Федеральное государственное образовательное бюджетное

учреждение высшего профессионального образования

«НАЦИОНАЛЬНЫЙ-ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Высшая школа бизнеса

Проект

по дисциплине “Количественные методы в экономике”

Выполнили:

Авдеева Мария

Барбаш Лев

Шохолова Елена

Москва - 2022

1. **Выбранная статья для анализа.**

За основу работы была выбрана статья [“Save lives or save livelihoods? A cross-country analysis of COVID-19 pandemic and economic growth”](https://www.sciencedirect.com/science/article/pii/S0167268122000804) из журнала “Journal of Economic Behavior and Organization”.

1. **Постановка исследовательской задачи. Актуальность темы. Формулировка интересных содержательных гипотез.**

**Исследовательская задача:**Выявить связь между различными показателями (ВВП на душу населения, плотность населения, инфраструктура здравоохранения, эффективность правительства и т.д.) и смертностью от COVID-19 с учетом межстрановых различий.

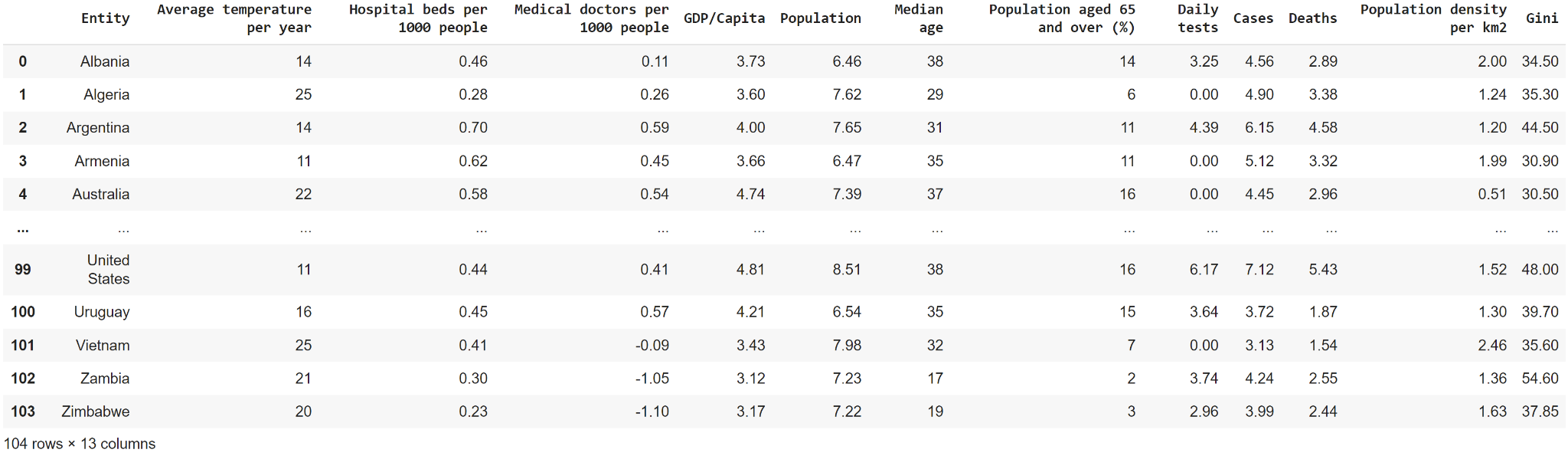
**Исследовательский вопрос:** Какие факторы могут повлиять на смертность во время пандемии COVID-19?

**Актуальность темы:**Пандемия коронавируса нанесла прямой ущерб экономикам стран мира через заболеваемость и смертность людей. Различные страны сообщают о разных количествах инфицированных COVID-19, причиной этому может быть неоднородность некоторых факторов риска и защиты от болезни. Влияние таких факторов, как ВВП, плотность населения, температура воздуха и прочее, на распространение и смертность от COVID-19 в разных странах недостаточно изучено, что и мотивирует наши межстрановые регрессионные анализы.

**Гипотезы:**

1. Модель, которая учитывает все показатели, будет лучше предсказывать смертность по стране чем модель, которая учитывает только показатели страны без данных о пандемии;
2. Показатель смертности от COVID-19 будет сильно коррелировать с показателем популяции страны;
3. Показатели температуры страны и количества больничных коек будут одними из самых важных защитных факторов в модели для предсказания смертности;
4. Показатели ВВП на душу населения и неравенства доходов населения будут одними из самых угрожающих факторов в модели для предсказания смертности.
5. **Источник данных, описание показателей (как рассчитываются и в чем измеряются). Анализ описательных статистик и графический анализ переменных.**

В качестве источника данных был выбран датасет [“COVID19-Dataset-with-100-World-Countries”](https://www.kaggle.com/datasets/sambelkacem/covid19-algeria-and-world-dataset) с сайта kaggle. Для его дальнейшего использования была проведена обработка данных. В частности, взяты данные только на 27 ноября 2020 года, как и в статье. Также были добавлены столбцы по плотности населения и коэффициенту Джини. Все данные были прологарифмированы для лучшего предсказания трендов, а не конкретных значений. Всего выбрано 104 страны и 12 различных показателей. Итоговая таблица для анализа показана на рисунке 1.



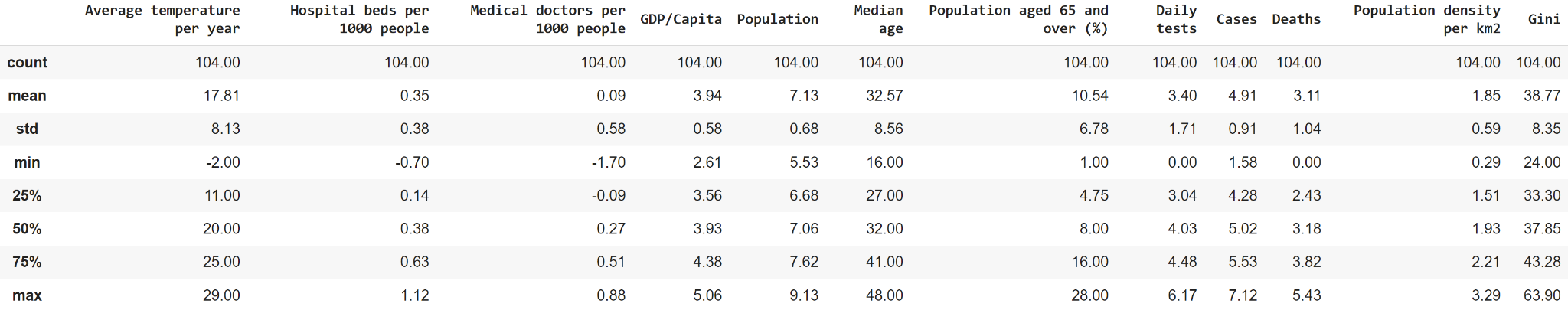
*Рисунок 1 - Таблица для анализа*

Ниже, в таблице 1, представлено описание всех показателей.

