**Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет**

**«Высшая школа экономики»**

ОТЧЁТ ПО МОДУЛЬНОЙ РАБОТЕ

На тему: “Моделирование статистических зависимостей”

Выполнил студент

Достовалов Максим Юрьевич

Преподаватель:

Балычева Юлия Евгеньевна

Нижний Новгород, 2023

Оглавление

[1. Введение 3](#_Toc150464035)

[2. Описание признаков 4](#_Toc150464036)

[3. Первичный анализ 5](#_Toc150464037)

[3.1. Описательные статистики 5](#_Toc150464038)

[3.2. Категориальные признаки 6](#_Toc150464039)

[3.3. Числовые признаки 6](#_Toc150464040)

[4. Предварительная обработка 9](#_Toc150464041)

[4.1. Разделение на тестовую и тренировочную выборки 9](#_Toc150464042)

[4.2. Кодирование категориальных признаков 9](#_Toc150464043)

[4.3. Масштабирование 10](#_Toc150464044)

[4.4. Удаление выбросов 10](#_Toc150464045)

[5. Корреляционный анализ 11](#_Toc150464046)

[5.1. Корреляции признаков между собой 11](#_Toc150464047)

[5.2. Корреляция целевой переменной с предикторами 12](#_Toc150464048)

[6. Анализ вздутия дисперсий 13](#_Toc150464049)

[7. Классическая модель линейной регрессии с мультиколлинеарностью 15](#_Toc150464050)

[8. Классическая линейная регрессия 21](#_Toc150464051)

[9. Линейная регрессия с регуляризацией 23](#_Toc150464052)

[9.1. Результаты моделирования 24](#_Toc150464053)

[10. Регрессия с полиномиальными признаками 25](#_Toc150464054)

[11. Итоговая модель 26](#_Toc150464055)

[11.1. Модель без удаления “незначимых” признаков 26](#_Toc150464056)

[11.2. Модель с удалёнными незначимыми признаками 28](#_Toc150464057)

[11.3. Интерпретация результатов 31](#_Toc150464058)

# Введение

В рамках исследования были использованы выборочные демографические данные с целью определения факторов, влияющих на продолжительность жизни населения в различных регионах. Источником данных послужила платформа “Kaggle”, которая является одним из крупнейших репозиториев данных, предоставляющих разнообразные данные для аналитических и исследовательских проектов. Ссылка на источник данных: [Life Expectancy (WHO) Fixed (kaggle.com)](https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated).

Регрессионный анализ — это статистический метод для исследования связи между зависимой переменной (в данном случае средняя ожидаемая продолжительность жизни) и одной или несколькими независимыми переменными (факторы, которые могут влиять на продолжительность жизни). Этот метод может быть полезен для исследования продолжительности жизни в разных регионах по следующим причинам:

* Определение влияющих факторов:

С помощью регрессионного анализа можно определить, какие конкретные переменные (например, уровень дохода, доступность медицинской помощи, уровень образования, экологические условия и т.д.) влияют на продолжительность жизни.

* Оценка взаимосвязей:

Регрессионный анализ позволяет оценить силу и направление взаимосвязи между зависимой переменной и каждым из предикторов. Это помогает понять, как изменение одного фактора может повлиять на продолжительность жизни.

* Прогнозирование:

Модель регрессии может быть использована для прогнозирования средней ожидаемой продолжительности жизни в регионе на основе значений независимых переменных.

Цель исследования заключается в разработке предиктивной модели, которая позволит оценить среднюю ожидаемую продолжительность жизни населения. В рамках данного анализа особое внимание будет уделено исключению из рассмотрения тех переменных, которые имеют прямую корреляцию с показателями смертности. Это решение обусловлено стремлением исключить непосредственное влияние факторов смертности на прогностическую способность модели и сосредоточиться на обнаружении других, менее очевидных детерминант продолжительности жизни. Поэтому переменные 'Infant\_deaths', 'Under\_five\_deaths', 'Adult\_mortality' будут исключены из анализа. Также переменная "Economy\_status\_Developing", которая представлена в виде двоичного признака, была исключена из набора предикторов. Это решение было обосновано наличием комплементарной переменной "Economy\_status\_Developed". Учитывая, что данные переменные образуют полный набор взаимоисключающих категорий, наличие одной из них достаточно для сохранения целостности информации о статусе экономики страны. Таким образом, отсутствие значения в переменной "Economy\_status\_Developed" (то есть, его равенство нулю) непосредственно указывает на то, что экономика страны является развивающейся.

# Описание признаков

Исходный датасет включает в себя 21 признак и 2864 наблюдения в виде собранной информации за год по конкретной стране. Повторений или пропущенных значений в данных не наблюдается.

Название столбцов и их интерпретация:

* `*Country*`: Категориальный признак, обозначающий страну.
* `*Region*`: Категориальный признак, обозначающий регион.
* `*Year*`: Год, когда были собраны данные.
* `*Infant\_deaths*`: Количество смертей среди младенцев.
* `*Under\_five\_deaths*`: Количество смертей детей младше пяти лет.
* `*Adult\_mortality*`: Уровень смертности среди взрослого населения.
* `*Alcohol\_consumption*`: Уровень потребления алкоголя на душу населения.
* `*Hepatitis\_B*`: Процент населения, вакцинированного против гепатита В.
* `*Measles*`: Количество случаев заболевания корью.
* `*BMI*`: Средний индекс массы тела населения.
* `*Polio*`: Процент населения, вакцинированного против полиомиелита.
* `*Total\_expenditure*`: Общие расходы на здравоохранение как процент от ВВП.
* `*Diphtheria*`: Процент населения, вакцинированного против дифтерии.
* `*Incidents\_HIV*`: Количество новых случаев ВИЧ на 1000 неинфицированных населения.
* `*GDP\_per\_capita*`: ВВП на душу населения.
* `*Population\_mln*`: Население страны в миллионах.
* `*Thinness\_ten\_nineteen\_years*`: Процент детей и подростков от 10 до 19 лет, испытывающих дефицит веса.
* `*Thinness\_five\_nine\_years*`: Процент детей от 5 до 9 лет, испытывающих дефицит веса.
* `*Schooling*`: Среднее количество лет обучения населения.
* `*Economy\_status\_Developed*`: Дамми-переменная, указывающая на развитый статус экономики.
* `*Economy\_status\_Developing*`: Дамми-переменная, указывающая на развивающийся статус экономики.
* `*Life\_expectancy*`: Средняя ожидаемая продолжительность жизни населения.

# Первичный анализ

## Описательные статистики

Для категориальных признаков:

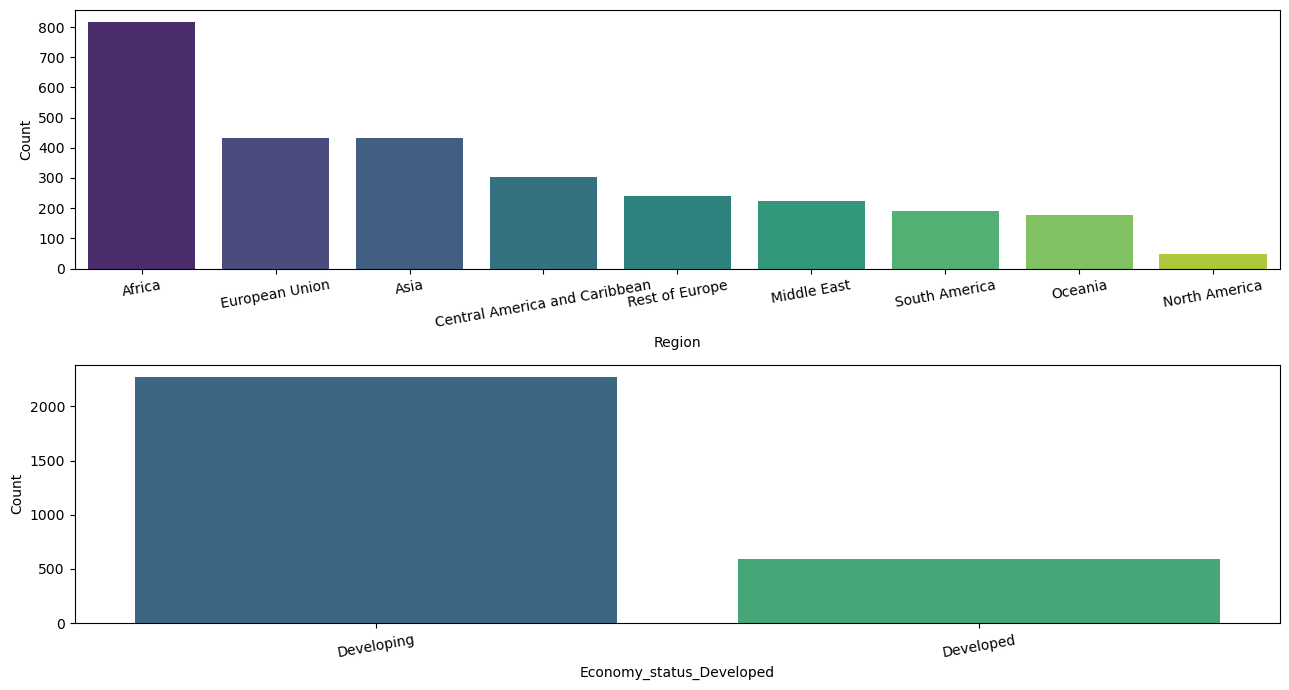
|  |  |  |  |
| --- | --- | --- | --- |
|  | Country | Region | Economy\_status\_Developed |
| Количество уникальных значений | 179 | 9 | 2 |
| Мода | Turkiye | Africa | Developing |
| Сколько раз мода встречается в данных | 16 | 816 | 2272 |

Для числовых признаков:

