**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего образования**

**"Национальный исследовательский университет "Высшая школа экономики"**

**Факультет мировой экономики и мировой политики**

**Образовательная программа «Мировая экономика»**

**ПРОЕКТ**

по дисциплине: “Эконометрика”

на тему:

«Факторы, определяющие внутристрановое неравенство по доходу»

Выполнили:

Олейников А.

Скоробогатова Т.

Камаданова А.

Группа № 213

Проверила: Рожкова К.В.

Москва, 2023 г.

**Содержание**

[**Введение 3**](#_ifhicvcpv6l)

[**Описание данных 6**](#_wndfg3wcv806)

[**Регрессионный анализ 12**](#_y8gf1oj8arbr)

[**Результаты второго выбранного метода 16**](#_ri1mga7ir1p5)

[**Заключение 18**](#_ddeqtlyr6i7a)

[**Список литературы: 19**](#_v3l5of3uj83q)

# **Введение**

В современном мире существует **проблема неравенства в доходах между различными социальными группами**. Сегодня эта экономическая дилемма актуальна и важна для анализа как никогда. Как в развитых, так и в развивающихся странах увеличивается финансовый разрыв между богатым и бедным населением, средний класс размывается на их фоне. Многие факторы такие как различия в заработных платах, доступ к образованию и здравоохранению, различия в жилищных условиях влияют на экономическое положение граждан одного государства. Уменьшение такого неравенства может способствовать более стабильному обществу, а также стимулировать экономический рост и развитие страны в целом.

**Цель нашего исследования** - рассмотреть влияние определенных факторов на уровень неравенства по доходу в разных странах мира.

Исходя из приведенной выше экономической проблемы, можно сформулировать такую гипотезу:

**Гипотеза:** Факторы социального положения, образования и экономического развития страны в равной степени оказывают влияние на уровень неравенства в государстве.

1. Статьи, в которых исследуется проблема внутристранового неравенства по доходу:
2. Malerba, G., & Spreafico, M. (2014). STRUCTURAL DETERMINANTS OF INCOME INEQUALITY IN THE EUROPEAN UNION: EVIDENCE FROM A PANEL ANALYSIS. *Rivista Internazionale Di Scienze Sociali*, *122*(1), 37–83. <http://www.jstor.org/stable/43830200>

В данной статье авторы рассматривают факторы, которые формируют неравенство в располагаемых доходах домохозяйств в Европейском Союзе. Не уделяют особое внимание проблема домохозяйства, учитывая недавние сокращения традиционно государственных пособий, изменения в структуре семьи и культурную валюту, касающиеся отношений между членами семьи и поколениями. Статья включает широкий набор индикаторов, определяющих неравенство доходов, такие как макроэкономические (в частности - экономический рост), институциональные, культурные и социальные факторы. Также они уделяют внимание государственным расходам, ведь они также влияют на располагаемый доход домохозяйств.

1. Afonso, António and Schuknecht, Ludger and Tanzi, Vito, Income Distribution Determinants and Public Spending Efficiency (May 26, 2010). Journal of Economic Inequality, Vol. 6, No. 3, pp. 367-389, 2010, Available at SSRN: <https://ssrn.com/abstract=1615962>

Авторы этой статьи в своей работе изучают факторы, определяющие распределение доходов и эффективность государственных расходов на снижение внутристранового неравентсва. Они изучают влияние государственных расходов, образования на распределение доходов в странах с развитой экономикой. Они также оценивают эффективность государственных расходов на перераспределение доходов, используя непараметрический подход Data Envelopment Analysis. Результаты показывают, что государственная политика существенно влияет на распределение доходов через социальные расходы и косвенно через высококачественное образование и надежные экономические институты. Таким образом, в стране уменьшается неравенство между различными социальными слоями общества.

1. **Исследовательский вопрос:** Оказывают ли равное влияние выбранные для анализа факторы на уровень неравенства в стране?

Как мы писали раньше за **гипотезу** мы взяли: Факторы социального положения, образования и экономического развития страны в равной степени оказывают влияние на уровень неравенства в государстве.

1. Переменные, которые будут задействованы в анализе, с точки зрения практического смысла и методологии их измерения.

