ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Московский институт электроники и математики им. А.Н. Тихонова

Шембель Даниил Альбертович, группа БИТ211

Модульная работа №1

по дисциплине «Прикладной Статистический анализ данных»

Дата сдачи отчета: 09.11.23

Москва 2023 г.

Оглавление

Важно

Чтобы запустить работу, нужно выбрать в качестве Kernel папку .venv внутри архива, там лежат все нужные библиотеки, в нужных версиях и конфигурациях. В других версиях библиотек возможны ошибки!

Актуальность

Регрессионный анализ экономических показателей с использованием указанных параметров имеет высокую актуальность в экономической науке и практике. Регрессионный анализ позволяет исследовать взаимосвязи между различными экономическими переменными и прогнозировать будущее экономическое развитие. Это критически важно для государственных институтов, банков, предприятий и инвесторов. Также регрессионный анализ позволяет изучать влияние различных факторов, таких как население, рост ВВП, инфляция, трудовая сила и инвестиции, на экономический рост и развитие. Если подойти к этому вопросу со стороны бизнеса или инвесторов, то и здесь регрессионный анализ может отлично помочь, при, например, анализе взаимосвязей между экономическими переменными и может помочь компаниям и инвесторам адаптировать свои бизнес-модели и стратегии к изменяющимся экологическим и экономическим условиям. Отойдя от одной страны, можно получить более статистически значимые данные о зависимостях, а также сравнить страны и регионы между собой, что полезно для исследователей, экономистов и международных организаций.

В целом, регрессионный анализ экономических показателей помогает нам лучше понимать экономические явления, делать прогнозы, принимать обоснованные решения и оценивать эффективность различных политик и программ в экономике. Это важный инструмент для исследователей, экономистов и принимающих решения в сфере экономики.

Исследуемые показатели

- Страна- показатель не будет участвовать в исследовании, однако присутствует в датасете для более удобного использования данного исследования в других, где страна может иметь значение.
- Ожидаемая продолжительность жизни при рождении (лет): Уровень ожидаемой жизни для новорожденных, выраженный в годах.
- Население, всего: Общее число людей, проживающих на определенной территории или стране.
- Прирост населения (годовой %): Процентный прирост населения за год, который определяется разницей между числом рожденных и числом умерших, а также миграцией.
- Миграция, всего: Общее число людей, перемещающихся внутри страны или между различными странами.
- Индекс развития человеческого потенциала (HCI) (шкала 0–1): Мера оценки развития человеческого потенциала в стране, учитывающая факторы, такие как образование, здравоохранение, производительность труда и другие социально-экономические показатели.
- ВВП на душу населения (доллары США): Общая величина экономической продукции, деленная на число людей в населении страны, выраженная в долларах США.
- Рост ВВП (годовой %): Процентное изменение ВВП страны за год, отражающее общую экономическую активность и производительность страны.

- Безработица, всего (% от общей численности рабочей силы) (смоделированная оценка МОТ): Процентное соотношение людей в трудоспособном возрасте, которые не имеют работы, к общему числу рабочей силы в стране.

- Инфляция, потребительские цены (годовые %): Процентное изменение среднего уровня цен на потребительские товары и услуги в стране за год.
- Выбросы CO2 (метрические тонны на душу населения): Общий объем выбросов углекислого газа CO2, производимый определенной страной, деленный на количество людей в населении.
- Площадь леса (% площади суши): Доля площади суши, занятая лесными угодьями, выраженная в процентах.
- Умышленные убийства (на 100 000 человек): Количество убийств, совершенных с преднамеренным умыслом, на 100 000 человек в населении страны.
- Статистические показатели эффективности (SPI): общий балл (шкала 0–100): Обобщенный показатель эффективности, оценивающий различные аспекты развития и благосостояния в стране, выраженный в виде числового балла от 0 до 100.
- Доля мест, занимаемых женщинами в национальных парламентах (%): Процентное соотношение женщин, занимающих места в национальных парламентах, к общему количеству членов парламента.
- Прямые иностранные инвестиции, чистый приток (% ВВП): Процентное соотношение суммы прямых иностранных инвестиций к ВВП страны, отражающее объем иностранных инвестиций в страну.