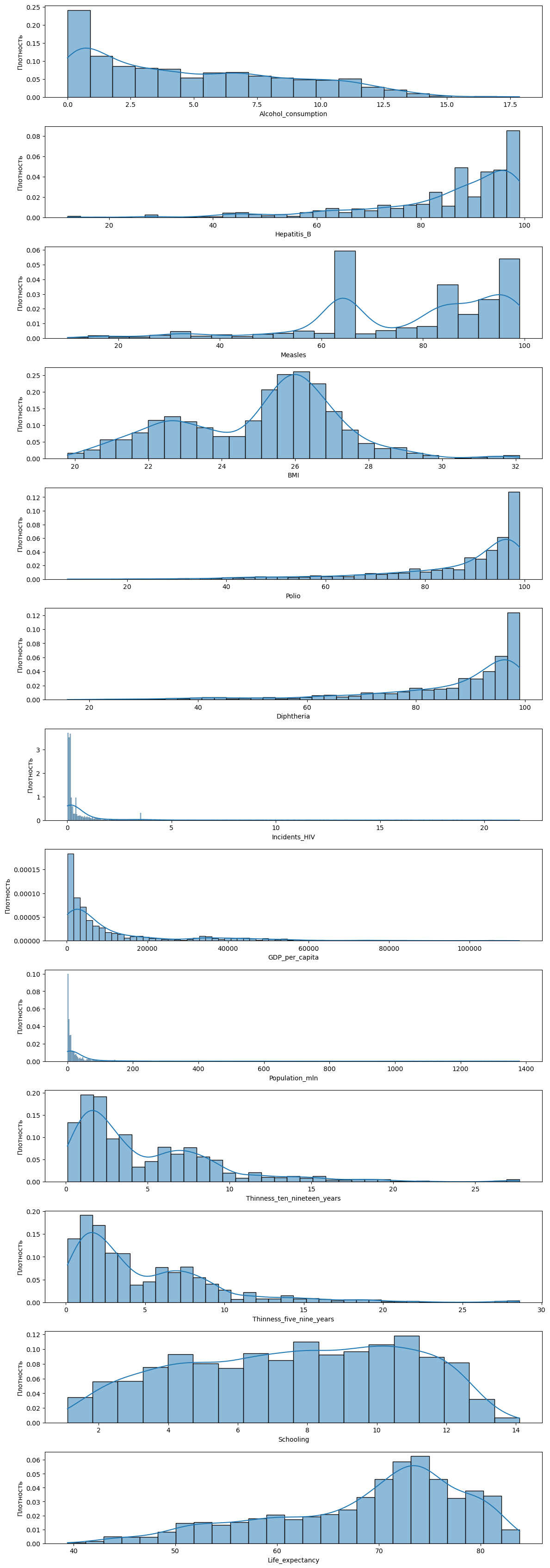
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Year | Alcohol\_consumption | Hepatitis\_B | Measles | BMI | Polio | Diphtheria |
| Среднее | 2007,5 | 4,820882 | 84,2926 | 77,34497 | 25,03293 | 86,49965 | 86,27165 |
| Стандартная ошибка | 4,610577 | 3,981949 | 15,99551 | 18,65969 | 2,193905 | 15,08036 | 15,53422 |
| Минимум | 2000 | 0 | 12 | 10 | 19,8 | 8 | 16 |
| 1 квартиль | 2003,75 | 1,2 | 78 | 64 | 23,2 | 81 | 81 |
| Медиана | 2007,5 | 4,02 | 89 | 83 | 25,5 | 93 | 93 |
| 3 квартиль | 2011,25 | 7,7775 | 96 | 93 | 26,4 | 97 | 97 |
| Максимум | 2015 | 17,87 | 99 | 99 | 32,1 | 99 | 99 |

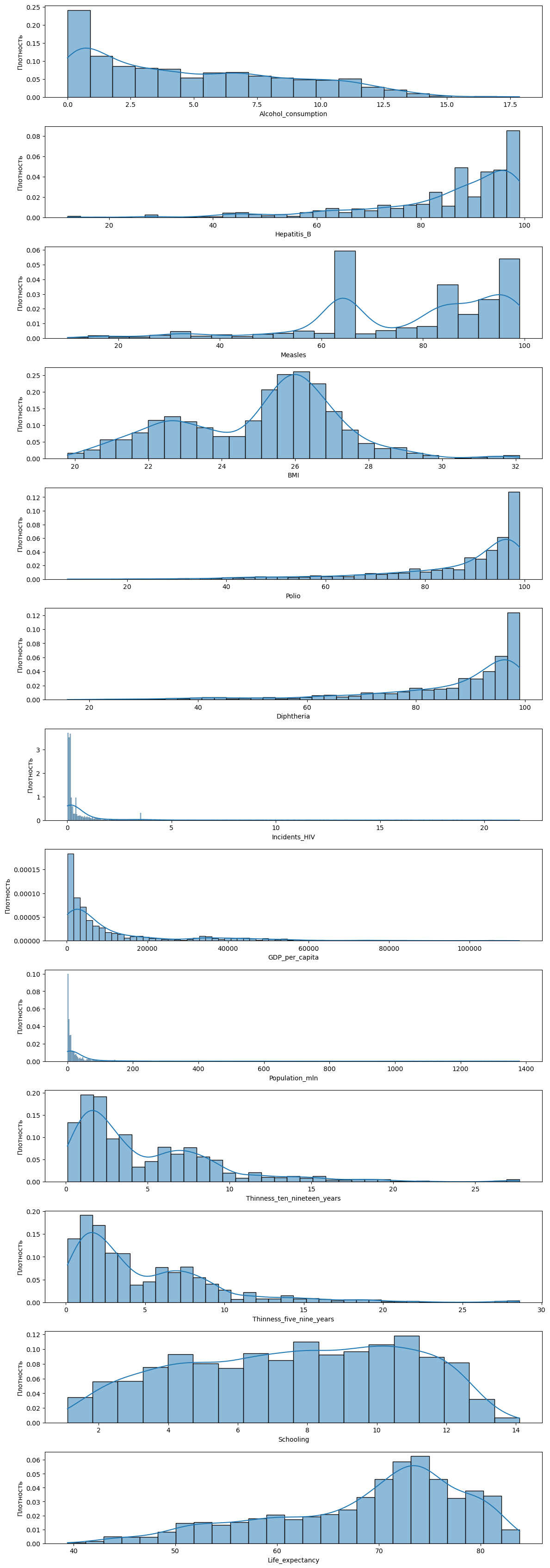
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Incidents\_HIV | GDP\_per\_capita | Population\_mln | Thinness\_ten\_nineteen\_years | Thinness\_five\_nine\_years | Schooling | Life\_expectancy |
| Среднее | 0,894288 | 11540,92 | 36,67591 | 4,865852 | 4,899825 | 7,632123 | 68,85608 |
| Стандартная ошибка | 2,381389 | 16934,79 | 136,4859 | 4,438234 | 4,525217 | 3,171556 | 9,405608 |
| Минимум | 0,01 | 148 | 0,08 | 0,1 | 0,1 | 1,1 | 39,4 |
| 1 квартиль | 0,08 | 1415,75 | 2,0975 | 1,6 | 1,6 | 5,1 | 62,7 |
| Медиана | 0,15 | 4217 | 7,85 | 3,3 | 3,4 | 7,8 | 71,4 |
| 3 квартиль | 0,46 | 12557 | 23,6875 | 7,2 | 7,3 | 10,3 | 75,4 |
| Максимум | 21,68 | 112418 | 1379,86 | 27,7 | 28,6 | 14,1 | 83,8 |

## Категориальные признаки



## Числовые признаки





На представленных диаграммах наблюдаются распределения исходных количественных характеристик. Основываясь на визуальном анализе, можно сделать вывод, что данные распределения существенно различаются и не демонстрируют свойств нормального распределения. Кроме того, отчетливо прослеживаются аномалии в виде экстремальных значений для большинства переменных, что необходимо принимать во внимание при последующем аналитическом исследовании. Эти выбросы могут существенно повлиять на оценку статистических моделей и их предсказательную способность. Следовательно, требуется тщательный анализ данных на предмет идентификации и обработки выбросов, что позволит улучшить качество моделирования и обеспечить более надежные результаты.

# Предварительная обработка

## Разделение на тестовую и тренировочную выборки

Для адекватной оценки производительности и обобщающей способности моделей, набор данных будет разделен на обучающую и тестовую выборки. Обучающая выборка будет использоваться для тренировки модели, в то время как тестовая выборка предназначена для ее валидации. Если модель демонстрирует высокую точность на обучающей выборке, но низкую на тестовой, это будет указывать на переобучение модели. Пропорция данных, выделенных для тестовой выборки, составляет 20% от общего объема датасета.

## Кодирование категориальных признаков

В ходе предварительной обработки данных, обнаружено наличие категориальных переменных, которые необходимо преобразовать в формат, приемлемый для дальнейшего моделирования. Метод преобразования "One-Hot-Encoding", хотя и является стандартным подходом для такой задачи, в данном случае был признан неоптимальным, учитывая его тенденцию к значительному увеличению размерности признакового пространства из-за генерации большого количества новых переменных. Это может привести к проблемам масштабируемости и ухудшению производительности моделей из-за "проклятия размерности".

Вместо этого был выбран метод "Label-Encoding", который устраняет указанную проблему, назначая уникальные целочисленные значения каждой категории внутри признака. Этот подход эффективно сокращает признаковое пространство, сохраняя при этом всю необходимую информацию для моделирования. Однако, стоит отметить, что "Label-Encoding" вносит предположение о наличии упорядоченной связи между категориями, что может не отражать реальную природу данных и искажать результаты моделирования. Таким образом, этот метод выбирается с учетом специфики данных и целей исследования, и принимается осознанное решение о его применении на основе анализа.

## Масштабирование

Для масштабирования признаков будет применяться процедура стандартизации, в ходе которой от каждого значения признака будет вычтено среднее значение этого признака в обучающей выборке, а результат деления будет нормирован на стандартное отклонение признака в обучающей выборке. Такой подход обеспечивает центрирование данных вокруг нуля и их масштабирование по стандартному отклонению. Для предотвращения утечки данных и переобучения, параметры стандартизации будут рассчитаны исключительно на основе обучающей выборки и затем применены как к обучающей, так и к тестовой выборке, обеспечивая таким образом сохранение общности обработки данных и корректность оценки производительности модели.