*Рост ВВП* в процентах в той или иной стране (количественная переменная) важен, чтобы понимать как быстро развивается экономика государства в рассматриваемом году. Кроме этого, также рассматривается к *какому типу “развитая” или “развивающаяся”* относятся выбранные страны (данная переменная была взята в качестве категориальной, был вариант взять определение уровня страны по доходу по классификации Всемирного Банка: низкий, ниже среднего, выше среднего, высокий, но тип страны нам показался более широким понятием, а ВВП на душу населения был включен количественной переменной),. *Доля ВВП, получаемая за счет использования природных ресурсов* важна для анализа, так как, возможно, доход от использования природных ресурсов проще сконцентрировать в отдельных руках, что приведет к повышению неравенства.В качестве количественной переменной также был взят коэффициент фертильности - сколько в среднем в течение жизни женщины в стране у нее может родиться детей. *Расходы на здравоохранение в процентах от ВВП* как индикатор отчасти может показывать качество человеческого капитала. *Продолжительность обязательного образования* также хороший социальный фактор для понимания развития человеческого капитала. Ведь в основном люди с более качественным образованием получают более высокий заработок. *Восприятие способности и возможности правительства разрабатывать и применять законы для развития частного сектора* оказывает непосредственное влияние на развитие бизнеса, что, в свою очередь, является местом для трудоустройства многих людей и отражает качество политических институтов. *Уровень безработицы (в %)* показывает, какая часть населения страны вынуждена жить на пособие по безработице, что тоже может служить фактором разрыва между богатыми и бедными. *Население, проживающее в сельской местности (в %)* отражает количество людей с небольшим заработком, так как в деревнях и селах большая часть населения зарабатывает очень мало. *ВВП на душу населения* также стоит включить в анализ, ведь он отражает уровень экономической активности страны и качество жизни людей. *Доля от общего дохода у 50% самых бедных*  и *доля от общего дохода у 10% самых богатых населения*  - два показателя, отношение между которыми будет использоваться для составления целевой переменной

Для исследования возьмем данные на 2021 год.

Мы считаем, что между выбранными переменными существует в большей мере линейная взаимосвязь. Например, чем больше расходов государства идет на здравоохранение, тем больше оказывается медицинских услуг граждан, тем они могут успешнее выполнять свою работу, тратя меньше времени на ожидание получения медицинской помощи. Чем быстрее принимаются законы правительством о развитие частного бизнеса и чем выше качество политических институтов, тем увереннее чувствуют себя люди и неравенство может быть меньше из-за успешной распределительной политики.

1. Методы, которыми мы будем пользоваться при анализе нашей проблемы:
2. Первый метод - регрессионный анализ с интерпретацией коэффициентов (линейная модель)

Для исследования будем рассматривать линейную регрессию, так как это математическая модель, с помощью которой можно описать связь нескольких переменных и попробовать спрогнозировать как тот или иной фактор будет влиять на внутристрановое неравенство.

1. Второй метод – Регрессионные деревья

Это алгоритм, который в результате представляет одно дерево принятия решений, которое позволяет проверить на наличие связей между переменными.

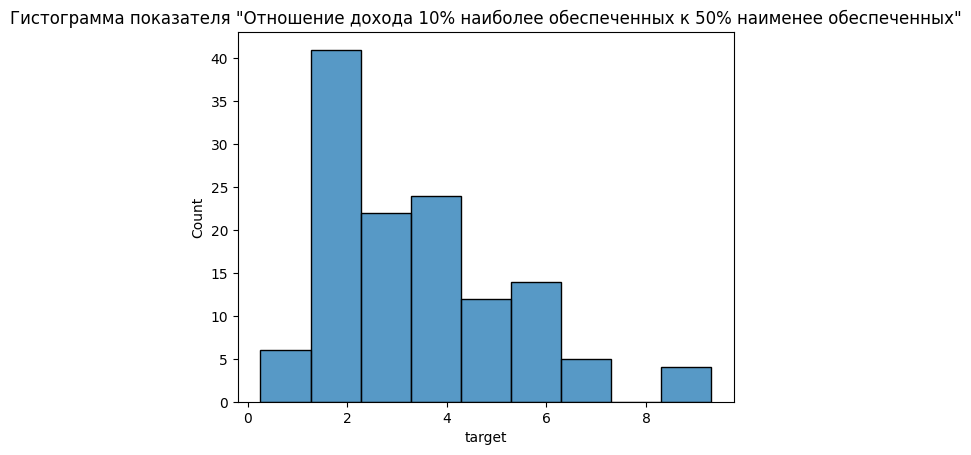
Ожидаемые результаты: главным образом, должна подтверждаться взаимосвязь между анализируемыми переменными. А также то, что факторы в равной степени оказывают влияние на уровень неравенства в странах.