*Таблица 1 - Описание показателей*

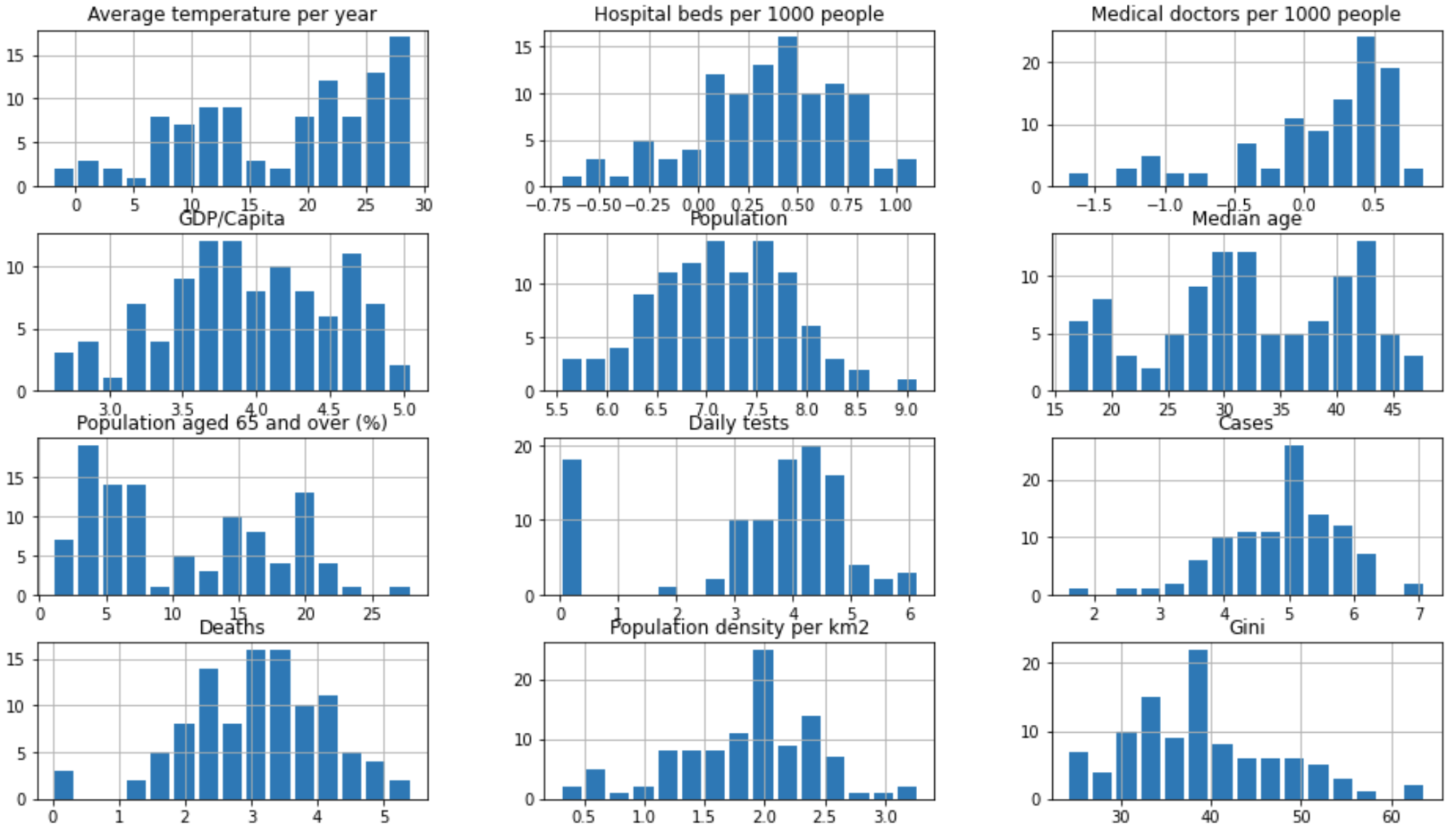
| **Показатель** | **Описание** |
| --- | --- |
| **Entity** | Страна, строковая переменная |
| **Average temperature per year** | Средняя температура за год в градусах Цельсия, числовая переменная |
| **Hospital beds per 1000 people** | Количество больничных коек на 1000 человек, числовая переменная |
| **Medical doctors per 1000 people** | Количество докторов на 1000 человек, числовая переменная |
| **GDP/Capita** | ВВП на душу населения, числовая переменная |
| **Population** | Население, числовая переменная |
| **Median age** | Медианный возраст, числовая переменная |
| **Population aged 65 and over (%)** | Население старше 65 лет в %, числовая переменная |
| **Daily tests** | Количество сделанных тестов на COVID-19, числовая переменная |
| **Cases** | Количество выявленных случаев COVID-19, числовая переменная |
| **Deaths** | Количество смертей, числовая переменная |
| **Population density per km2** | Плотность населения на км2, числовая переменная |
| **Gini** | Значение коэффициента Джини, числовая переменная |

На рисунке 2 представлена таблица описательных статистик всех числовых переменных.