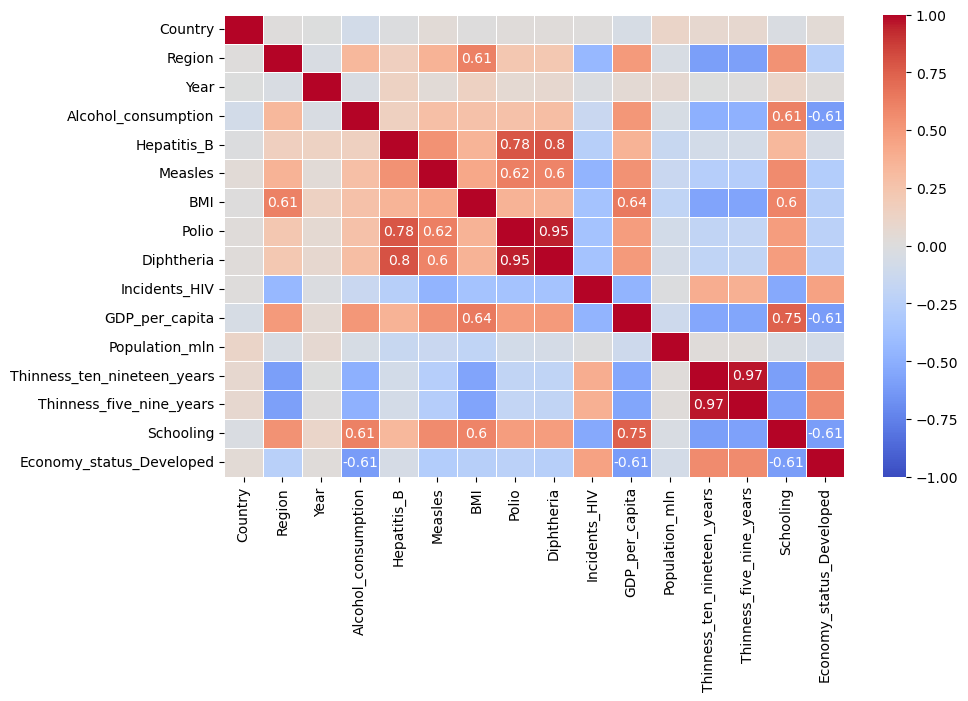
В связи с тем, что различные признаки могут быть измерены в несоответствующих единицах измерения, применяется масштабирование данных для обеспечения одинакового масштаба всех переменных. Это необходимо для корректной оценки и сопоставления вклада каждого признака в предсказание целевой переменной.

## Удаление выбросов

Чтобы минимизировать потенциальное влияние выбросов на оценку коэффициентов регрессионной модели и их стандартные ошибки, применяется метод исключения данных, превышающих три стандартных отклонения от среднего, в рамках обучающей выборки. Это позволяет снизить вероятность переобучения модели на аномальные значения. Однако в тестовой выборке выбросы сохраняются, чтобы можно было оценить реальную предсказательную способность модели и её поведение в условиях присутствия потенциальных аномалий в данных, что отражает возможные условия её практического применения.

# Корреляционный анализ

## Корреляции признаков между собой



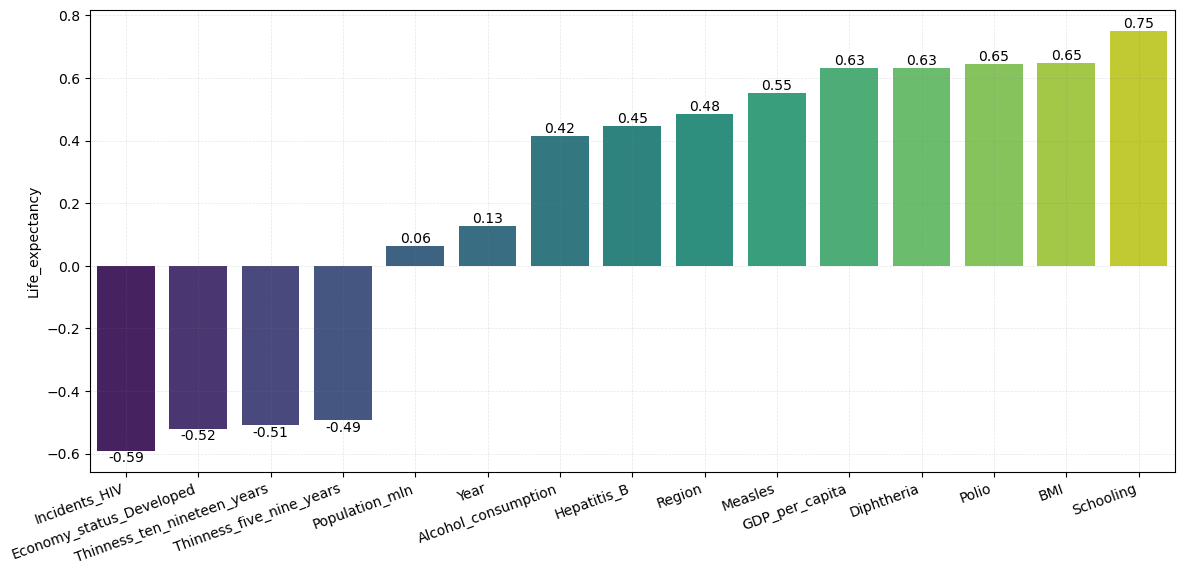
На представленном изображении иллюстрируется корреляционная матрица в формате тепловой карты, демонстрирующая степень взаимосвязей между разнообразными исследуемыми переменными. Каждая клетка этой матрицы содержит коэффициент корреляции Спирмана для соответствующих пар переменных. В контексте анализа, использование ранговой корреляции Спирмана оказывается предпочтительным, учитывая наличие кодированных категориальных признаков в наборе данных. Данный коэффициент корреляции изменяется в диапазоне от -1 до +1, где значение +1 указывает на абсолютную прямую зависимость, 0 отражает отсутствие взаимосвязи, а -1 обозначает абсолютную обратную зависимость между переменными. На карте числовые обозначения присвоены только тем корреляциям, величина которых по модулю превышает 0.6, поскольку более низкие значения не представляют интерес для данного анализа. Красный цвет на карте символизирует положительную корреляцию, в то время как синий цвет отражает отрицательную. Интенсивность окраски пропорциональна силе корреляционной связи.

Из значимых корреляций следует отметить:

* Присутствует высокая положительная корреляция между уровнями вакцинации от полиомиелита (Polio) и дифтерии (Diphtheria), достигающая значения в 0.95, что может свидетельствовать о том, что улучшение в области вакцинопрофилактики способствует совместному снижению распространенности этих заболеваний.
* Значения корреляции между показателями худобы детей в возрасте от пяти до девяти лет (Thinness\_five\_nine\_years) и от десяти до девятнадцати лет (Thinness\_ten\_nineteen\_years) составляют 0.97, что указывает на схожесть в распределении этих показателей истощения среди детей данных возрастных категорий.
* Образовательный уровень (Schooling) тесно связан с валовым внутренним продуктом на душу населения (GDP\_per\_capita) с коэффициентом корреляции 0.75, что может отражать наличие прямой связи между уровнем образования и экономическим благополучием нации.

Существуют сильные корреляции, что может говорить о мультиколлинеарности в данных.

## Корреляция целевой переменной с предикторами



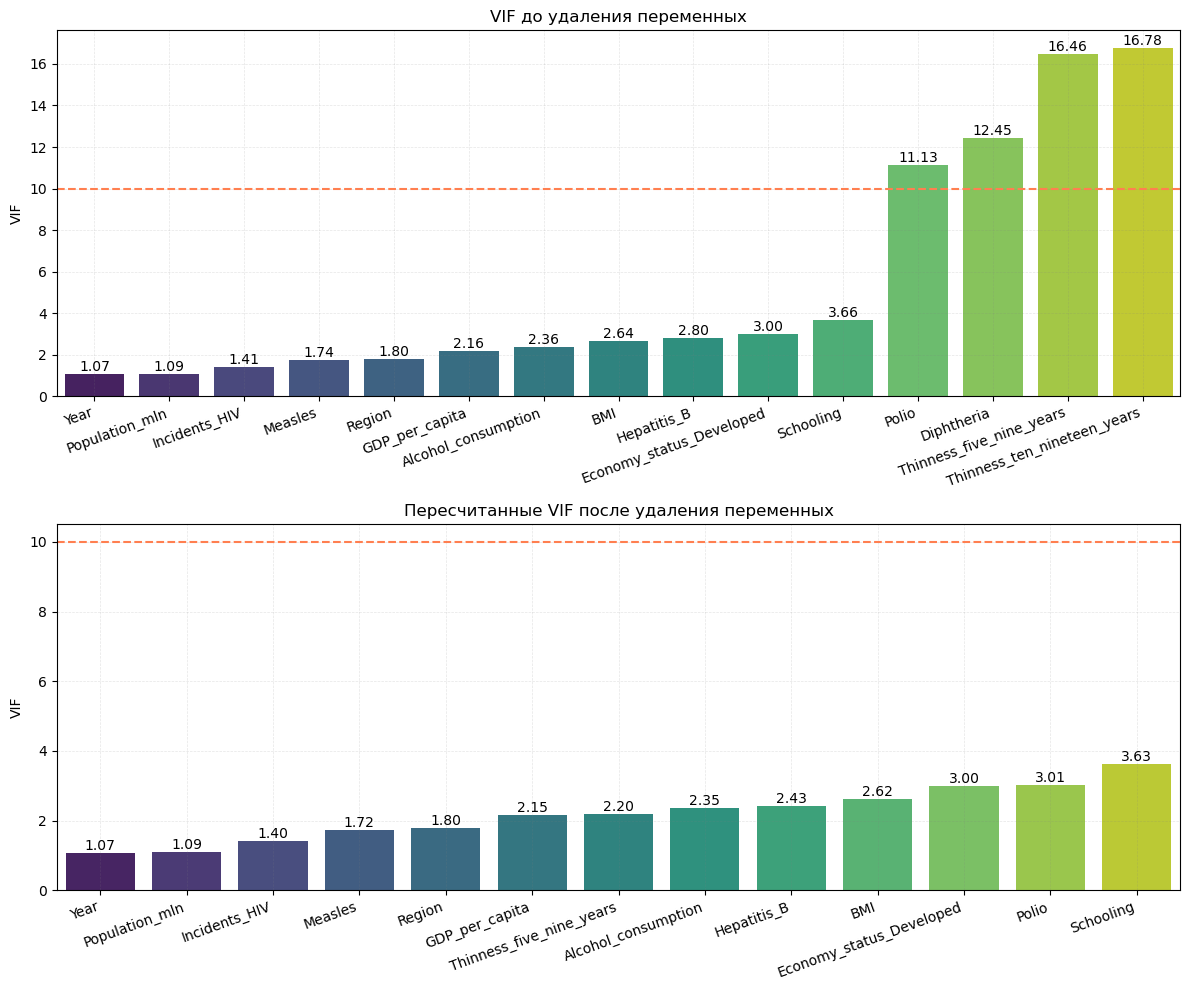
На предоставленном изображении изображена столбчатая диаграмма, которая показывает корреляцию между ожидаемой продолжительностью жизни (life-expectancy) и рядом предикторных переменных (предполагаемых факторов, влияющих на целевую переменную).