Вступление

В современном мире изучение и анализ экономических показателей играют важную роль в исследованиях, направленных на понимание и прогнозирование развития государств и регионов. Один из наиболее существенных показателей, характеризующих экономическое благополучие, - это ВВП на душу населения. Для поиска и анализа зависимости между данными показателями и ВВП на душу населения, будет построена модель на основе методов регрессионного анализа. Цель данного исследования состоит в том, чтобы выявить статистически значимые факторы, которые оказывают влияние на уровень ВВП на душу населения. Это позволит лучше понять, какие экономические переменные и показатели играют ключевую роль в формировании экономического благополучия населения.

Исследование основывается на доступных статистических данных, включающих широкий спектр экономических, социальных и экологических показателей и ВВП на душу населения для различных стран.

Результаты данного исследования могут иметь важное практическое применение и быть полезными для экономических аналитиков, правительственных органов, бизнес-сектора и других заинтересованных сторон, помогая принимать более обоснованные экономические решения и разрабатывать эффективные стратегии развития.

Начальные предположения

- Ожидаемая продолжительность жизни при рождении (лет): Должно быть сильно положительное влияние, так как продолжительность жизни говорит об уровне общедоступного здравоохранения, следовательно, при наличии сильной экономики будет и сильная система здравоохранения и наоборот.
- Население, всего: население должно иметь очевидно отрицательное влияние на ВВП на душу населения, ведь это должна быть буквально обратная зависимость.
- Прирост населения (годовой %): Осмелюсь предположить, что тут должно наблюдаться отрицательное влияние, из-за глобального тренда развитых стран к отрицательному приросту населения- чем больше люди начинают задумываться о карьере и работе, тем сильнее становится экономика страны, тем сильнее падает уровень рождаемости в стране и, соответственно, прирост населения.
- Миграция, всего: Предполагается, что должна быть положительная зависимость, ведь чем больше людей приезжают в государство, тем большим бенефициаром выступает экономика данной страны.
- Индекс человеческого потенциала (HCI) (шкала 0–1): Индекс развития человеческого потенциала в большинстве своём показывает сколько в среднем лет тратится жителями страны на обучение. Вполне логичным мне кажется предположение о том, что чем лучше образовано и развито население, тем лучше будет развита экономика и ВВП на душу населения.
- Инфляция, потребительские цены (годовые %): Чем меньше инфляция, тем стабильнее экономика страны, а, следовательно, должно быть отрицательное влияние на ВВП на душу населения.
- Выбросы CO2 (метрические тонны на душу населения): По количеству выбросов CO2 можно судить об уровне развития промышленности в стране, чем лучше развита промышленность, тем лучше будет развита и экономика страны, поэтому предполагаемое влияние должно быть положительным.
- Умышленные убийства (на 100 000 человек): Чем лучше развита правовая система и образование в стране, тем меньше будет показатель преднамеренных убийств. Предположу, что чем что развитие экономики и правовой системы — это связанные понятия, поэтому должна наблюдаться обратная зависимость.
- Доля мест, занимаемых женщинами в национальных парламентах (%): Доля мест, занимаемых женщинами в национальных парламентах- один из способов судить о вовлеченности населения в трудовую деятельность, а, следовательно, и в экономику, поэтому я думаю результат влияния должен быть положительный.
- Прямые иностранные инвестиции, чистый приток (% ВВП): Чем больше инвестиций в экономику, тем больше и прирост, который она показывает, поэтому здесь должно быть сильное положительное влияние.

Первичный анализ

Так как данные изначально находятся в достаточно сыром виде, прежде всего они должны быть очищены от пустых значений, которых в датасете присутствует, достаточно много, а также отобрать только интересующие исследование параметры. В итоге изначальный массив данных был уменьшен до

143 стран по 14 интересующим нас показателям (в изначальной выборке было 217 стран по 24 различным признакам).

