данными и предсказанными моделью видно, что в некоторых случаях она все равно работает не идеально.

.

4 Азия

Азия в данном случае кардинально отличается от других континентов, так как на ней расположено очень много стран с различным государственным строем и уровнем жизни. Например в Японии очень высокий уровень жизни и доступная медицина, в Индии наоборот очень большой процент бедности и нет доступа к врачебной помощи, также есть закрытые страны. Из за того что в Азии расположены координально разные страны у нас недостаточно данных для того, чтобы построить качественную модель с линейной регрессией.

	Coefficient
aged_70_older	0.527609
human_development_index	0.372702
diabetes_prevalence	0.375497
gdp_per_capita	-0.000179

При применении простой линейной регресси не удалось найти параметр от которого бы смертность зависила хотя бы с коэффицентом $R^2 > 0,2$. Что еще сильнее показывает отличие от остальных континентов.

Перейдем к линейной множественной регресии. Максимальный коэффицент которого удалось достигнуть $R^2 = 0,34$. К сожалению, для Азии не удалось построить качественную модель.

5 Волны COVID-19 на разных континентах

Для начала рассмотрим как протекала болезнь в разных странах. Выясним,сколько на каждом континенте было волн заболеваемости и какая из них была наиболее смертностной

Посмотрим на графики построенные для каждого континента.

Синия кривая - количество новых случае за день, разделенное на 100.

Черная кривая - количество новых смертей за день

Рис. 5: Европа и Африка

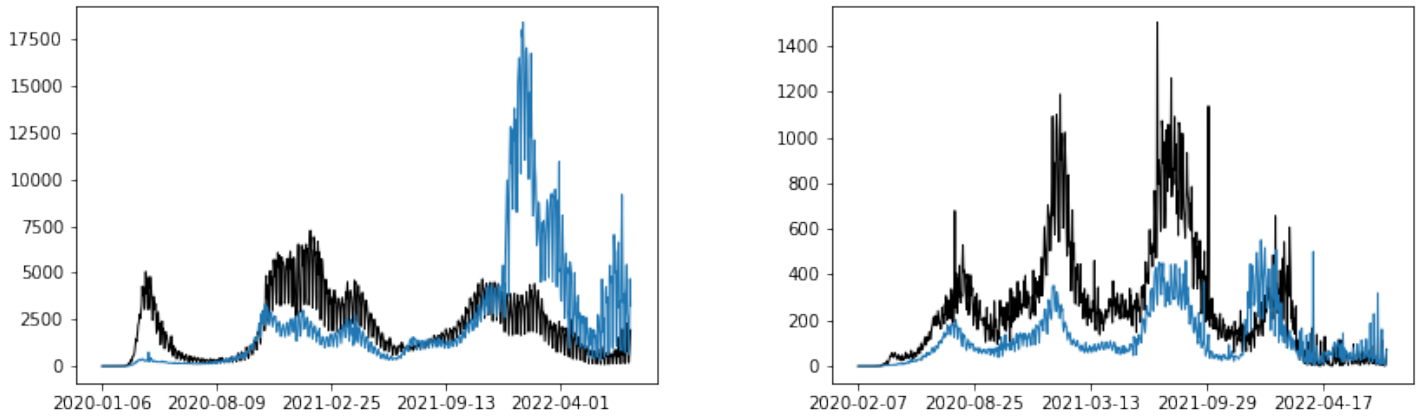
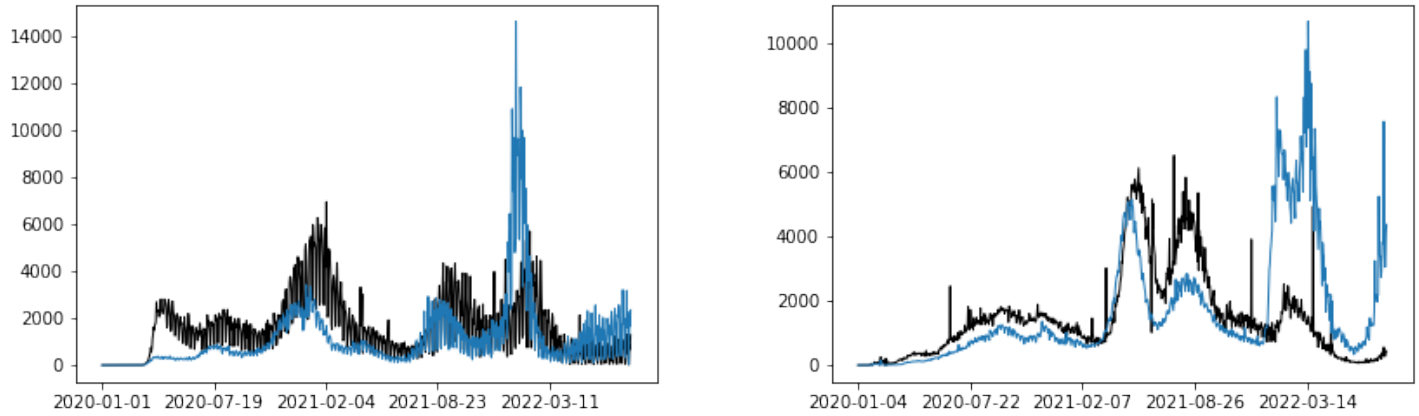


Рис. 6: Северная Америка и Азия



Первое, что важно заметить, что количество волн на континентах разное. В Европе на синий кривой видно 5 пиков: три маленьких и 2 больших, значит волн COVID-19 было 5. В Африке было 4 волны, в Северной Америке и Азии волны не так сильно выделяются, но все равно можно выделить 4 волны в Америке и 6 волн в Азии.

Теперь перейдем к смертности. В Европе максимальное количество смертей было в первые три волны, в Африке смертнось во все волны стабильно высокая как и в Северной Америке. В Азии максимальная смертность была в 3 и 4 волны.

6 Влияет ли курение на смертность?

На прямую курение не влияет на смертность от COOVID-19 ни на одном континенте, кроме Европы. В моделях с множественной регрессией на других континентах видно что в них даже не везде присутствует процент курящих мужчин или женщин. Обьяснением этого является то, что на других континентах существуют признаки, которые перекрывают влияние курения. Например в Африке большой процент бедности, в Европе же хороший уровень жизни и доступная медицина, что позволяет курению сильнее влиять на процент смертности. Таким образом можно сделать вывод, что действительно при заболевании ковидом у двух людей, которые имеют доступ к качественному лечению, вероятность выжить больше у того кто не курит.

7 Вывод

Таким образом мы построили 3 вполне рабочие модели для Африки, Северной Америки и Европы с помощью множественной линейной регресии. Также взглянули на то как протекала эпидемия на разных континентах и проанализировали влияние смертности и других факторов. Наиболее интересным открытием является обратная зависимость смертности от больных диабетом в Европе. В данной работе страны обьединялись по территориальному признаку, но в большинстве случаев это также означало схожесть в уровне жизни и культуре, что как раз и повлияло на возможность создать модели с помощью линейно регресси. В Азии же где страны, культуры и народы очень разнобразные для качественной модель требуется учитывать другие факторы.