Вот краткое описание корреляций, представленных на диаграмме:

* Отрицательные корреляции (отображены фиолетовыми столбцами):
* Наибольшая отрицательная корреляция (-0.59) наблюдается между переменной "Incidents\_HIV" (частота случаев ВИЧ) и ожидаемой продолжительностью жизни, что может указывать на влияние распространенности ВИЧ на сокращение продолжительности жизни.
* Следующие три столбца показывают отрицательную связь между экономическим статусом страны ("Economy\_status\_Developed"), индексами истощения для двух возрастных групп ("Thinness\_ten\_nineteen\_years" и "Thinness\_five\_nine\_years") и ожидаемой продолжительностью жизни, с коэффициентами от -0.52 до -0.49.
* Положительные корреляции (отображены зелеными и желтым столбцами):
* Самая низкая из отображенных положительных корреляций — "Alcohol\_consumption" (употребление алкоголя) с коэффициентом 0.06.
* Следующие столбцы показывают более высокие положительные корреляции для "Hepatitis\_B", "Region", "Measles" (корь), "GDP\_per\_capita" (ВВП на душу населения), "Diphtheria" (дифтерия) и "Polio" (полиомиелит), с коэффициентами от 0.42 до 0.65.
* Наивысшая положительная корреляция (0.75) показана для "Schooling" (школьное образование), что предполагает сильную связь между уровнем образования населения и ожидаемой продолжительностью жизни.

Значения коэффициентов корреляции представлены на вершинах каждого столбца на диаграмме, что облегчает их интерпретацию.

# Анализ вздутия дисперсий



На представленных графиках визуализированы результаты оценки мультиколлинеарности с использованием индекса вздутия дисперсии каждой из независимых переменных в модели регрессии. Индекс VIF служит мерой степени увеличения дисперсии коэффициента оценки, вызванной коллинеарностью, и является индикатором мультиколлинеарности между предикторами. Традиционно, значение VIF, превышающее 10, рассматривается как указание на значительную мультиколлинеарность.

На верхнем графике показаны значения VIF до удаления переменных, где два предиктора "Diphtheria" и "Thinness\_ten\_nineteen\_years" демонстрируют особенно высокие значения VIF, 12.45 и 16.78 соответственно. Такие значения свидетельствуют о существенной мультиколлинеарности, что может отрицательно сказаться на статистической устойчивости оценок модели, увеличивая стандартные ошибки коэффициентов и снижая точность оценок.

После исключения этих переменных из модели, как видно на нижнем графике, значения VIF для оставшихся переменных были пересчитаны. Результаты показывают, что удаление вышеупомянутых переменных привело к снижению мультиколлинеарности, что видно по сокращению значений VIF для всех предикторов. Все оставшиеся значения VIF оказались ниже установленного порогового значения 10, что свидетельствует о приемлемой степени мультиколлинеарности.

Эти результаты можно интерпретировать так, что присутствие переменных "Diphtheria" и "Thinness\_ten\_nineteen\_years" вносило значительную мультиколлинеарность в модель. Их исключение улучшило оценочные характеристики регрессионной модели, повышая её надежность и уменьшая вероятность искажения выводов из-за коллинеарности.

# Классическая модель линейной регрессии с мультиколлинеарностью

В качестве начального этапа, предшествующего оптимизации модели, будет осуществлён тщательный анализ текущей спецификации регрессионной модели, включая все предикторы, даже те, которые проявляют признаки мультиколлинеарности. Целью данного анализа является детальная оценка общего качества и адекватности модели, а также анализ распределения остатков. Этот процесс включает в себя оценку соответствия модели на основе статистических критериев и проверку гомоскедастичности, нормальности распределения остатков и отсутствия автокорреляции. Особое внимание будет уделено идентификации любых необычных паттернов в остатках, которые могут указывать на потенциальные проблемы в спецификации модели, такие как неправильно включенные или исключенные переменные, неправильная функциональная форма или проблемы с выборочными данными.

Такой комплексный первоначальный анализ необходим для обеспечения того, чтобы любые дальнейшие шаги по модификации модели, включая устранение мультиколлинеарности, были основаны на прочном понимании исходной спецификации и её поведения. Это также поможет в избежании переобучения модели и обеспечит, что последующие изменения приведут к улучшению её предсказательной способности и интерпретируемости результатов.

Анализ качества модели:

1. R­2 =0.819

Значение коэффициента детерминации указывает на то, что модель хорошо подходит для данных: примерно 81.9% вариации зависимой переменной объясняется вариациями независимых переменных, включенных в модель. Это достаточно высокое значение. Оно предполагает, что модель обеспечивает значительную точность в предсказаниях.

1. Prob (F-statistic) = 0

Значение Prob (F-statistic) равное 0 означает, что гипотеза о том, что все коэффициенты регрессионной модели равны нулю (то есть модель в целом не имеет статистической значимости) отвергается на любом разумном уровне значимости. С практической точки зрения, это означает, что по крайней мере один из предикторов имеет статистически значимое влияние на зависимую переменную.

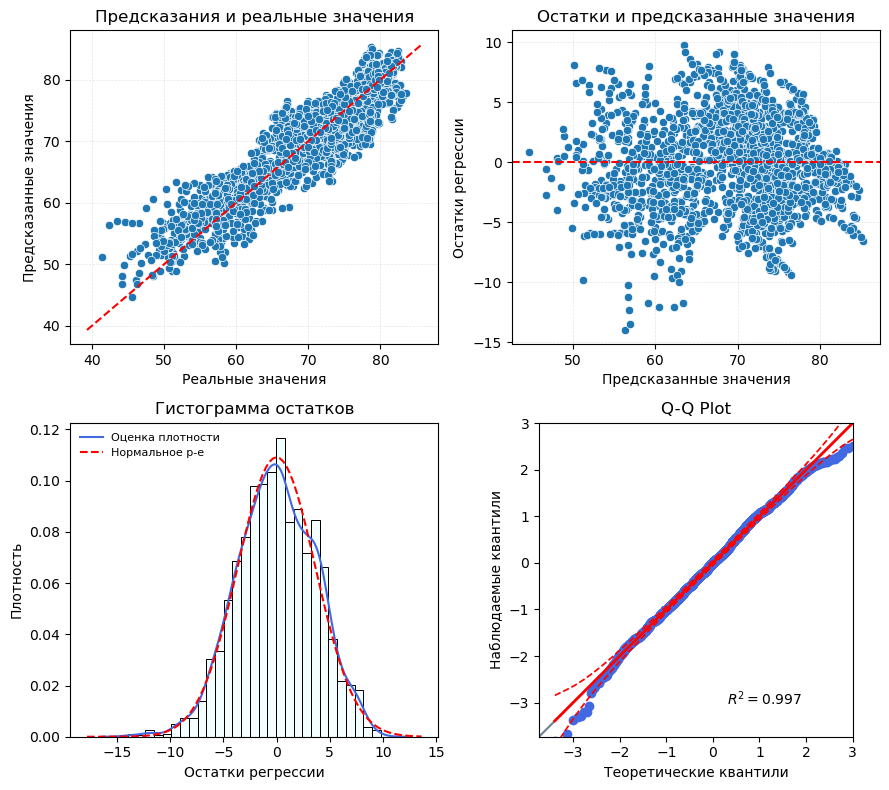
Был проведён ряд статистических тестов, в том числе:

* Тест Шапиро-Уилка:
* Этот тест применяется для проверки нулевой гипотезы (H0) о том, что остатки модели регрессии следуют нормальному распределению против альтернативной гипотезы (H1), согласно которой распределение остатков отличается от нормального.
* Тест Бреуша-Пагана и Тест Уайта:
* Оба теста направлены на диагностику гомоскедастичности остатков.
* Тест Бреуша-Пагана и Тест Уайта проверяют нулевую гипотезу (H0) о наличии гомоскедастичности, что подразумевает постоянство дисперсии остатков регрессионной модели.
* Альтернативная гипотеза (H1) для этих тестов предполагает присутствие гетероскедастичности, т.е. переменности дисперсии остатков в зависимости от уровня предсказываемой переменной или величины остатков.

В рамках исследования был установлен стандартный порог статистической значимости на уровне 0.05, что соответствует 5%. Это означает, что для того, чтобы отклонить нулевую гипотезу в любом из проведённых статистических тестов, рассчитанное p-значение должно быть меньше 0.05.

В случае подтверждения альтернативных гипотез каждого из тестов, исследователь может столкнуться с необходимостью корректировки модели регрессии с целью устранения обнаруженных нарушений предположений классической линейной регрессии (согласно теореме Гаусса-Маркова оценки будут эффективными в классе линейный несмещённых оценок).

|  |  |  |  |
| --- | --- | --- | --- |
|  | P-values | Отклоняем H0 | Вывод |
| Тест Шапиро-Уилка | 1.682600e-04 | True | Остатки не распределены нормально |
| Тест Бреуша-Пагана | 6.954985e-23 | True | Присутствует гетероскедастичность |
| Тест Уайта | 4.123269e-71 | True | Присутствует гетероскедастичность |



Эти четыре графика представляют собой типичные диагностические инструменты, используемые для оценки качества и адекватности линейной регрессионной модели.