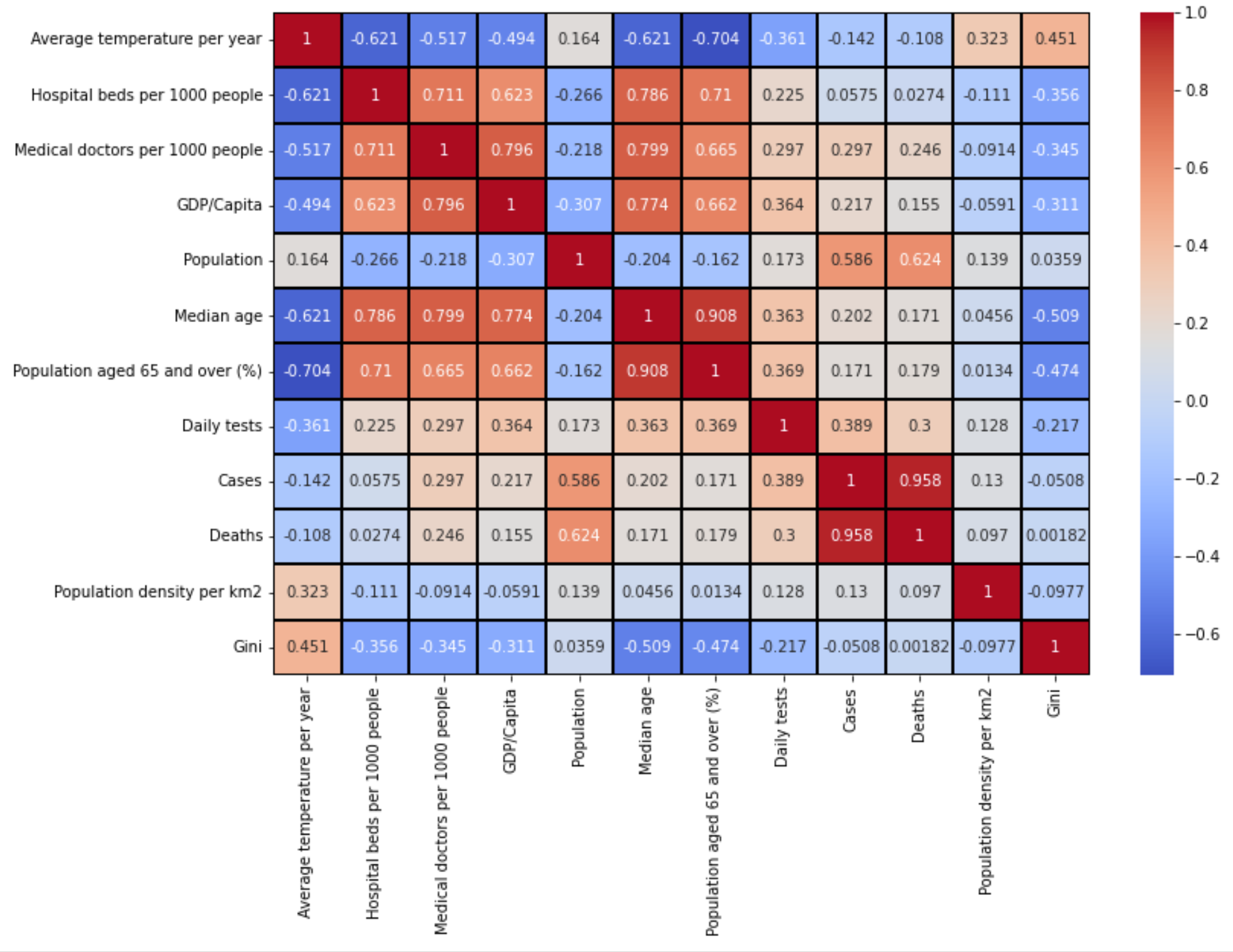
*Рисунок 2 - Описательные статистики переменных*

На рисунке 3 можно видеть гистограммы всех числовых переменных.



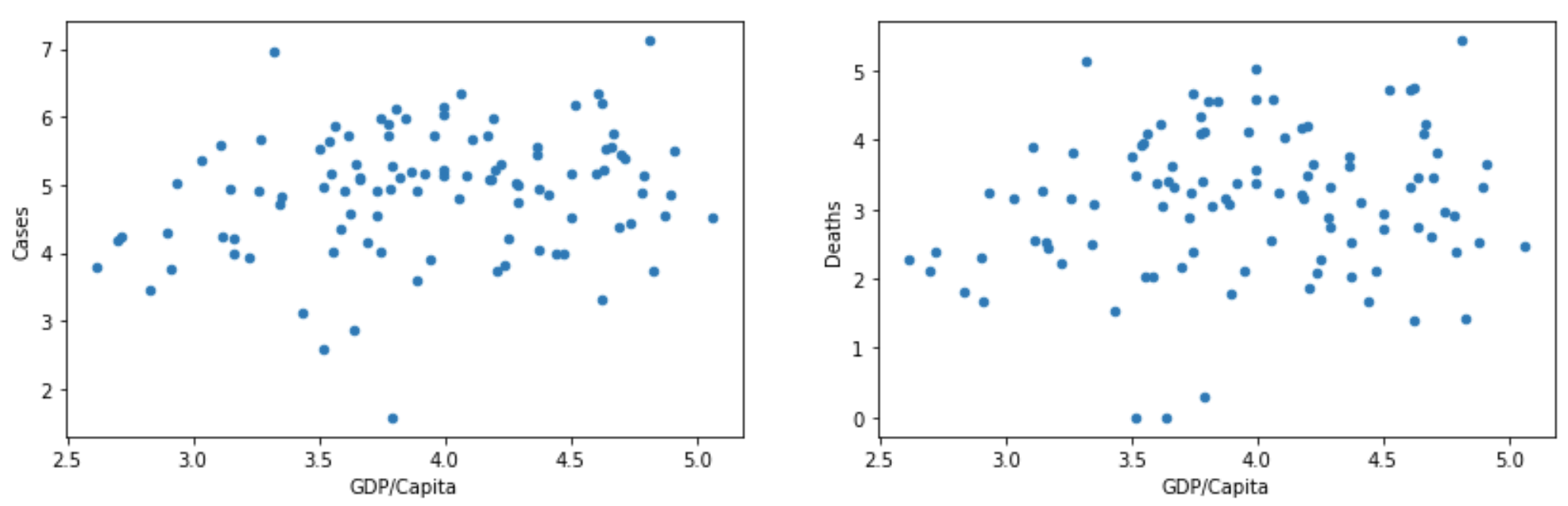
*Рисунок 3 - Гистограммы числовых переменных*

На рисунке 4 можно видеть корреляционную матрицу всех числовых переменных.



*Рисунок 4 - Корреляционная матрица числовых переменных*

Далее, на основе корреляционной матрицы числовых переменных можно построить графики рассеивания, чтобы лучше увидеть взаимосвязь отдельных пар показателей. Например, графики рассеивания показателей выявленных случаев и ВВП, смертей и ВВП на рисунке 5 (больше графиков можно увидеть к приложенному файлу с кодом).