	Life expectancy at birth, (years)	Population, total	Population growth (annual %)	Migration, total	Human Capital Index (HCI) (scale 0-1)	GDP per capita (US$) (dependent variable)	GDP growth (annual %)	Unemployment, total (% of total labor force) (modeled ILO estimate)	Inflation, consumer prices (annual %)	CO2 emissions (metric tons per capita)	Forest area (% of land area)	Intentional homicides (per 100,000 people)	Statistical performance indicators (SPI): Overall score (scale 0-100)	Proportion of seats held by women in national parliaments (%)	Foreign direct investment, net inflows (% of GDP)
0	62.0	41128771.0	2.5	-183672	0.4	363.7	-20.7	11.7	2.3	0.2	1.9	4.0	49.8	27.0	0.1
1	76.0	2775634.0	-1.3	-10612	0.6	6802.8	4.8	11.8	6.7	1.5	28.8	2.0	75.4	36.0	7.6
2	76.0	44903225.0	1.6	-18797	0.5	4273.9	3.1	11.6	9.3	3.7	0.8	2.0	55.1	8.0	0.0
5	62.0	35588987.0	3.1	29089	0.4	2998.5	3.0	10.2	25.8	0.6	53.4	4.0	54.9	34.0	-5.8
8	72.0	2780469.0	-0.4	-12825	0.6	7014.2	12.6	12.6	8.6	2.4	11.5	2.0	82.2	36.0	5.1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
209	70.0	326740.0	2.4	-197	0.5	3010.3	1.8	2.1	2.3	0.4	36.3	0.0	40.7	2.0	4.3
211	74.0	98186856.0	0.7	-992	0.7	4163.5	8.0	1.9	3.2	3.7	46.7	2.0	66.0	30.0	4.3
214	64.0	33696614.0	2.1	-101468	0.4	676.9	0.8	13.6	8.1	0.3	1.0	6.0	36.8	0.0	-1.3
215	61.0	20017675.0	2.8	9015	0.4	1487.9	4.7	6.1	11.0	0.4	60.3	5.0	59.0	15.0	0.4
216	59.0	16320537.0	2.0	-25005	0.5	1267.0	3.4	7.9	104.7	0.5	45.1	6.0	61.7	31.0	0.6

143 rows × 15 columns

После первого этапа “очистки”, следует провести дальнейший анализ для того, чтобы уменьшить мультиколлинеарность в данных. Для этого рассчитаем таблицу корреляций признаков между собой:

	Life expectancy at birth, (years)	Population, total	Population growth (annual %)	Migration, total	Human Capital Index (HCI) (scale 0-1)	GDP growth (annual %)	Unemployment, total (% of total labor force) (modeled ILO estimate)	Inflation, consumer prices (annual %)	CO2 emissions (metric tons per capita)	Forest area (% of land area)	Intentional homicides (per 100,000 people)	Statistical performance indicators (SPI): Overall score (scale 0-100)	Proportion of seats held by women in national parliaments (%)	Foreign direct investment, net inflows (% of GDP)
Life expectancy at birth, (years)	1.000000	-0.092247	-0.275436	-0.213634	0.873021	0.054807	-0.273559	-0.133150	0.526529	0.040692	-0.346187	0.683958	0.217006	0.151296
Population, total	-0.092247	1.000000	0.041886	0.290279	-0.055118	-0.027720	-0.112759	0.016975	-0.013265	-0.017614	0.069227	0.113694	-0.064527	-0.165230
Population growth (annual %)	-0.275436	0.041886	1.000000	0.072510	-0.327914	0.270473	-0.094057	-0.055776	-0.166494	-0.077553	0.011618	-0.309751	0.030310	-0.002818
Migration, total	-0.213634	0.290279	0.072510	1.000000	-0.104655	-0.005421	-0.021333	0.023625	-0.062302	-0.026838	0.142831	-0.075563	-0.147283	-0.049656
Human Capital Index (HCI) (scale 0-1)	0.873021	-0.055118	-0.327914	-0.104655	1.000000	0.023942	-0.291490	-0.137302	0.526901	0.059792	-0.316395	0.799257	0.261598	0.177514
GDP growth (annual %)	0.054807	-0.027720	0.270473	-0.005421	0.023942	1.000000	0.002691	-0.165980	0.059168	0.126582	0.067965	-0.008665	0.050856	0.331844
Unemployment, total (% of total labor force) (modeled ILO estimate)	-0.273559	-0.112759	-0.094057	-0.021333	-0.291490	0.002691	1.000000	0.185656	-0.238880	-0.124109	0.350816	-0.254847	0.001095	0.040077
Inflation, consumer prices (annual %)	-0.133150	0.016975	-0.055776	0.023625	-0.137302	-0.165980	0.185656	1.000000	-0.132311	-0.085808	-0.060899	-0.110375	-0.120422	-0.083720
CO2 emissions (metric tons per capita)	0.526529	-0.013265	-0.166494	-0.062302	0.526901	0.059168	-0.238880	-0.132311	1.000000	-0.200486	-0.209314	0.313418	-0.036759	-0.033265
Forest area (% of land area)	0.040692	-0.017614	-0.077553	-0.026838	0.059792	0.126582	-0.124109	-0.085808	-0.200486	1.000000	0.160749	0.053560	0.034148	0.074005
Intentional homicides (per 100,000 people)	-0.346187	0.069227	0.011618	0.142831	-0.316395	0.067965	0.350816	-0.060899	-0.209314	0.160749	1.000000	-0.278611	0.003188	-0.008273
Statistical performance indicators (SPI): Overall score (scale 0-100)	0.683958	0.113694	-0.309751	-0.075563	0.799257	-0.008665	-0.254847	-0.110375	0.313418	0.053560	-0.278611	1.000000	0.369241	0.051644
Proportion of seats held by women in national parliaments (%)	0.217006	-0.064527	0.030310	-0.147283	0.261598	0.050856	0.001095	-0.120422	-0.036759	0.034148	0.003188	0.369241	1.000000	0.104698
Foreign direct investment, net inflows (% of GDP)	0.151296	-0.165230	-0.002818	-0.049656	0.177514	0.331844	0.040077	-0.083720	-0.033265	0.074005	-0.008273	0.051644	0.104698	1.000000