# **Описание данных**

Большая часть датасета была создана путем ручного отбора индикаторов с сайта Всемирного банка (https://data.worldbank.org/). При этом для части переменных Всемирный Банк выступал в качестве агрегатора данных, предоставляя информацию, полученную из других источников: расходы на здравоохранение - из ВОЗ, суммарный коэффициент фертильности - из статистических агентств стран и статистического подразделения ООН, продолжительность обязательного образования в стране - ЮНЕСКО, уровень безработицы - Международная организация труда. Также для получения зависимой переменной использовались данные World Inequality Database (<https://wid.world/data/>), а именно доли дохода, получаемые 10% самого обеспеченного и 50% наименее обеспеченного населения страны. Причиной решения использовать для измерения неравенства в стране отношение дохода 10% наиболее обеспеченных к 50% наименее обеспеченных, а не такого общепризнанного показателя, как коэффициент Джини, стала недостаточность статистических данных по последнему - для многих стран последнее известное значение коэффициента Джини датировалось 2014 годом.

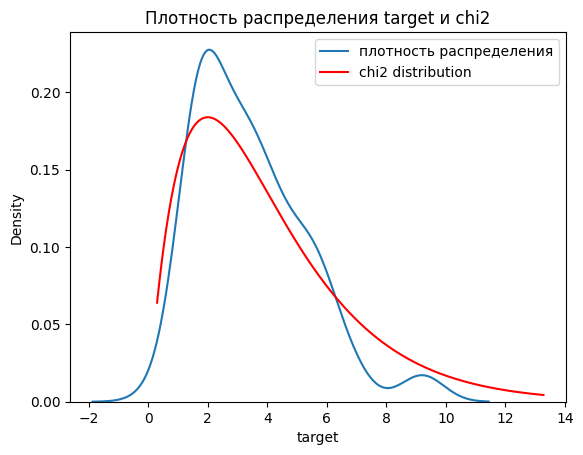
Датасет содержит в себе информацию по 132 странам и 9 показателям за 2021 + 1 показатель за 2020 год (это показатель расходов на здравоохранение, так как на наш взгляд информация о данном показателе могла стать важной объясняющей переменной, а более свежая информация отсутствовала на сайте Всемирного Банка. Однако мы считаем, что можно использовать информацию по данной переменной за 2020 год, так как данный период был аномальным для всех стран из-за пандемии и расходы на здравоохранение достигали наибольших значений, то есть в определенной степени это оценка сверху для 2021 года, но при этом часть средств, потраченных на здравоохранение в 2020 году могла в том числе использоваться и в следующие периоды).

Была построена гистограмма частотности зависимой переменной (Рисунок 1).



**Рисунок 1 - Гистограмма показателя “Отношение дохода 10% самых богатых к 50% самых бедных”**. Источник: составлено авторами на основе данных <https://wid.world/data/>

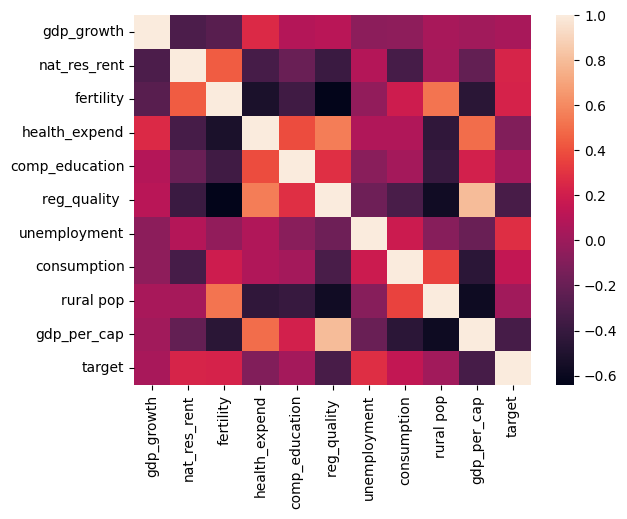
Визуально данное распределение напоминает Хи^2, так как наибольшее значение смещено к левому краю графика, для подтверждения данного предположения был построен график плотности распределения и распределения Хи^2 с 4 степенями свободы (Рисунок 2)



**Рисунок 2 - Плотность распределения target и chi2**

При расчете описательных статистик было получено, что среднее значение целевой переменной составляет 3.68 и медиана - 3.19, что может говорить о существовании выбросов, так как наибольшее значение данного показателя 12.67. Интересным фактом является то, что среднее значение и медиана показателя “reg\_quality” не равны 0 (0.08 и -0.08 соответственно), так как по информации Всемирного Банка данный показатель был отнормирован к нормальному распределению. Объяснением несоответствия может служить факт исключения из выборки части стран, у которых были пропущенные значения по хотя бы одной из переменных. Полная информация о корреляции между переменными содержится в программном коде.