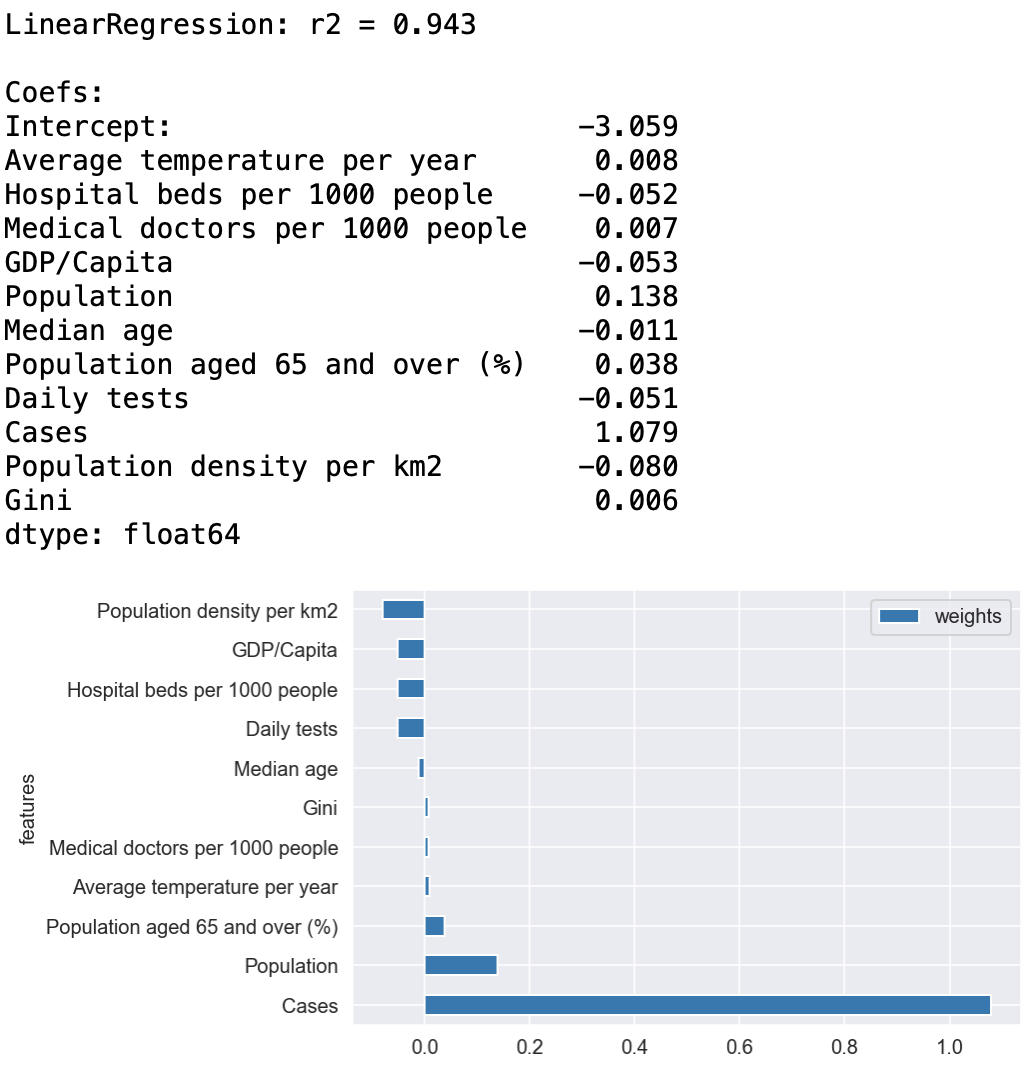


*Рисунки 5 - Графики рассеивания*

1. **Построение эконометрических моделей. Корректное оценивание и применение эконометрических методов. Проверка предпосылок теоремы Гаусса-Маркова и гипотезы о нормальности случайной ошибки. Выбор наилучшей модели.**

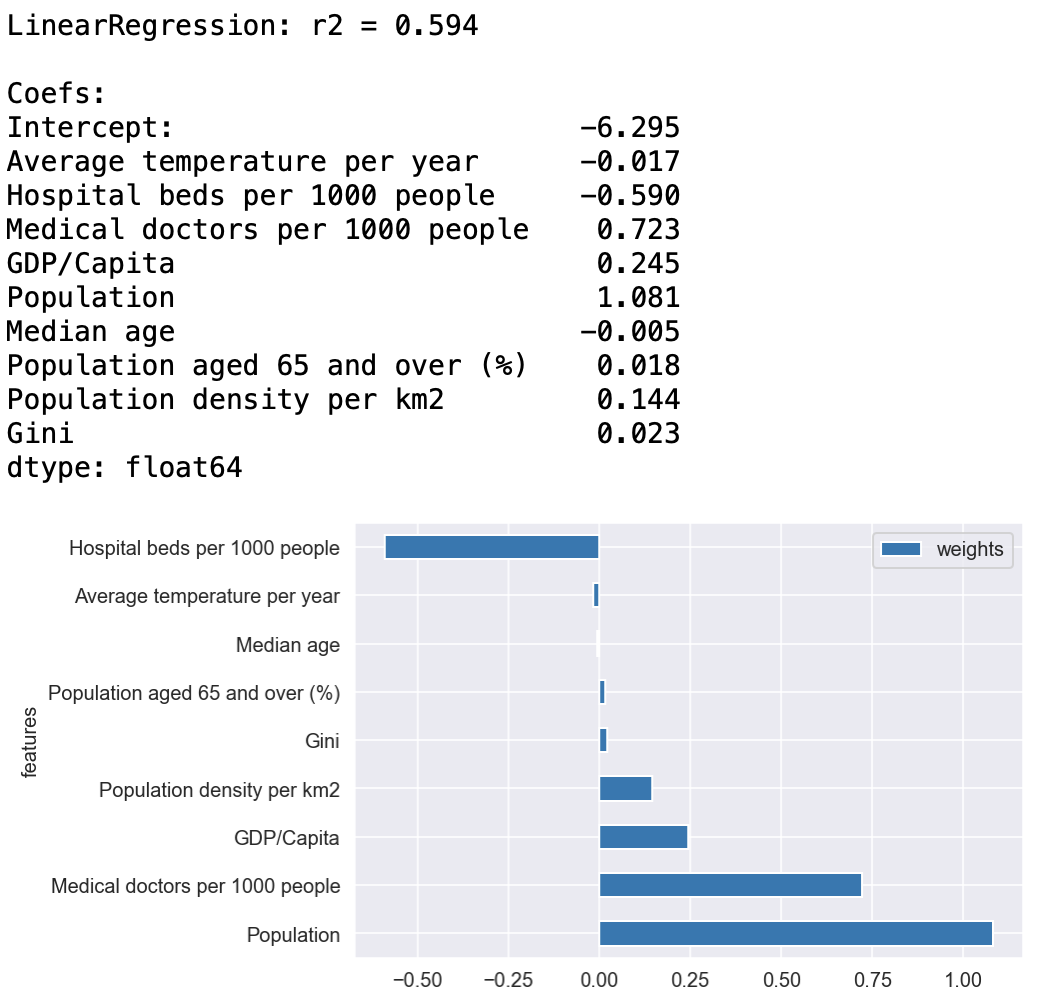
На основании выбранных гипотез и анализа переменных были предложены три линейные модели. Каждая из них подойдет для определенных целей исследования и предсказания.

Первая модель учитывает все показатели и на основе их предсказывает смертность в стране. На рисунке 6 можно видеть коэффициенты при показателях в модели.