1. График предсказаний и реальных значений (Predicted vs. Actual Values):

* На графике представлены реальные значения по сравнению с предсказанными моделью значениями.
* Красная пунктирная линия представляет собой идеальную подгонку, где предсказанные значения равны реальным.
* Точки в основном расположены вдоль этой линии, что указывает на то, что модель обладает хорошей предсказательной способностью.

1. График остатков и предсказанных значений (Residuals vs. Predicted Values):

* На этом графике остатки (разница между реальными и предсказанными значениями) показаны относительно предсказанных значений.
* Идеальный случай — случайные, непаттернированные остатки вокруг горизонтальной оси (красная пунктирная линия).
* В этом случае остатки кажутся довольно случайно распределенными вокруг нуля, хотя при более высоких предсказанных значениях заметно некоторое увеличение дисперсии остатков, что может указывать на гетероскедастичность.

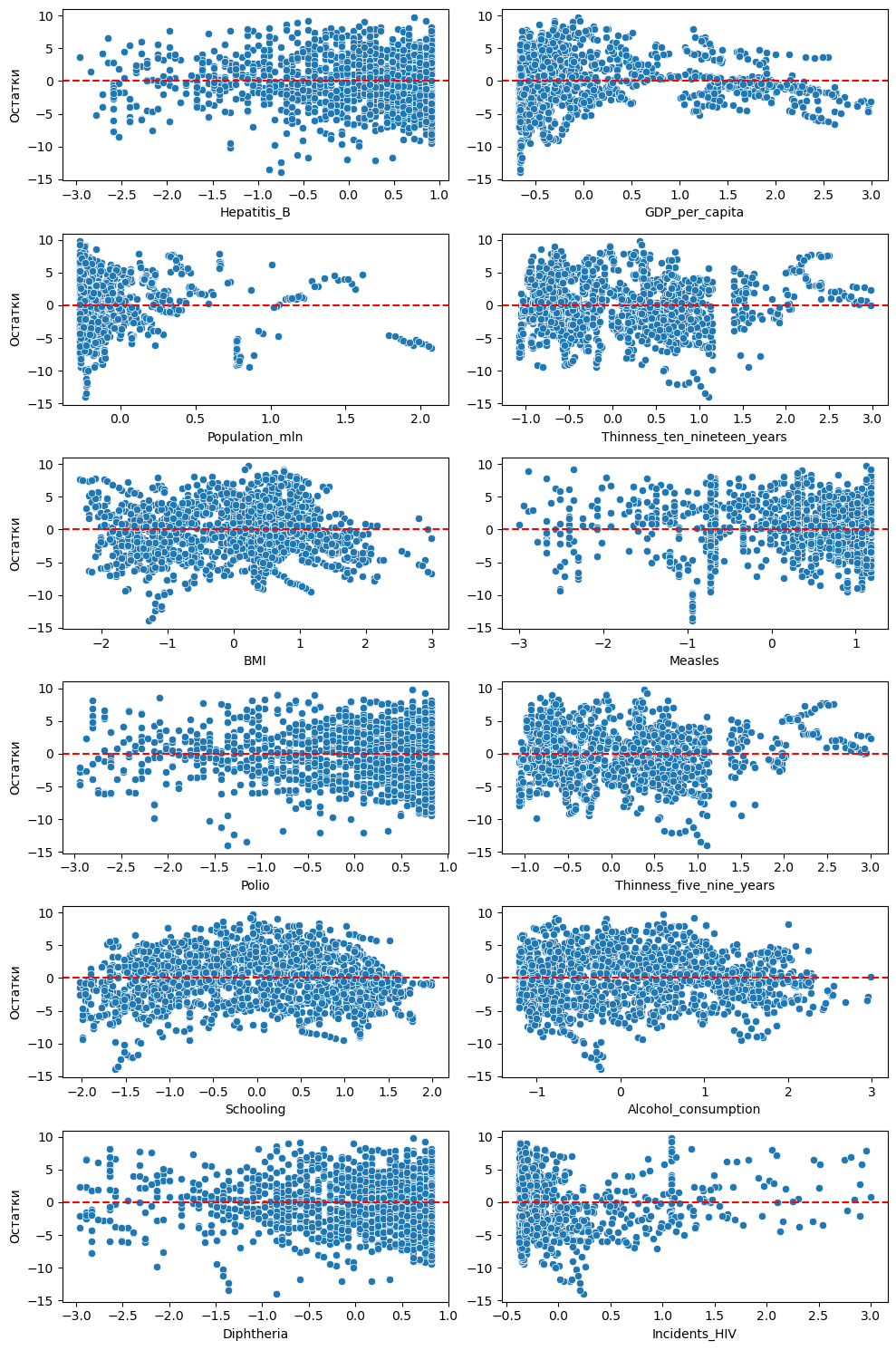
1. Гистограмма остатков:

* Гистограмма показывает распределение остатков модели.
* Кривая нормального распределения (красная линия) наложена на гистограмму для визуального сравнения.
* Остатки кажутся приблизительно нормально распределенными, что является хорошим признаком для линейной регрессии, однако, есть небольшие отклонения в хвостах.

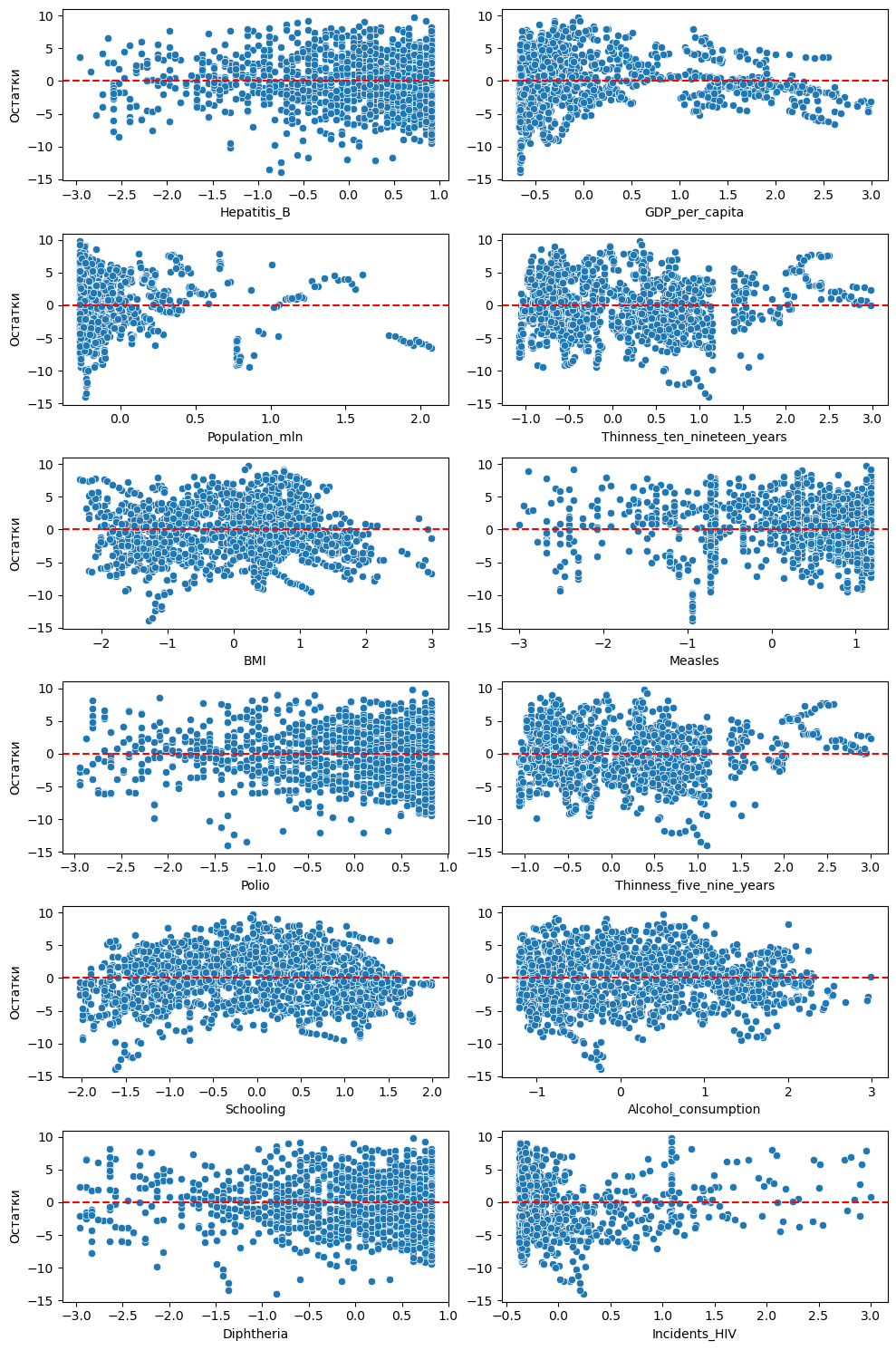
1. Q-Q Plot (график квантиль-квантиль):

* Этот график используется для оценки, насколько хорошо распределение остатков соответствует нормальному распределению.
* Точки следуют вдоль красной пунктирной линии, которая представляет идеальное нормальное распределение, за исключением крайних данных в хвостах.
* Это может указывать на то, что большинство остатков соответствует нормальному распределению, но существуют некоторые выбросы или аномальные значения, особенно в хвостах распределения.

Таким образом, для выявления признаков, наиболее сильно влияющих на мультиколлинеарность, был проведён анализ графиков зависимости остатков регрессионной модели от каждой независимой переменной. Это позволило оценить степень искажения оценок регрессионных коэффициентов, обусловленную взаимной корреляцией объясняющих переменных.



В целом, остатки не показывают существенной систематической ошибки для большинства переменных, за исключением потенциальных нелинейностей или выбросов в данных. Также стоит отметить необычный паттерн для переменной с популяцией в стране. Данные в ней распределены с сильным перевесом количества низких наблюдений. При этом для больших наблюдений мы видим, что присутствует в среднем большая ошибка регрессии. Можно сказать, что модель хорошо предсказывает уровень жизни стран с низким населением.