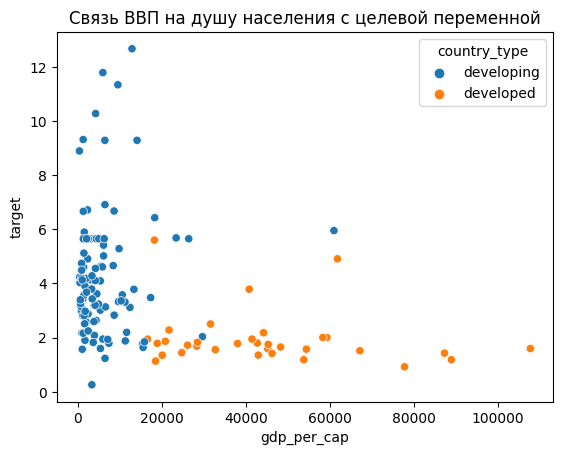
Для первичного анализа связи между переменными была построена тепловая карта корреляции (рисунок 3).

****

**Рисунок 3 - Тепловая карта корреляции**

На данной карте можно видеть, что наиболее выраженная линейная связь с целевой переменной у “reg\_quality” - качества правительства и “gdp\_per\_cap” (значения корреляции -0.33 и -0.34 соответственно), то есть изначальное предположение об отрицательной связи неравенства по доходу в стране с качеством управления на высшем уровне в определенной степени подтверждается корреляцией. Для других объясняющих переменных корреляция с целевой по модулю не превышает 0.16. Стоит отметить существование ярко выраженной корреляции между объясняющими переменными: расходами на здравоохранение и качеством управления с уровнем фертильности, долей сельского населения с ВВП на душу населения и качеством управления. Данный факт может говорить о возможном существовании мультиколлинеарности, который будет необходимо иметь в виду при проведении регрессионного анализа.

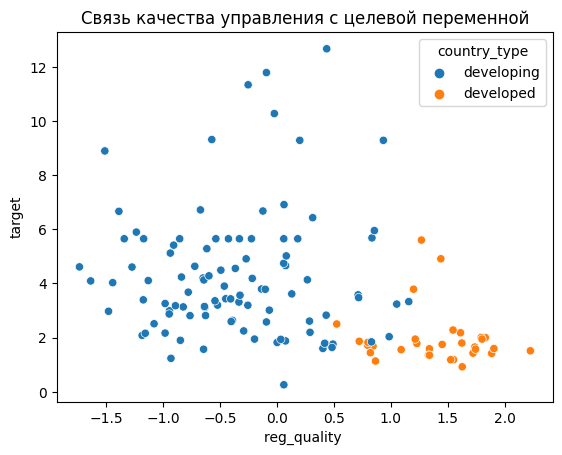
Для выявления характера связи между целевой переменной и ВВП на душу населения была построена диаграмма рассеивания, существование отрицательной корреляции позволяет ожидать распределение точек в пределах общего тренда (рисунок 4).



**Рисунок 4 - Связь ВВП на душу населения с целевой переменной**

Однако график демонстрирует, что отрицательная связь между показателями объясняется наличием в выборке развитых стран, характеризуемых в большинстве своем низким уровнем неравенства и высоким ВВП на душу населения, в то время как основная часть наблюдений - развивающиеся страны- располагаются в пределах от о до $20000 на душу населения, то есть, скорее всего, при исключении из датасета развитых стран, корреляция была бы меньше.