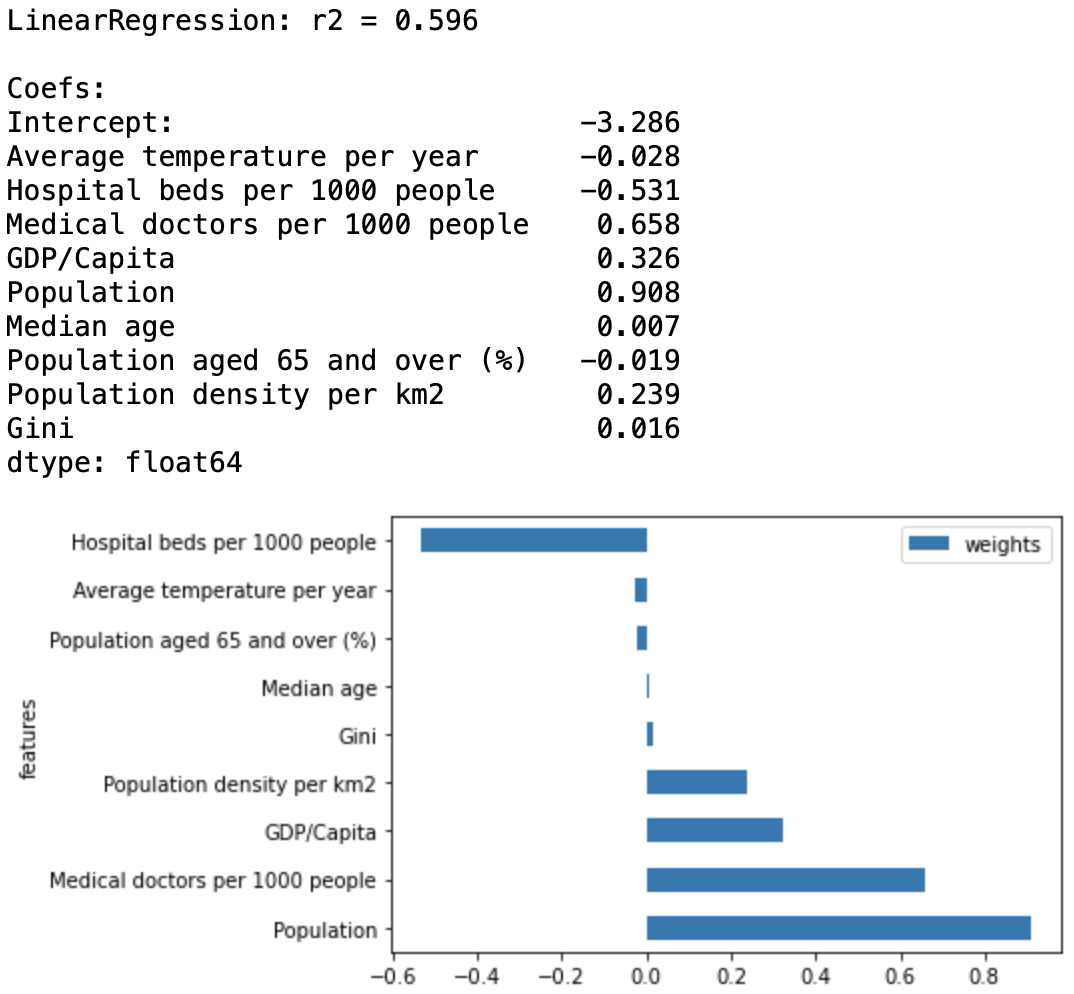
*Рисунок 6 - Коэффициенты первой модели*

Вторая модель предсказывает смертность в стране на основе общих статистических данных по странам, не учитывающих показатели, связанные с пандемией. Ее коэффициенты представлены на рисунке 7.



*Рисунок 7 - Коэффициенты второй модели*

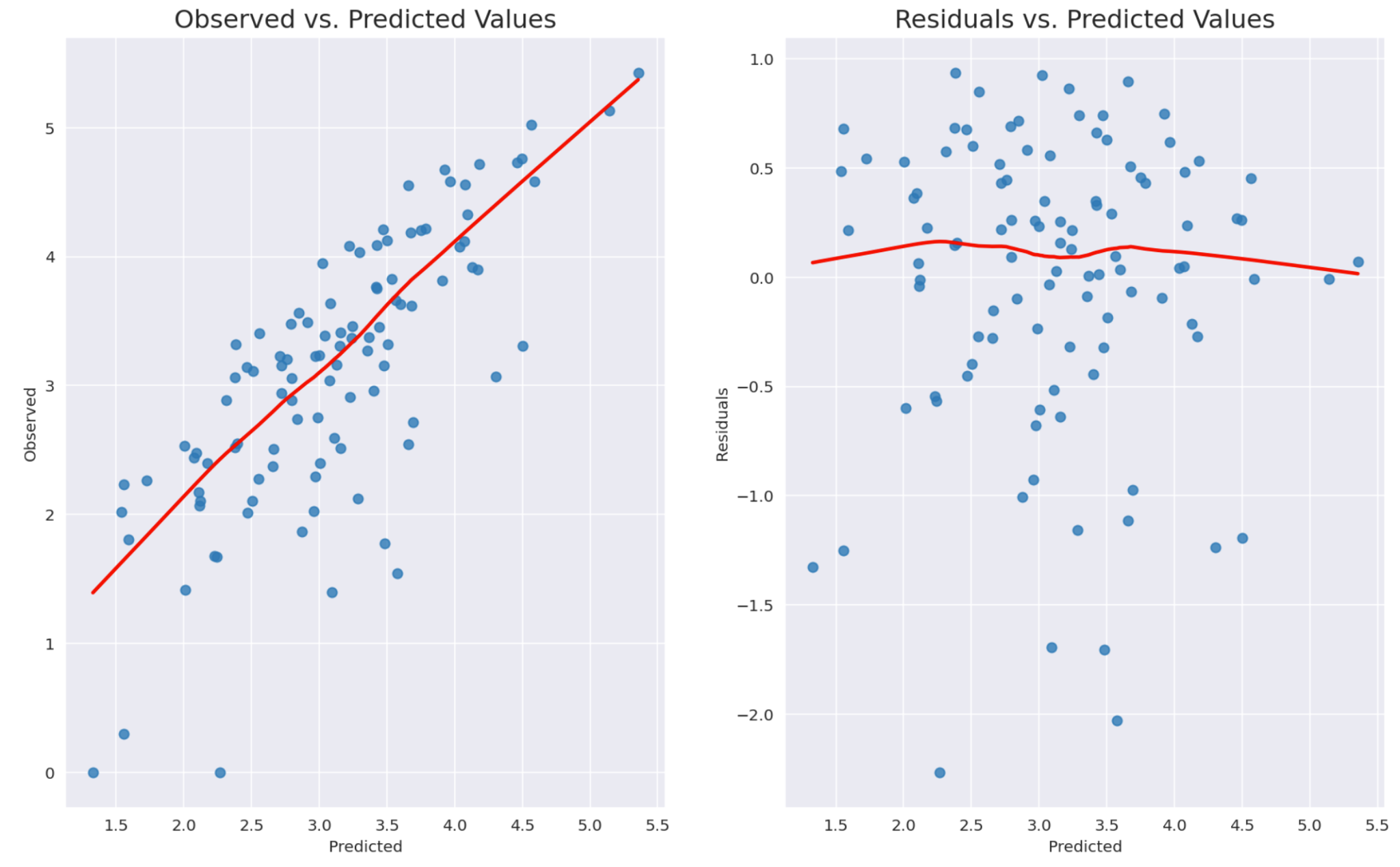
Третья модель, коэффициенты которой показаны на рисунке 8, на основе общих статистических данных по странам, не учитывающих показатели, связанные с пандемией предсказывает случаи заражения в стране.



*Рисунок 8 - Коэффициенты третьей модели*

Теперь проверим все предпосылки теоремы Гаусса-Маркова. Ниже приведен пример для второй модели.