Также таблицу корреляций признаков и зависимой переменной:

Life expectancy at birth, (years)	0.689743
Population, total	-0.045267
Population growth (annual %)	-0.043346
Migration, total	-0.053219
Human Capital Index (HCI) (scale 0-1)	0.697425
GDP per capita (US$) (dependent variable)	1.000000
GDP growth (annual %)	0.083543
Unemployment, total (% of total labor force) (modeled ILO estimate)	-0.273214
Inflation, consumer prices (annual %)	-0.171565
CO2 emissions (metric tons per capita)	0.604699
Forest area (% of land area)	-0.065355
Intentional homicides (per 100,000 people)	-0.252983
Statistical performance indicators (SPI): Overall score (scale 0-100)	0.542524
Proportion of seats held by women in national parliaments (%)	0.262908
Foreign direct investment, net inflows (% of GDP)	0.057079

И коэффициент инфляции дисперсии (VIF):

	Variable	VIF
0	Life expectancy at birth, (years)	92.918242
1	Population, total	1.685526
2	Population growth (annual %)	1.704395
3	Migration, total	1.158107
4	Human Capital Index (HCI) (scale 0-1)	117.364661
5	GDP per capita (US$) (dependent variable)	3.653210
6	GDP growth (annual %)	1.697747
7	Unemployment, total (% of total labor force) (...	3.978568
8	Inflation, consumer prices (annual %)	1.586061
9	CO2 emissions (metric tons per capita)	3.615116
10	Forest area (% of land area)	3.703576
11	Intentional homicides (per 100,000 people)	1.922857
12	Statistical performance indicators (SPI): Over...	74.841312
13	Proportion of seats held by women in national ...	7.224595
14	Foreign direct investment, net inflows (% of GDP)	1.728611

По значениям VIF, а также по значениям из матрицы корреляций отчётливо видно, что некоторые признаки не должны находиться в дальнейшем исследовании из-за отчётливо видимой мультиколлиниарности, которую они вносят в данные, поэтому из дальнейшего исследования будут исключены следующие показатели: Статистические показатели эффективности (SPI): Общий балл (шкала 0–100)», «Индекс человеческого капитала (HCI) (шкала 0–1)», «Ожидаемая продолжительность жизни при рождении, (лет)». В итоге в дальнейшем будет проведена работа с датасетом из 143 стран по 11 признакам.