Также для переменных “ВВП” и “уровень образования” видна нелинейная зависимость, которую не может описать линейная модель.

Основываясь на анализе остатков модели регрессии, который включает тестирование на гетероскедастичность и соответствующую визуализацию, можно сделать вывод о наличии гетероскедастичности в данных. Это указывает на то, что полученные оценки регрессионных коэффициентов могут не быть оптимальными с точки зрения эффективности. В таких случаях, для снижения уровня гетероскедастичности и повышения точности оценок, целесообразно применять трансформацию переменных, в частности, логарифмирование признаков, которые демонстрируют сильную концентрацию наблюдений в области низких значений. Такое преобразование может помочь стабилизировать дисперсию ошибок и улучшить адекватность и точность модели.

Для целей коррекции гетероскедастичности и улучшения нормальности распределения ошибок в модели регрессии были отобраны переменные: 'Population\_mln', 'Incidents\_HIV', 'GDP\_per\_capita', 'Thinness\_five\_nine\_years', 'Alcohol\_consumption'. Логарифмическое преобразование данных переменных было выполнено до разбиения исходного датасета на обучающую и тестовую выборки. Разделение исходного набора данных на обучающий и тестовый поднаборы было осуществлено путем клонирования индексов из предварительно разделенных выборок. Этот метод позволяет поддерживать одинаковые наборы данных для обучения и тестирования различных моделей, что обеспечивает точность последующего сравнительного анализа. Важно отметить, что такой подход гарантирует согласованность и сопоставимость результатов моделирования на разных этапах исследования, включая сравнение моделей с начальной, которая содержала мультиколлинеарность среди объясняющих переменных.

После логарифмирования и разбиения данных были применены дальнейшие этапы предварительной обработки. Это включает в себя стандартизацию переменных, что является важным шагом для устранения влияния различий в масштабах измерения. Также была проведена процедура очистки данных от выбросов для повышения робастности модели.

При всех этапах предобработки данных особое внимание уделялось предотвращению переобучения модели. В этой связи, преобразования и очистка данных проводились с использованием только информации из обучающей выборки, чтобы исключить утечку данных из тестового набора и обеспечить независимость оценки модели.

# Классическая линейная регрессия

В рамках исследования было разработано четыре варианта модели классической линейной регрессии, каждая из которых отличается по степени учёта мультиколлинеарности и преобразованию данных:

1. Стандартная линейная регрессионная модель с учетом всех предикторов, включая те, между которыми присутствует мультиколлинеарность, без внесения корректировок на нее.
2. Модифицированная линейная регрессионная модель без признаков, способствующих мультиколлинеарности, тем самым устраняя её возможное влияние.
3. Оптимизированная линейная регрессионная модель с логарифмированными независимыми переменными, что потенциально снижает эффект гетероскедастичности и улучшает интерпретируемость коэффициентов.
4. Линейная регрессионная модель с логарифмированием как зависимой, так и всех независимых переменных, обеспечивая оценку эластичности одних переменных по отношению к другим.

Для наглядности и удобства сравнения результатов эффективности каждой из моделей представлены в структурированной таблице:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Оценка моделей | | | | | |
| Логарифмированные признаки | Модель | Тренировочная выборка | | | Тестовая выборка | | |
| R^2 | MAE | MAPE | R^2 | MAE | MAPE |
|  |
| Нет | Классическая линейная регрессия с мультиколлинеарностью | 0,819 | 2,94 | 0,044 | 0,782 | 3,313 | 0,05 |  |
| Классическая линейная регрессия | 0,819 | 2,948 | 0,044 | 0,78 | 3,327 | 0,051 |  |
| Есть | Классическая линейная регрессия | 0,866 | 2,337 | 0,035 | 0,865 | 2,628 | 0,04 |  |
| Линейная регрессия с логарифмированной целевой переменной | 0,856 | 2,433 | 0,036 | 0,857 | 2,718 | 0,041 |  |

В представленной таблице приводится сравнительный анализ эффективности различных моделей линейной регрессии на основе метрик качества, применяемых как на тренировочной, так и на тестовой выборках. Оценки моделей базируются на наличии или отсутствии логарифмирования предикторов (независимых переменных) и целевой переменной. Для каждой модели рассчитаны следующие показатели:

* Коэффициент детерминации (R^2), отражающий долю дисперсии зависимой переменной, объясненную моделью.
* Средняя абсолютная ошибка (MAE), представляющая собой среднее абсолютное отклонение прогнозируемых значений от фактических.
* Средний процент абсолютного отклонения (MAPE), измеряющий относительную ошибку прогноза в процентах.

Результаты оценки четырех моделей представлены следующим образом:

1. Модель классической линейной регрессии с мультиколлинеарностью без логарифмирования признаков показала коэффициент детерминации R^2 равный 0,819 на тренировочной выборке и 0,782 на тестовой, что указывает на высокую объясняющую способность модели на обучающих данных и незначительное снижение этой способности на тестовых данных.
2. Классическая линейная регрессия без логарифмированных признаков имеет аналогичные значения R^2, но с незначительно большими ошибками по MAE и MAPE на тестовой выборке.
3. При логарифмировании признаков классическая линейная регрессия показывает улучшение коэффициента детерминации до 0,866 на тренировочной выборке и до 0,865 на тестовой, сопровождаемое снижением ошибок MAE и MAPE.
4. Линейная регрессия с логарифмированной целевой переменной демонстрирует несколько более низкие значения R^2, но все еще высокие, со сравнительно малыми значениями ошибок MAE и MAPE по сравнению с классической линейной регрессией с мультиколлинеарностью.

Эти результаты подчеркивают потенциальные преимущества логарифмирования признаков в улучшении точности и устойчивости линейных регрессионных моделей.

# Линейная регрессия с регуляризацией

В рамках исследования был осуществлён регрессионный анализ, в котором применялись два алгоритма регуляризации: Lasso и Ridge. Эти методы являются инструментами для решения проблемы переобучения и мультиколлинеарности в линейной регрессии. Алгоритм Lasso способствует отбору переменных, обнуляя коэффициенты менее значимых признаков, тем самым упрощая модель и снижая сложность. В свою очередь, Ridge регрессия уменьшает размер коэффициентов, что минимизирует риск переобучения, сохраняя при этом все переменные в модели.

Была использована модель линейной регрессии без предварительного удаления переменных с высокими значениями VIF, основываясь на предпосылке, что регуляризационные методы, такие как Lasso и Ridge, могут эффективно снизить влияние мультиколлинеарности, минимизируя вздутие дисперсии коэффициентов.

Для оптимизации гиперпараметров регуляризации был применён инструмент автоматической настройки гиперпараметров Optuna, который позволяет автоматически итеративно исследовать пространство параметров для поиска наилучшего набора значений. В частности, для модели Lasso искались значения параметра alpha, который контролирует силу регуляризации, в пределах от 0.0001 до 50. Аналогичная процедура была выполнена для модели Ridge с тем же диапазоном для alpha. Optuna применяла кросс-валидацию с пятикратным разбиением для оценки качества модели с различными значениями гиперпараметров, используя коэффициент детерминации R^2 в качестве метрики качества. Этот подход позволяет не только найти оптимальные значения гиперпараметров, но и обеспечить, что модель обладает хорошей обобщающей способностью, что оценивалось по устойчивости показателей качества на разных подвыборках данных.

В результате такого методического подхода к подбору гиперпараметров были найдены наилучшие конфигурации для обеих регуляризированных моделей:

* Ridge: {'alpha': 1.156}
* Lasso: {'alpha': 0.005}

Для модели Ridge регрессии найденное значение alpha равное 1.156 указывает на то, что регуляризационный штраф умеренный. Для модели Lasso регрессии значение alpha равное 0.005 говорит о том, что регуляризационный штраф относительно низкий, но всё же присутствует.

## Результаты моделирования

Для обеспечения объективности анализа эффективности моделей, оснащённых механизмом регуляризации, производится их сравнение с базовыми моделями, построенными на предыдущих этапах исследования. Этот процесс сопоставления необходим для того, чтобы определить, вносит ли введение регуляризации улучшения в предсказательную способность моделей. В частности, будет проанализировано изменение величин коэффициента детерминации (R²), средней абсолютной ошибки (MAE) и средней абсолютной процентной ошибки (MAPE) до и после применения регуляризации, что позволит оценить улучшение или ухудшение точности и обобщающей способности моделей.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Логарифмированные признаки | Удалены признаки с большим VIF | Модель | Тренировочная выборка | | | Тестовая выборка | | |
| R^2 | MAE | MAPE | R^2 | MAE | MAPE |
|
| Нет | Нет | Классическая линейная регрессия | 0,819 | 2,940 | 0,044 | 0,782 | 3,313 | 0,050 |
| Да | Классическая линейная регрессия | 0,819 | 2,948 | 0,044 | 0,780 | 3,327 | 0,051 |
| Есть | Классическая линейная регрессия | 0,866 | 2,337 | 0,035 | 0,865 | 2,628 | 0,040 |
| Линейная регрессия с логарифмированной целевой переменной | 0,856 | 2,433 | 0,036 | 0,857 | 2,718 | 0,041 |
| Нет | Lasso регрессия | 0,868 | 2,330 | 0,035 | 0,864 | 2,634 | 0,040 |
| Ridge регрессия | 0,868 | 2,328 | 0,035 | 0,865 | 2,626 | 0,040 |

В рамках анализа результатов, полученных на тренировочных и тестовых выборках, обнаружено, что модели с регуляризацией Lasso и Ridge демонстрируют превосходные показатели на тренировочных данных по сравнению с базовыми моделями. Однако, при переходе к тестовой выборке, обе модели с регуляризацией выявляют аналогичные значения метрик качетсва, что указывает на их сопоставимую обобщающую способность с моделями, где были исключены мультиколлинеарные признаки. В контексте сравнения моделей Lasso и Ridge, хотя различия в средней абсолютной ошибке (MAE) на тестовой выборке незначительны, преимущество отдаётся модели Ridge. Такое решение обусловлено величиной коэффициента регуляризации, который в случае Ridge регрессии больше, что свидетельствует о более сильном воздействии на регуляризацию модели. Это может указывать на потенциально большую устойчивость модели к переобучению по сравнению с Lasso, что делает Ridge регрессию более предпочтительной в данных обстоятельствах.