На рисунке 9 проведен тест на линейность модели. Линии почти прямые, можно сказать, что линейность скорее выполняется.



*Рисунок 9 - Проверка на линейность второй модели*

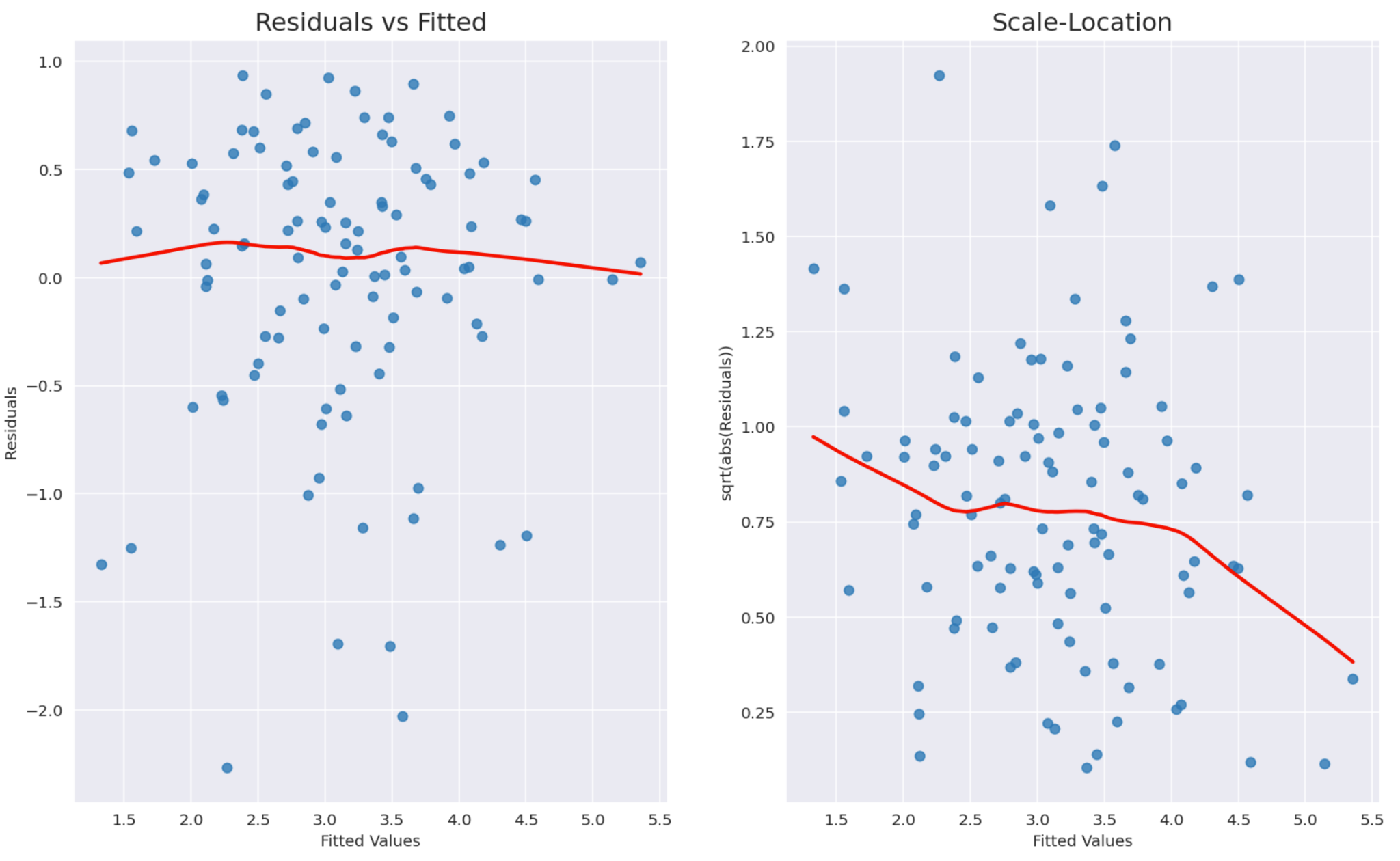
Проверка на равенство средних остатков нулю показала очень маленькое число, стремящееся к нулю (8.986401366629392e-15). Делаем вывод, что условие выполняется.

На рисунке 10 можно видеть таблицу, показывающее значение vif. При отсутствии мультиколлинеарности значения должны быть около 1, в нашем случае это условие не выполняется.



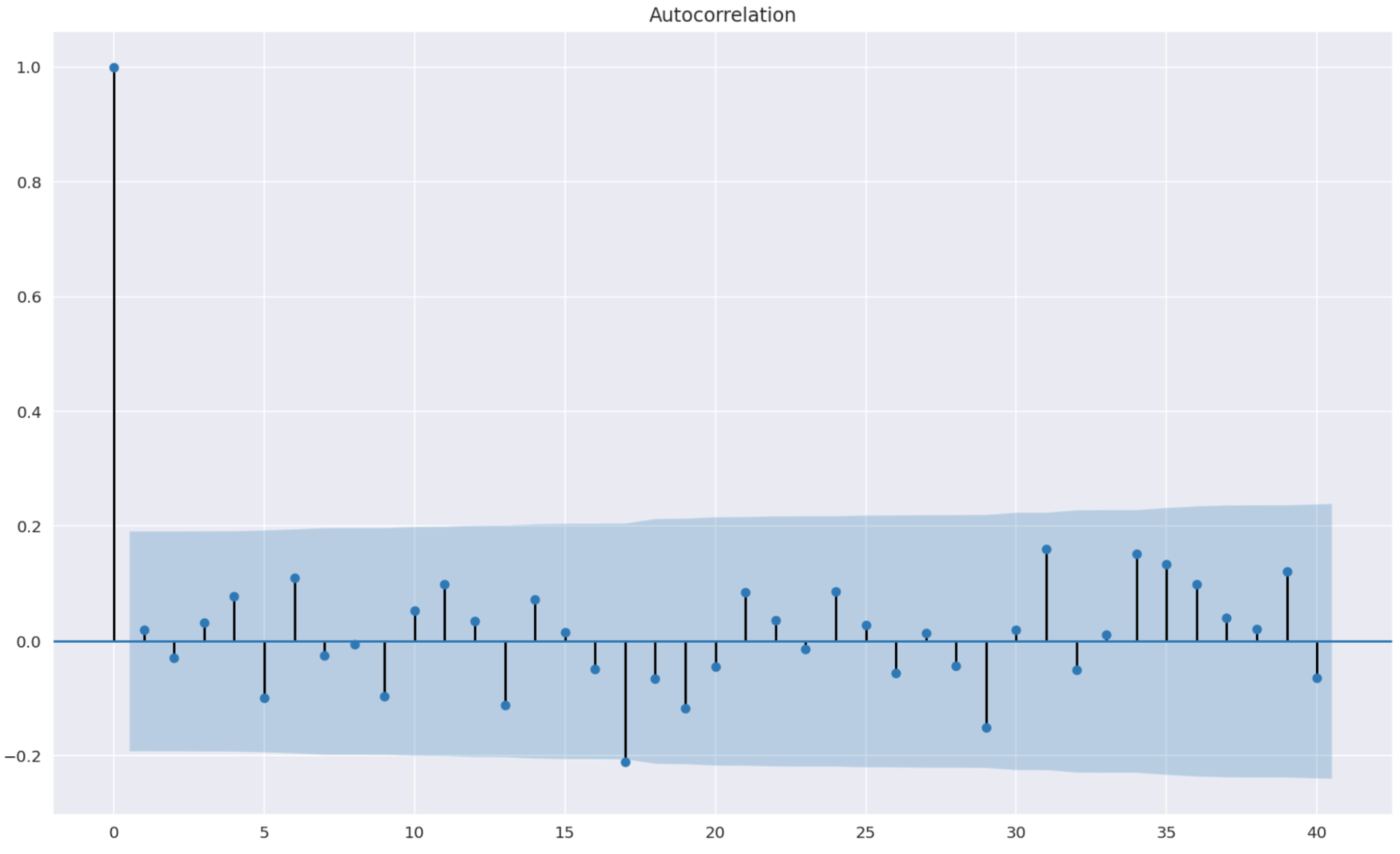
*Рисунок 10 - Проверка на отсутствие мультиколлинеарности второй модели*