	Population, total	Population growth (annual %)	Migration, total	GDP per capita (US$) (dependent variable)	GDP growth (annual %)	Unemployment, total (% of total labor force) (modeled ILO estimate)	Inflation, consumer prices (annual %)	CO2 emissions (metric tons per capita)	Forest area (% of land area)	Intentional homicides (per 100,000 people)	Proportion of seats held by women in national parliaments (%)	Foreign direct investment, net inflows (% of GDP)
0	41128771.0	2.5	-183672	363.7	-20.7	11.7	2.3	0.2	1.9	4.0	27.0	0.1
1	2775634.0	-1.3	-10612	6802.8	4.8	11.8	6.7	1.5	28.8	2.0	36.0	7.6
2	44903225.0	1.6	-18797	4273.9	3.1	11.6	9.3	3.7	0.8	2.0	8.0	0.0
5	35588987.0	3.1	29089	2998.5	3.0	10.2	25.8	0.6	53.4	4.0	34.0	-5.8
8	2780469.0	-0.4	-12825	7014.2	12.6	12.6	8.6	2.4	11.5	2.0	36.0	5.1
...	...	...	...	...	...	...	...	...	...	...	...	...
209	326740.0	2.4	-197	3010.3	1.8	2.1	2.3	0.4	36.3	0.0	2.0	4.3
211	98186856.0	0.7	-992	4163.5	8.0	1.9	3.2	3.7	46.7	2.0	30.0	4.3
214	33696614.0	2.1	-101468	676.9	0.8	13.6	8.1	0.3	1.0	6.0	0.0	-1.3
215	20017675.0	2.8	9015	1487.9	4.7	6.1	11.0	0.4	60.3	5.0	15.0	0.4
216	16320537.0	2.0	-25005	1267.0	3.4	7.9	104.7	0.5	45.1	6.0	31.0	0.6

143 rows × 12 columns

Регрессионные модели

Для построения регрессионных моделей в дальнейшем будут использованы две библиотеки: scikit-learn и statsmodels. Также выборка будет разделена на тестовую и предназначенную для обучения модели, причем тестовая выборка будет выбираться случайным образом, но каждый раз будет составлять 20% от всей выборки. Базовая модель линейной регрессии показывает относительно плохой результат коэффициента детерминации R^2 = 0.2962076157961222.

```
X = data.drop('GDP per capita (US$) (dependent variable)', axis=1)
y = data['GDP per capita (US$) (dependent variable)']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
score = model.score(X_test, y_test)
score
```

✓ 0.1s

```
0.2962076157961222
```

Однако, это вполне предсказуемый результат ведь данные слишком различны по шкале и могут отличаться в сотни и тысячи раз. Поэтому вполне логичным следующим шагом является логарифмирование данных с целью привести их к вполне сравнимой между собой шкале, и построение новой модели на логарифмированных данных.

```
def replace_values(value):
    if value > 0:
        return np.log(value)
    elif value < 0:
        return -1 * np.log(abs(value))
    else:
        return 0

log_data = data.map(replace_values)
```

✓ 0.0s

```
X = log_data.drop('GDP per capita (US$) (dependent variable)', axis=1)
y = log_data['GDP per capita (US$) (dependent variable)']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model_log = LinearRegression()
model_log.fit(X_train, y_train)
score = model_log.score(X_test, y_test)
score
```

✓ 0.0s

```
0.8312296269301478
```

Модель на логарифмированных данных показывает результат сильно лучше и коэффициент детерминации R^2 = 0.8312296269301478, что является отличным результатом.

Нормализация данных с помощью методов StandartScaler и MinMaxScaler из библиотеки sklearn.preprocessing, к сожалению не дала ожидаемых результатов и обе модели на нормализованных данных показали сравнимые с базовой моделью результаты коэффициента детерминации R^2= 0.29620761581765964, для MinMaxScaler и R^2= 0.2962076158176601, для StandartScaler, отличия в результатах этих двух моделях объясняется различными методами нормализации: MinMaxScaler приводит значения к определенному диапазону, обычно от 0 до 1. Он вычитает минимальное значение из каждого элемента и затем делит на разность между максимальным и минимальным значениями в наборе данных. Формула выглядит следующим образом:

x_scaled = (x - x_min) / (x_max - x_min)