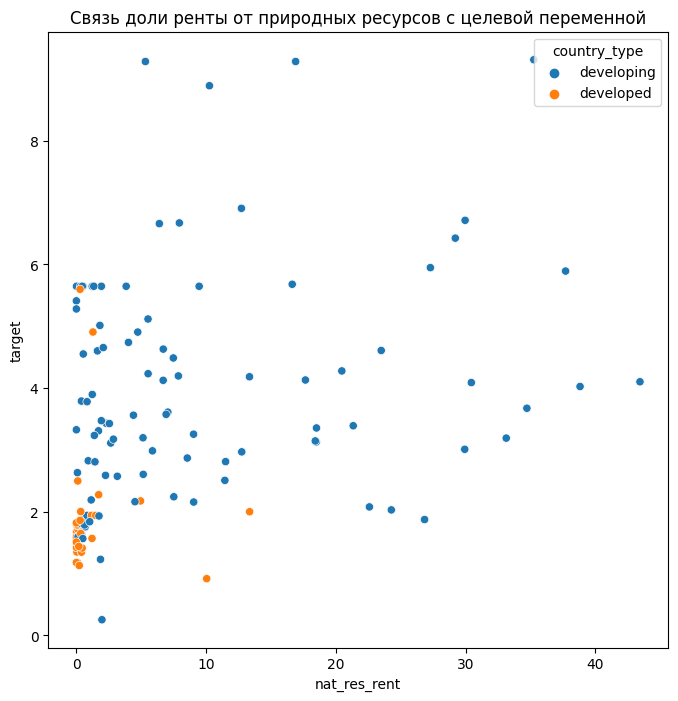
При этом для переменной качество управления отрицательная зависимость с целевой переменной наблюдается не только для развитых стран (рисунок 5).



**Рисунок 5 - Связь качества управления с целевой переменной**

В целом за исключением отдельных наблюдений можно видеть существование линейной связи: с ростом качества управления снижается значение отношения доходов 10% самых обеспеченных к 50% наименее обеспеченных -то есть население становится более равномерным по доходу, чем было при низком качестве управления.

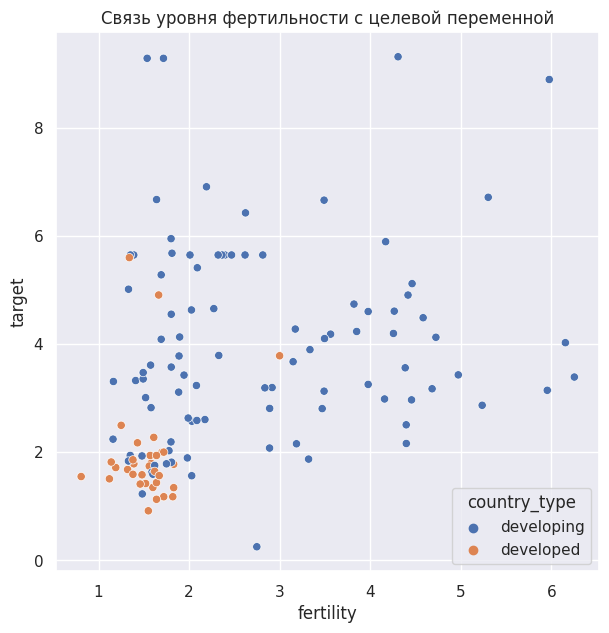
Так как низкая корреляция не обязательно означает низкий уровень связи между переменными (она может быть нелинейной, и тогда корреляция будет низкой, но взаимосвязь способна существовать) было решено построить графики связи целевой переменной с несколькими другими объясняющими переменными. Для доли ренты от природных ресурсов в ВВП ожидается, что при росте данного показателя будет расти и неравенство, так как высокий доход от ресурсов проще концентрировать в руках небольшого числа людей (рисунок 6).



**Рисунок 6 - Связь доли ренты от природных ресурсов с целевой переменной**

Однако для большинства наблюдений доля ренты от природных ресурсов находится от 0 до 10% и также есть значительный разброс целевого показателя для данных стран, но с ростом доли ренты от природных ресурсов нельзя проследить четкий тренд увеличения неравенства по доходу.

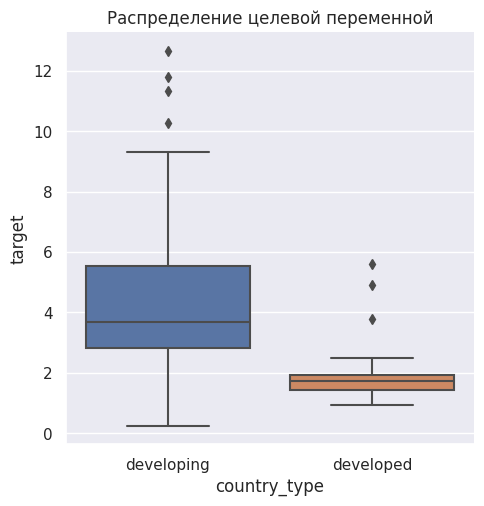
Для уровня фертильности ожидалось увидеть нелинейную связь с уровнем неравенства, так как значительная фертильность ти может говорить о нахождении страны на этапе первого демографического перехода и необходимости иметь столько детей для работы в сельском хозяйстве, а доход отдельного индивида ниже, чем в случае с более современной структурой семьи. То есть рост уровня фертильности может приводить к росту неравенства темпами, превышающими линейные (рисунок 7).