Таким образом, хотя различия между Lasso и Ridge регрессиями могут казаться маргинальными, выбор в пользу Ridge может быть оправдан при более детальном рассмотрении свойств регуляризации, в частности, её способности к снижению дисперсии оценок и предотвращению переобучения.

# Регрессия с полиномиальными признаками

Предварительный визуальный анализ данных выявил потенциальные нелинейные отношения между зависимой переменной и независимыми переменными. В целях более глубокого исследования этих нелинейностей, представляется целесообразным модифицировать спецификацию исходной модели. Это предполагает включение в модель дополнительных переменных, которые представляют собой квадраты исходных предикторов и их попарные произведения. Такой подход может привести к значительному увеличению числа предикторов и соответствующему росту размерности признакового пространства, что потенциально увеличивает риск столкнуться с "проклятием размерности". Это явление характеризуется снижением эффективности алгоритмов машинного обучения при резком увеличении количества признаков по сравнению с количеством наблюдений.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Логарифмированные признаки | Удалены признаки с большим VIF | Модель | Тренировочная выборка | | | Тестовая выборка | | |
| R^2 | MAE | MAPE | R^2 | MAE | MAPE |
|
| Нет | Нет | Классическая линейная регрессия | 0,819 | 2,940 | 0,044 | 0,782 | 3,313 | 0,050 |
| Да | Классическая линейная регрессия | 0,819 | 2,948 | 0,044 | 0,780 | 3,327 | 0,051 |
| Есть | Классическая линейная регрессия | 0,866 | 2,337 | 0,035 | 0,865 | 2,628 | 0,040 |
| Линейная регрессия с логарифмированной целевой переменной | 0,856 | 2,433 | 0,036 | 0,857 | 2,718 | 0,041 |
| Нет | Lasso регрессия | 0,868 | 2,329 | 0,035 | 0,865 | 2,631 | 0,040 |
| Ridge регрессия | 0,868 | 2,328 | 0,035 | 0,865 | 2,626 | 0,040 |
| Да | Регрессия с полиномиальными признаками | 0,934 | 1,648 | 0,025 | 0,726 | 2,519 | 0,042 |

На основе полученных результатов можно констатировать, что модель демонстрирует улучшенные показатели на обучающем наборе данных по всем рассматриваемым метрикам. В то же время, анализируя результаты на тестовой выборке, обнаруживается некоторое снижение коэффициента детерминации (R^2) и возрастание средней абсолютной процентной ошибки (MAPE), хотя средняя абсолютная ошибка (MAE) при этом уменьшилась.

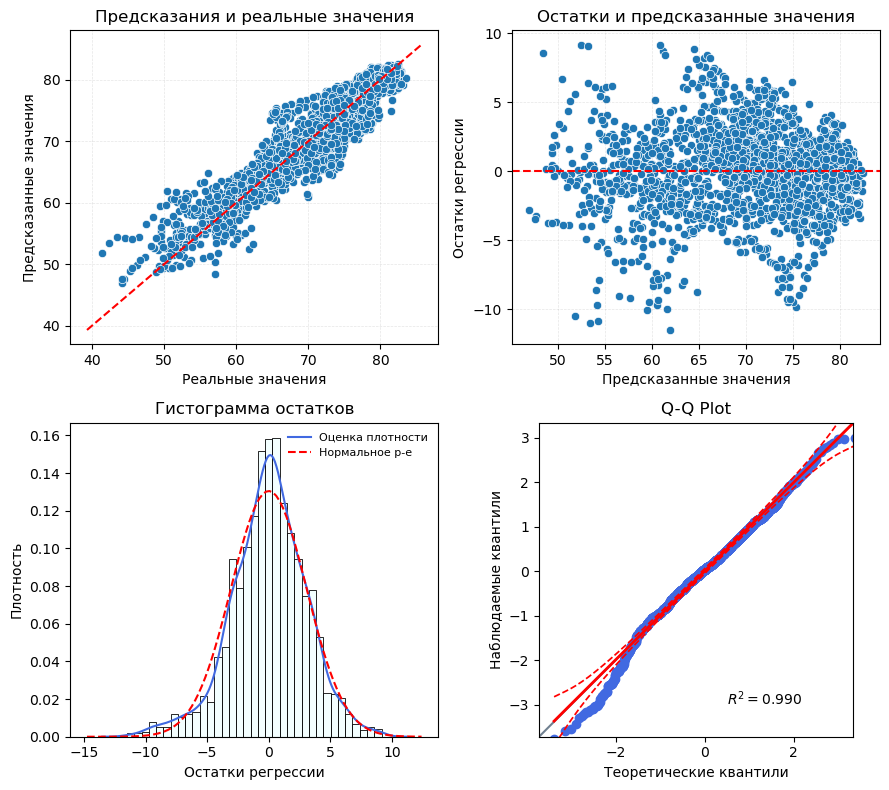
Такие разнонаправленные тенденции в изменении показателей точности могут быть свидетельством того, что модель лучше приспособилась к предсказанию экстремальных значений, потеряв при этом точность в оценке данных, близких к среднему уровню. Исходя из всего перечисленного, можно предположить, что наблюдается явление переобучения модели: повышение её сложности привело к "запоминанию" шума или структурных особенностей обучающего набора данных в ущерб способности генерализации на новые данные.

# Итоговая модель

Основной задачей для итоговой модели было поставлено получить простую модель с минимальным числом признаков для экономической интерпретации и анализа взаимосвязей.

## Модель без удаления “незначимых” признаков

В ходе сравнительного анализа, в качестве финальной модели было принято решение использовать классический метод линейной регрессии, который был модифицирован путём исключения факторов, вызывающих мультиколлинеарность, и применения логарифмического преобразования к некоторым переменным. Выбор данной модели вместо более сложной Ridge регрессии обусловлен её большей интерпретируемостью и простотой, и схожестью результатов на тестовой выборке, что подчёркивает отсутствие преимущества использования метода с регуляризацией в данном контексте.



|  |  |  |  |
| --- | --- | --- | --- |
|  | P-values | Отклоняем H0 | Вывод |
| Тест Шапиро-Уилка | 1.072522e-10 | True | Остатки не распределены нормально |
| Тест Бреуша-Пагана | 1.944986e-30 | True | Присутствует гетероскедастичность |
| Тест Уайта | 3.937329e-98 | True | Присутствует гетероскедастичность |

На основе проведённых статистических тестов была выявлена гетероскедастичность остатков в рассматриваемой модели, что подчёркивает необходимость корректировки стандартных ошибок для обеспечения надёжности выводов о значимости коэффициентов регрессии. В ответ на эту проблему, для оценки стандартных ошибок будут применены значения, скорректированные с использованием поправки HC1. Однако следует осознавать, что при отклонении от нормальности распределения остатков, уверенность в результаты тестов уменьшается. Несмотря на предположение о сходимости распределения остатков к нормальному в соответствии с Центральной Предельной Теоремой при увеличении объёма выборки, данное предположение остаётся лишь аппроксимацией, учитывая, что на практике мы ограничены конечным набором данных. В свете этих соображений, предполагается с осторожностью относиться к выводам тестов на значимость коэффициентов, признавая внесённое допущение о нормальности распределения остатков как необходимость для интерпретации результатов в рамках доступной выборки.

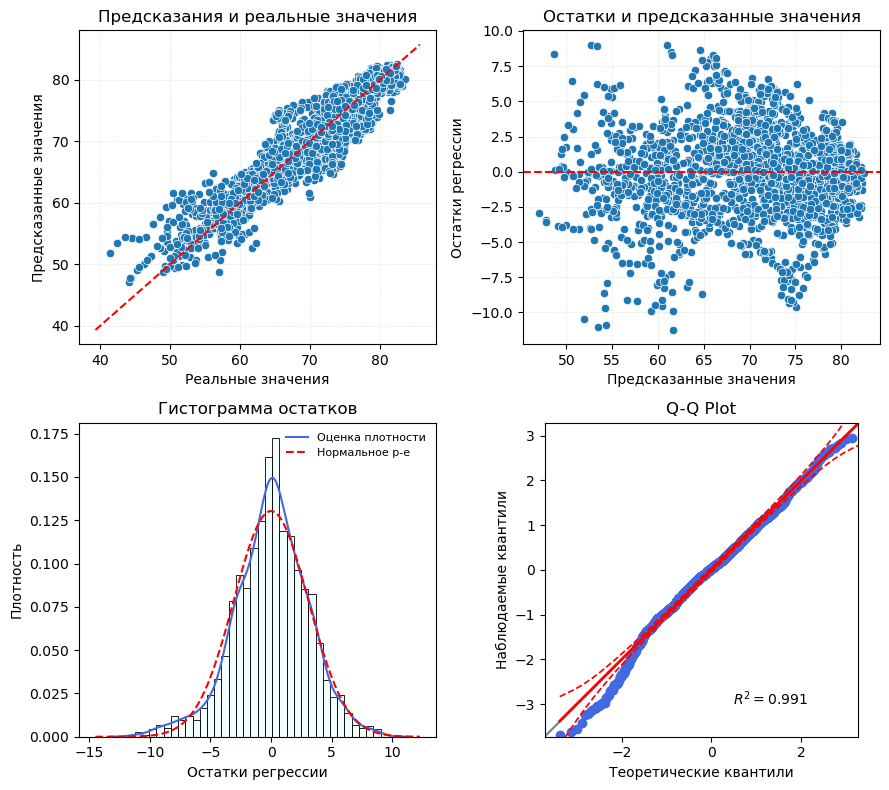
## Модель с удалёнными незначимыми признаками