StandardScaler, с другой стороны, нормализует данные, вычитая среднее значение и деля на стандартное отклонение. Формула для StandardScaler выглядит следующим образом:

x_scaled = (x - mean) / std

Где:

x - исходное значение данных

x_scaled - нормализованное значение данных

x_min - минимальное значение в наборе данных

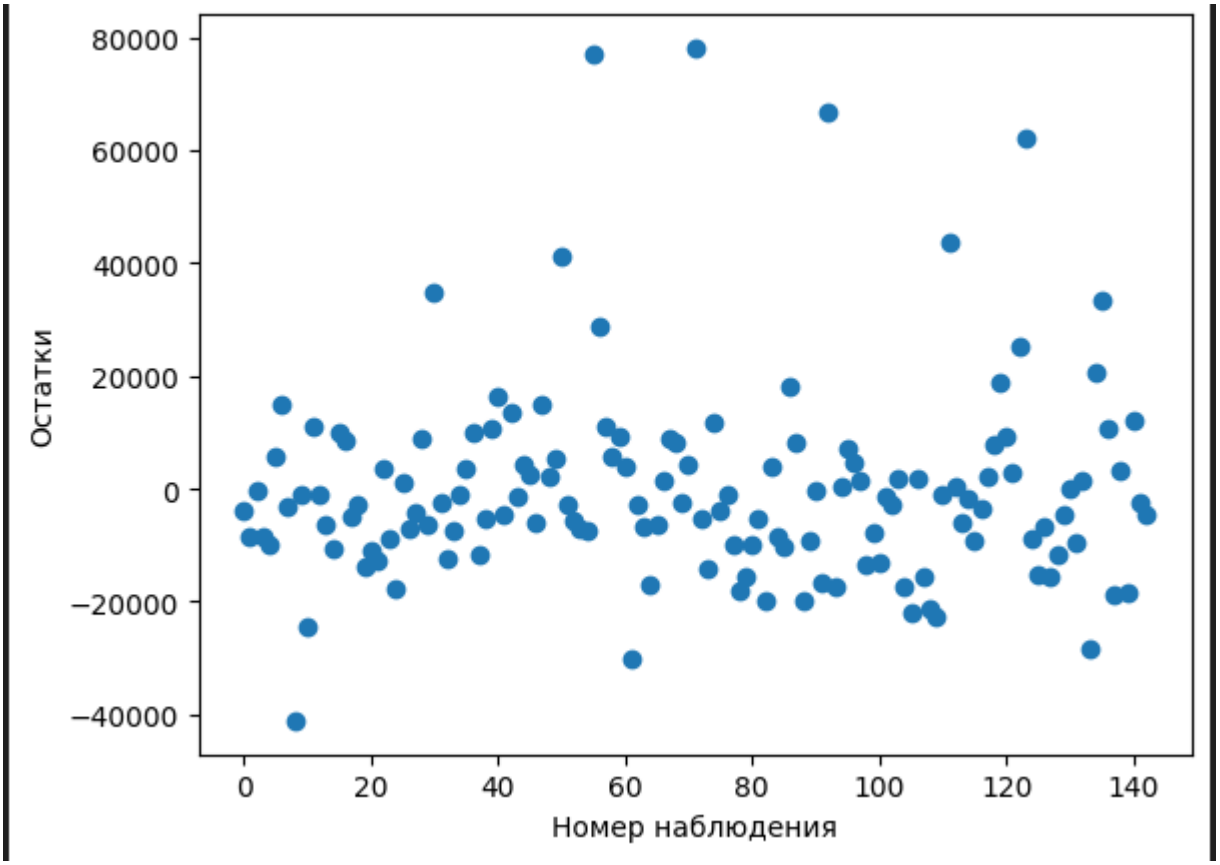
x_max - максимальное значение в наборе данных

mean - среднее значение в наборе данных

std - стандартное отклонение в наборе данных

Выявление гетероскедастичности в данных

Для первичного выявления гетероскедастичности в данных были построены графики остатков (общих и по каждому отдельному параметру)