На рисунке 11 проведен тест на гомоскедастичность остатков. Линии почти прямые, имеем пограничный случай. Для большей точности были проведены тесты Бройша-Пагана и Гольдфельда-Квандта. Первый показал гомоскедастичность (показатель p-value 0.287585), а второй - гетероскедастичность (показатель p-value 0.018616). Также имеем пограничный случай.



*Рисунок 11 - Проверка на гомоскедастичность остатков второй модели*

На рисунке 12 проведен тест на отсутствие автокорреляции остатков. Как можно видеть, первое значение выходит за рамки установленного доверительного интервала, а остальные находятся в его пределах. Можно сказать, что автокорреляции нет. Для дополнительной проверки был проведен тест Дарбина-Ватсона. Он показал значение примерно 1.95, что значит действительно, автокорреляции нет.



*Рисунок 12 - Проверка на отсутствие автокорреляции остатков второй модели*

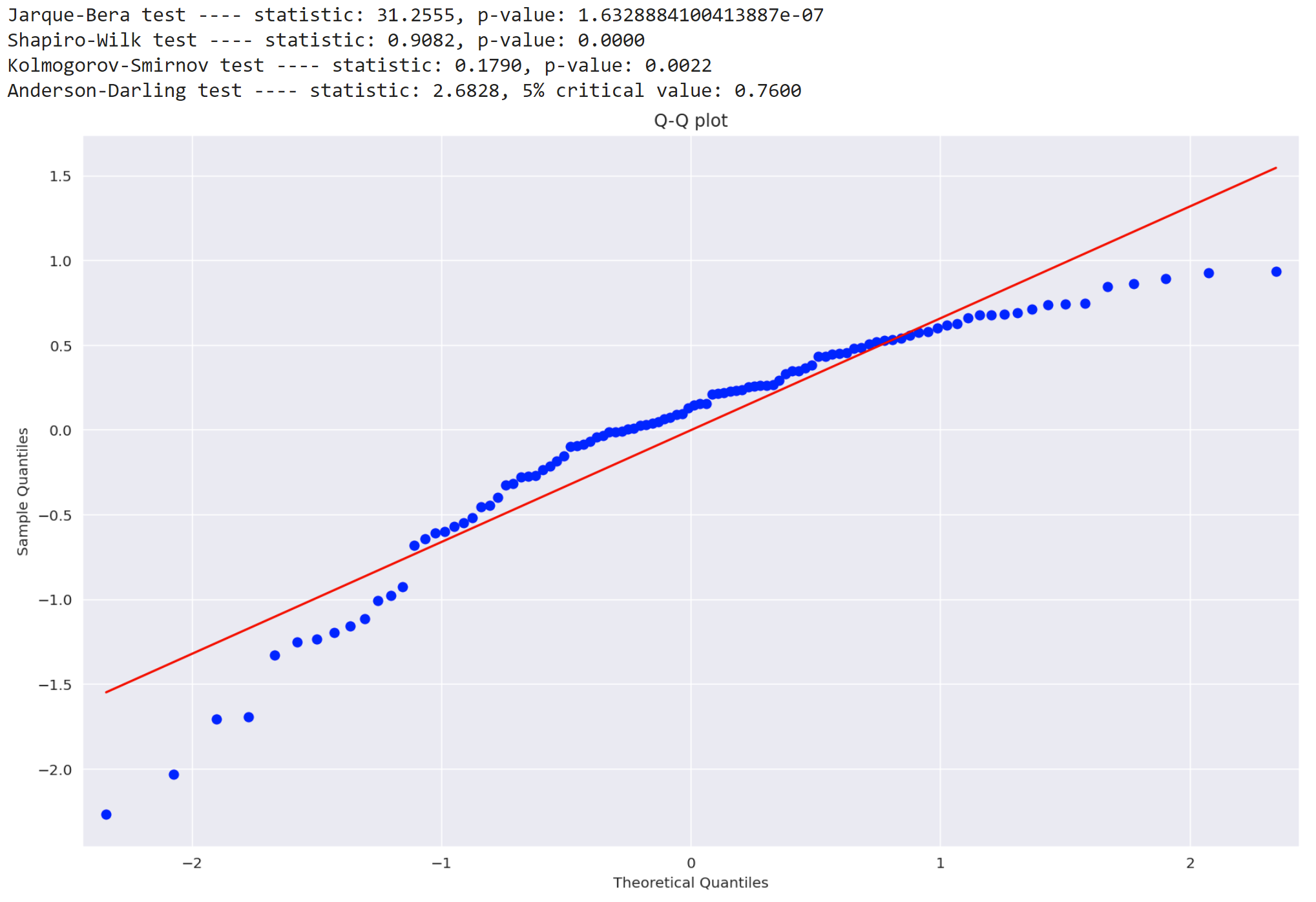
Результаты проверки предпосылок теоремы Гаусса-Маркова для трех моделей представлена в таблице 2.

*Таблица 2 - Проверка предпосылок Гаусса-Маркова*

| **Условие/Модель** | **Модель 1** | **Модель 2** | **Модель 3** |
| --- | --- | --- | --- |
| **Линейность модели** | Не выполняется | Скорее выполняется | Скорее выполняется |
| **Равенство средних остатков нулю** | Выполняется | Выполняется | Выполняется |
| **Отсутствие мультиколлинеарности** | Не выполняется | Не выполняется | Не выполняется |
| **Гомоскедастичность остатков** | Не выполняется | Пограничный случай | Пограничный случай |
| **Отсутствие автокорреляции остатков** | Скорее выполняется | Скорее выполняется | Скорее выполняется |

Более подробные пояснения о получении данных результатов, а также графики и отчеты можно увидеть в приложенном коде работы.