Проведя анализ значимости коэффициентов с использованием тестов на нулевую гипотезу о равенстве коэффициентов нулю и применяя поправку HC1 Уайта для коррекции на гетероскедастичность, были исключены те признаки, чьи P-значения превышали установленный порог значимости в 0.05. В результате была сформирована модифицированная регрессионная модель:

Итоговая таблица со всеми проанализированными моделями:

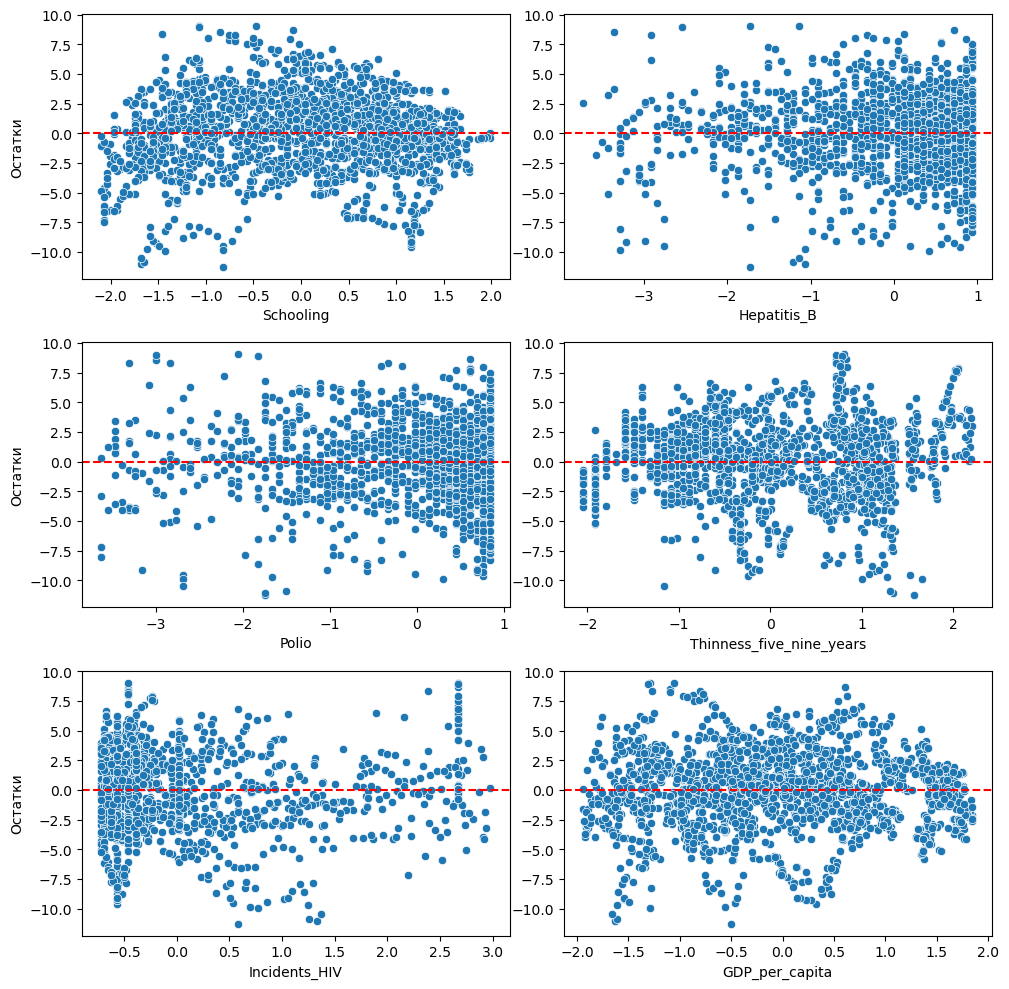
| Логарифмированные признаки | Удалены признаки с большим VIF | Модель | Тренировочная выборка | | | Тестовая выборка | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| R^2 | MAE | MAPE | R^2 | MAE | MAPE |
|  |
| Нет | Нет | Классическая линейная регрессия с мультиколлинеарностью | 0,819 | 2,940 | 0,044 | 0,782 | 3,313 | 0,050 |  |
| Да | Классическая линейная регрессия | 0,819 | 2,948 | 0,044 | 0,780 | 3,327 | 0,051 |  |
| Есть | Классическая линейная регрессия | 0,866 | 2,337 | 0,035 | 0,865 | 2,628 | 0,040 |  |
| Линейная регрессия с логарифмированной целевой переменной | 0,856 | 2,433 | 0,036 | 0,857 | 2,718 | 0,041 |  |
| Нет | Lasso регрессия | 0,868 | 2,329 | 0,035 | 0,865 | 2,631 | 0,040 |  |
| Ridge регрессия | 0,868 | 2,328 | 0,035 | 0,865 | 2,626 | 0,040 |  |
| Да | Регрессия с полиномиальными признаками | 0,934 | 1,648 | 0,025 | 0,726 | 2,519 | 0,042 |  |
| Регрессия с удалёнными незначимыми признаками | 0.865 | 2,345 | 0,035 | 0,864 | 2,643 | 0,040 |  |

Анализ метрик качества позволяет заключить, что итоговая модель демонстрирует показатели, которые, хотя и немного уступают по эффективности более ранним, более сложным моделям, тем не менее сохраняют сопоставимый уровень производительности.

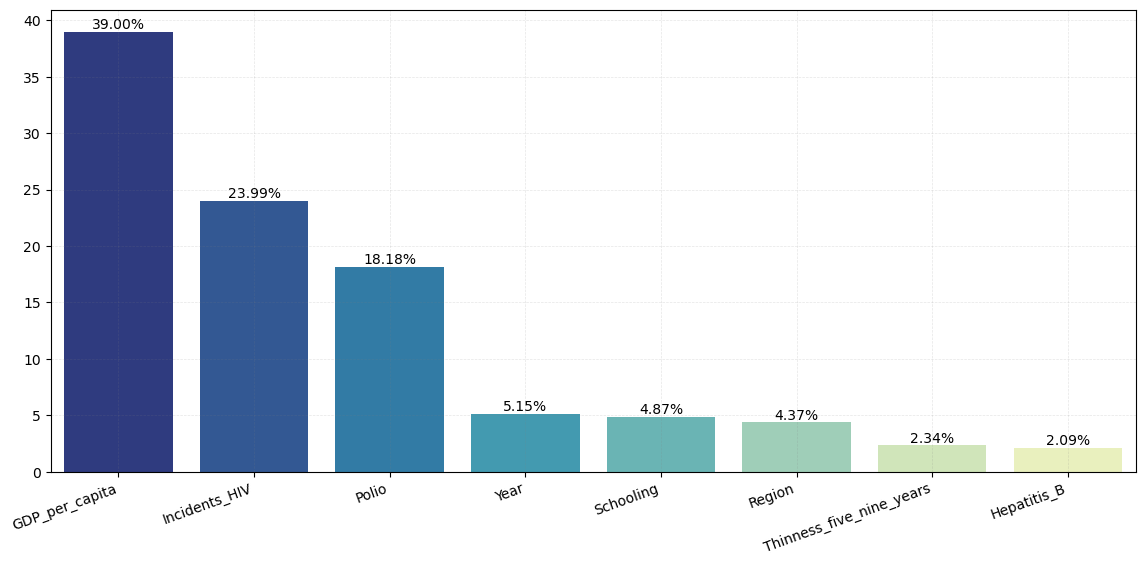


Из анализа распределения остатков модели следует, что они демонстрируют более высокую концентрацию вокруг нулевого значения, относительно нормального распределения. Кроме того, обнаруживается лёгкое увеличение частоты встречаемости выбросов в области значений около -10. Так же, как и предшествующие модели, текущая модель продолжает проявлять признаки гетероскедастичности остатков. Вопреки этому обстоятельству, структурные особенности данных ограничивают возможности дальнейшей коррекции гетероскедастичности в рамках применённой модели.

Остатки регрессии относительно каждой переменной в модели:



## Интерпретация результатов



Диаграмма, представленная на графике, показывает относительное влияние различных предикторов на целевую переменную, которой в данном случае является "средняя продолжительность жизни". Вот экономическая интерпретация результатов:

* ВВП на душу населения (39,00%): Это наиболее влиятельный предиктор. По мере увеличения ВВП на душу населения, что указывает на более высокий уровень среднего дохода, это обычно ведет к лучшей медицинской помощи, питанию и условиям жизни, что может значительно увеличить среднюю продолжительность жизни.
* Число случаев ВИЧ (23,99%): Отрицательный коэффициент в уравнении регрессии указывает на то, что с увеличением числа случаев ВИЧ продолжительность жизни снижается. ВИЧ/СПИД может радикально сокращать продолжительность жизни из-за его воздействия на здоровье и смертность.
* Полиомиелит (18,18%): Положительный коэффициент для уровня вакцинации от полиомиелита подразумевает, что более высокое покрытие вакцинацией связано с более долгой продолжительностью жизни, что отражает эффективность общественных инициатив по предотвращению опасных заболеваний.
* Год (5,15%): Со временем улучшения в медицинской технологии, инфраструктуре здравоохранения и уровне жизни способствуют увеличению продолжительности жизни. Положительный коэффициент указывает на то, что более поздние годы связаны с более высокой продолжительностью жизни.
* Образование (4,87%): Образование способствует продолжительности жизни, поощряя более здоровые практики, улучшая возможности трудоустройства и увеличивая доступ к медицинской помощи. Положительный коэффициент здесь указывает на то, что более высокий уровень образования коррелирует с более долгой продолжительностью жизни.
* Регион (4,37%): Различные регионы могут иметь разные продолжительности жизни из-за различий в здравоохранении, культуре, окружающей среде и экономических условиях. Положительный коэффициент предполагает, что продолжительность жизни может сильно различаться по регионам.
* Худоба у детей в возрасте от пяти до девяти лет (2,34%): Недоедание в ранние годы может привести к долгосрочным проблемам со здоровьем, негативно влияя на продолжительность жизни. Отрицательный коэффициент отражает эту связь.
* Гепатит В (2,09%): С отрицательным коэффициентом увеличение заболеваемости гепатитом В негативно влияет на продолжительность жизни. Гепатит В может вызывать хронические заболевания печени, которые представляют собой значительную медицинскую нагрузку.