По графикам очевидно присутствует гетероскедастичность, а также очевидно наличие выбросов в данных, поэтому было принято решение провести вторичную чистку данных с удалением некоторого количества выбросов, для этого в каждом столбце был посчитан iqr (межквартильный размах) и были определены границы (верхняя и нижняя), за пределами которых значение считалось выбросом, после проведения такой операции в каждом столбце, были выброшены из датасета все страны, которые имели бы выбросы более 2 раз. Полный алгоритм приведен ниже:

```python
def detect_outliers(column):
    q1 = column.quantile(0.25)
    q3 = column.quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - 1.5*iqr
    upper_bound = q3 + 1.5*iqr
    outliers = column[(column < lower_bound) | (column > upper_bound)]
    return outliers


outliers_count = {}

for column in data.columns:
    outliers = detect_outliers(data[column])
    if not outliers.empty:
        for index in outliers.index:
            country = data_to_display.loc[index, 'Country']
            if country in outliers_count:
                outliers_count[country] += 1
            else:
                outliers_count[country] = 1

countries_to_remove = []
print("Страны с максимальным количеством выбросов:")
sorted_outliers = sorted(outliers_count.items(), key=lambda x: x[1], reverse=True)
for country, count in sorted_outliers:
    if count>2:
        print(f"Страна: {country}, Количество выбросов: {count}")
        countries_to_remove.append(country)
```

✓ 0.0s

```
Страны с максимальным количеством выбросов:
Страна: united-states, Количество выбросов: 4
Страна: lebanon, Количество выбросов: 4
Страна: nigeria, Количество выбросов: 3
Страна: australia, Количество выбросов: 3
Страна: sri-lanka, Количество выбросов: 3
Страна: guyana, Количество выбросов: 3
Страна: south-sudan, Количество выбросов: 3
```

```python
indices = data_to_display[data_to_display['Country'].isin(countries_to_remove)].index
indices
```

✓ 0.0s

```
Index([10, 83, 110, 144, 178, 180, 206], dtype='int64')
```

```python
new_data = data.drop(index=indices)
new_data
```

✓ 0.0s

По итогам вторичной очистки были удалены 7 стран, после чего графики остатков были перестроены, а также заново были построены модели линейной регрессии (обычная и с логарифмированными данными), результат обычной модели немного улучшился, однако модель на логарифмированных данных ухудшила результат: новые коэффициенты детерминации R^2 получились равными 0.37420789762365736 и 0.7191064200210606, соответственно.

Статистические тесты на гетероскедастичность

Для определения наличия гетероскедастичности были также проведены два статистических теста: тест Бройша-Пагана и тест Уайта, уровень значимости был принят за 0,05 или 5%. Тесты были взяты из библиотеки statsmodels.stats.diagnostic. P-value получились равными 0.06113733229058492 0.004765459122166954, для теста Бройша-Пагана и Уайта соответственно. Тест Уайта показал однозначное наличие гетероскедастичности, что подтверждается графиками, однако тест Бройша-Пагана дал результат близкий к уровню значимости и по этому тесту нельзя точно утверждать о наличии гетероскедастичности. Далее были построены 2 новые модели с поправкой Уайта и были получены следующие результаты R^2: 0.6051708096411431 для обычной модели, и 0.9785872101878553 для модели на логарифмированных данных. Также для сравнения была построена модель на логарифмированных данных с добавлением константы, однако ее результат значительно хуже:

```
log_data = new_data.map(replace_values)
X = log_data.drop('GDP per capita (US$) (dependent variable)', axis=1)
X = sm.add_constant(X, prepend=True)
y = log_data['GDP per capita (US$) (dependent variable)']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = sm.OLS(y_train, X_train).fit(cov_type='HC3')
y_pred = model.predict(X_test)
rsquared = model.rsquared
print(rsquared)
coefficients = model.params
print(coefficients)
print(model.summary())
```

✓ 0.0s

```
0.8021891604143574
```

Регуляризация данных методами Ridge и Lasso

Также была выполнена регуляризация методами Ridge и Lasso. Регуляризация - это процесс добавления дополнительных членов в функционал цели модели для управления переобучением. Методы Ridge и Lasso являются распространенными способами регуляризации в линейной регрессии.

Метод Ridge добавляет регуляризационный член (сумму квадратов весов модели) к функционалу цели модели линейной регрессии. Этот член штрафует большие значения весов и помогает предотвратить переобучение. Ridge регуляризация стремится минимизировать сумму квадратов ошибок и сумму квадратов весов. Параметр регуляризации (алфа) регулирует влияние регуляризационного члена на модель. Метод Lasso также добавляет регуляризационный член к функционалу цели модели линейной регрессии, но использует сумму абсолютных значений весов вместо суммы квадратов. Lasso регуляризация имеет свойство "отбора признаков", что означает, что она может автоматически установить веса для некоторых признаков на ноль, делая их несущественными для модели. Это позволяет производить отбор наиболее информативных признаков и упрощать модель. Параметр регуляризации (алфа) также регулирует влияние регуляризационного члена на модель. Коэффициенты полученные данными методами: 0.8164668827435873 методом Lasso и 0.8346480757422984 методом Ridge.