Для проверки гипотезы о нормальности случайной ошибки построим для каждой модели график Q-Q plot, а также проведем тесты на нормальность Жака-Бера, Шапиро-Уилка, Колмогорова-Смирнова, Андерсона-Дарлинга. Пример проверки на нормальность второй модели показан на рисунке 13.



*Рисунок 13 - Проверка на нормальность второй модели*

Результаты проверки нормальности всех моделей представлены в таблице 3.

*Таблица 3 - Проверка гипотезы о нормальности случайной ошибки*

| **Модель** | **Модель 1** | **Модель 2** | **Модель 3** |
| --- | --- | --- | --- |
| **Тест** | **Гипотеза о нормальности** | | |
| **Жака-Бера** | Отвергается | Отвергается | Отвергается |
| **Шапиро-Уилка** | Отвергается | Отвергается | Отвергается |
| **Колмогорова-Смирнова** | Отвергается | Отвергается | Отвергается |
| **Андерсона-Дарлинга** | Отвергается | Отвергается | Отвергается |
| **Q-Q plot** | Отвергается | Отвергается | Отвергается |

Более подробные пояснения о получении данных результатов, а также графики и отчеты можно увидеть в приложенном коде работы.

На основе всех проведенных тестов и анализа статьи был сделан вывод о том, что наилучшая модель из всех нами представленных - вторая. Она близка к модели из статьи, так как по ней мы предсказываем смертность исходя только из различных показателей страны, не учитывая статистику по COVID-19. Также, несмотря на отвергнутые гипотезы о нормальности, она показывает лучшие результаты среди остальных моделей.

1. **Содержательная интерпретация полученных выводов. Все коэффициенты должны быть проинтерпретированы в соответствии с их значимостью и функциональной формой модели.**

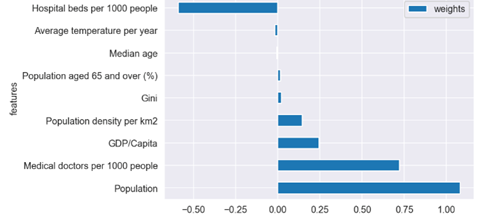
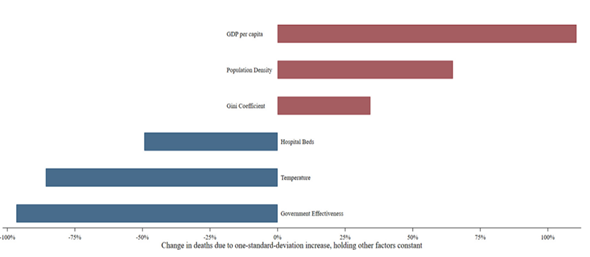
Интерпретация коэффициентов второй модели (оценка смертности по общим страновым признакам) в соответствии с их значимостью (расположены в порядке возрастания p-value):

* **Нулевой коэффициент** и **население** - наиболее значимые признаки (p-value = 0.000) очерчивают базовую смертность от последствий китайского коронавируса в стране, на которую в дальнейшем влияют различные факторы риска и защитные факторы.
* **Число врачей на 1000 человек** - значимый признак (p-value = 0.003), положительно коррелирующий с показателями развитости экономики и здравоохранения страны (например: ВВП на душу населения, затраты на здравоохранение и тд.). Как известно, эпидемия затронула развитые страны с развитой медициной в значительно большей степени чем развивающиеся и бедные, поэтому в данной модели число врачей выступает важным фактором риска, зависящим от разных переменных, как включенных, так и не включенных в модель.
* **Индекс Джини -** умеренно значимый признак (p-value = 0.026), свидетельствующий о большем риске смертности в странах с более неравным распределением доходов населения.
* **Число больничных коек на 1000 человек** - умеренно значимый признак (p-value = 0.072), являющийся одним из важнейших защитных факторов страны. Чем больше зданий и помещений выделено под места для ухода за тяжелобольными тем выше общее противостояние заражению и сокращение базового значения смертности в стране.
* **Среднее значение температуры в год** - несмотря на незначительность для полученный модели (p-value = 0.233), значение температуры является важным защитным фактором эпидемии, так как вирус распространяется медленнее в странах с более жарким климатом.
* **ВВП на душу населения** - несмотря на незначительность для полученной модели (p-value = 0.265), данный признак - важный фактор риска, так как он свидетельствует о рыночном производстве, потреблении, социальной деятельности и взаимодействию между людьми в той или иной стране, высокое значение такого показателя благоприятно влияет на распространение вируса и приводит к большей смертности.
* **Плотность населения на кв. км** - как известно, эпидемия затронула наиболее заселенные регионы и города мира (восточное побережье США, южная Италия, Англия и тд.), поэтому, несмотря на незначительность для модели (p-value = 0.296), плотность населения является важным фактором риска.
* **Доля населения старше 65 лет в %** - пожилые и люди старческого возраста наиболее уязвимы для китайского коронавирусу, но несмотря на это показатель их доли в населении страны является незначительным (p-value = 0.537) и маловажным фактором риска в исследовании заражения и смертности.
* **Медианный возраст** - наименее значимый (p-value = 0.870) и наименее влиятельный защитный фактор в исследовании, при удалении из модели значение r2, округленное до 3-х знаков после запятой, не изменится.

Для проверки функциональной формы модели был проведен тест Рамсея, который показал значение p-value при F-статистике равное примерно 0,64. Это значит, что линейная модель задана верно.

1. **Сравнение результатов проведенного исследования с результатами статьи.**

На рисунке 14 представлено сравнение значений коэффициентов моделей.



*Рисунок 14 - Сравнение коэффициентов моделей*

Авторы статьи сделали выводы, что при прочих равных, в стране с более высоким ВВП на душу населения, более высокой плотностью населения, более высоким коэффициентом Джини, меньшим количеством больничных коек, более низкой температурой и более низкой эффективностью правительства, как правило, больше смертей. Наше исследование показало, что страны с более высокой плотностью населения, более высоким ВВП на душу населения, большим неравенством доходов населения и более низкой температурой в регионе и меньшим количеством больничных коек, по прогнозам, будут более уязвимыми к глобальной пандемии. В частности, стоит отметить, что проблема занижения данных как о совокупных инфекциях, так и о совокупных смертях, повлияла на наши эмпирические результаты.

Авторы статьи определили, что для смертности ВВП на душу населения, плотность населения и неравенство доходов являются тремя наиболее важными факторами риска; эффективность правительства, температура и больничные койки являются тремя наиболее важными защитными факторами. Мы обнаружили, что ВВП на душу населения, плотность населения, коэффициент Джини и количество врачей являются тремя наиболее важными факторами риска, тогда как температура региона и количество больничных коек являются наиболее важными факторами защиты от последствий заражения COVID-19.