Выбор оптимальной модели

Лучшей моделью стала, модель с поправкой Уайта для логарифмированных данных, которая показала коэффициент детерминации ≈0.98, что является прекрасным результатом, ниже приведено полное описание модели (в том числе коэффициенты регрессии):

```
Dep. Variable:     GDP per capita (US$) (dependent variable)   R-squared (uncentered):                   0.979
Model:                                                 OLS   Adj. R-squared (uncentered):              0.977
Method:                                      Least Squares   F-statistic:                              824.7
Date:                                     Wed, 08 Nov 2023   Prob (F-statistic):                    9.30e-92
Time:                                             20:05:22   Log-Likelihood:                         -194.52
No. Observations:                                      114   AIC:                                      405.0
Df Residuals:                                          106   BIC:                                      426.9
Df Model:                                                8
Covariance Type:                                       HC3
=============================================================================================================
                                                coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Population, total                             0.4687      0.031     14.940      0.000       0.407       0.530
Population growth (annual %)                  0.2054      0.142      1.449      0.147      -0.072       0.483
Migration, total                              0.0097      0.018      0.532      0.594      -0.026       0.045
GDP growth (annual %)                         0.0316      0.141      0.224      0.823      -0.245       0.308
Inflation, consumer prices (annual %)         0.0568      0.152      0.374      0.708      -0.241       0.354
CO2 emissions (metric tons per capita)        1.0716      0.183      5.852      0.000       0.713       1.430
Intentional homicides (per 100,000 people)    0.0451      0.121      0.373      0.709      -0.192       0.282
Foreign direct investment, net inflows (% of GDP)  0.2820  0.126      2.244      0.025       0.036       0.528
==============================================================================
Omnibus:                       45.870   Durbin-Watson:                   2.061
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              168.064
Skew:                           1.356   Prob(JB):                    3.20e-37
Kurtosis:                       8.294   Cond. No.                        23.1
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors are heteroscedasticity robust (HC3)
```

Однако эта модель не будет являться оптимальной из-за странных полученных коэффициентов зависимости. Очевидно, что ВВП на душу населения просто не может иметь сильно положительный коэффициент связи с населением, поэтому оптимальной будет модель с регуляризацией методом Ridge:

```
R^2:0.8346480757422984
Population, total: -0.048514024831435146
Population growth (annual %): 0.018714968575976446
Migration, total: 0.039358931711342596
GDP growth (annual %): 0.032329714798717074
Inflation, consumer prices (annual %): -0.09431731470019043
CO2 emissions (metric tons per capita): 0.7516787920806346
Intentional homicides (per 100,000 people): -0.10875176189053819
Foreign direct investment, net inflows (% of GDP): 0.10863780697978576

Mean Squared Error (MSE): 0.43313592436961973
Root Mean Squared Error (RMSE): 0.6581306286518047
Mean Absolute Error (MAE): 0.5287492408066616
```

Результаты данной модели вполне могут быть объяснены с экономической точки зрения- самое сильное влияние на ВВП на душу населения имеют выбросы CO2, что говорит об уровне развития производств в стране. Сейчас большая часть мирового производства завязана на использовании переработки топливных продуктов, которые и влияют на выбросы CO2. Влияние иностранных инвестиций также прямо влияет на ВВП на душу населения, ведь чем больше инвестиций поступает в экономику, тем больше и быстрее она растёт. Умеренное отрицательное влияние количества умышленных убийств говорит, о том, что в более экономически успешных странах, лучше развита и правовая система, пресекающая использование первобытного правила кто сильнее, тот и прав,

откуда и низкий показатель количества умышленных убийств. Ситуация с инфляцией, скорее аналогична ситуации с иностранными инвестициями в экономику страны, экономически сильные страны, отличаются и стабильностью экономических систем, отсюда и умеренно отрицательное влияние.