**Рисунок 7 - Связь уровня фертильности с целевой переменной**

По рисунку можно отметить существование двух трендов: с ростом фертильности от 1 до 3 наблюдается сильный линейные рост неравенства, а при росте от 3 до 6 общий возрастающий тренд не так ярко выражен.

Для выделения выбросов в целевой переменной был построен график boxplot в зависимости от уровня развития страны (рисунок 8)

****

**Рисунок 8 - Распределение целевой переменной**

По графику можно сделать вывод, что уровень неравенства в развивающихся странах выше (среднее значение около 4, для развитых - около 2), и также присутствуют выбросы как для развитых, так и для развивающихся стран. Было принято решение удалить 4 выброса для развивающихся стран (Коста-Рика -12.67, ЮАР - 11.78, Мексика 11.33, Намибия 10.27), так как они являются экстремальными для всей выборки в целом и могут значительно изменить результаты построения модели. При этом для выбросов в развитых странах было принято решение оставить их в итоговом датасете, так как среди всех наблюдений присутствует только 33 развитых страны - возможный негативный эффект от удаления 10% этой части выборки был оценен более значительным, чем наличие данных значений у развитых стран в пределах стандартных чисел для развивающихся.

Таким образом, датасет для построения моделей содержит 128 наблюдений.

# **Регрессионный анализ**

По итогам составления линейной регрессионной модели, включавшей в себя 10 количественных переменных и одну качественную - уровень развития страны (закодированную 1, если страна развивающаяся и 0, если страна развитая), был получен следующий результат (таблица 1):

| Переменная | коэффициент | стандартная ошибка | t-статистика | p-value |
| --- | --- | --- | --- | --- |
| const | -0.1555 | 1.528 | -0.102 | 0.919 |
| gdp\_growth | 0.0680 | 0.034 | 1.993 | 0.049 |
| nat\_res\_rent | 0.0371 | 0.021 | 1.803 | 0.074 |
| fertility | 0.3833 | 0.175 | 2.188 | 0.031 |
| health\_expend | 0.0922 | 0.077 | 1.196 | 0.234 |
| comp\_education | 0.0197 | 0.068 | 0.290 | 0.772 |
| reg\_quality | -0.0607 | 0.309 | -0.196 | 0.845 |
| unemployment | -0.0168 | 0.036 | -0.466 | 0.642 |
| consumption | 0.0136 | 0.014 | 0.950 | 0.344 |
| rural pop | -0.0282 | 0.010 | -2.764 | 0.007 |
| gdp\_per\_cap | -6.326e-06 | 1.33e-05 | -0.476 | 0.635 |
| country\_type\_developing | 1.8058 | 0.626 | 2.887 | 0.005 |

**Таблица 1 -** Результат построения регрессионной модели

Сама модель получилась адекватной, так как при проверке гипотезы против альтернативы вида:

F-статистика, получаемая по формуле: F-stat = F(n-1,n-k-1), где ESS - объясненная сумма квадратов отклонений, RSS - сумма остатков квадратов регрессии, n - число наблюдений, k - число объясняющих переменных, составила 6.113, а p-value - 7.76\*10^(-8), то есть на любом разумном уровне значимости нулевая гипотеза о том, что все коэффициенты равны 0, отвергается в пользу альтернативы. R^2 составил 0,367 (нормированный R^2 равен 0,307), но при этом у модели присутствует только 4 переменных значимых на уровне значимости в 5% - это gdp\_growth, fertility, rural pop, country\_type\_developing. У остальных переменных значение p-value больше 0.05.

Причиной данного факта может быть недостаточное число наблюдений - в выборке присутствуют только 128 стран и, возможно, если бы наблюдений было больше, модель смогла бы лучше обучиться. Но главным объяснением мы считаем наличие корреляции между факторами, как было упомянуто на этапе описания данных, существовал риск возникновения мультиколлинеарности из-за наличия высокой корреляции между отдельными переменными и из-за нее часть переменных могла оказаться незначимой: их влияние было опосредованно включено в итоговую модель через другие переменные. Для проверки наличия мультиколлинеарности был проведен анализ с помощью VIF, при котором строятся линейные регрессии зависимости одной из объясняющих переменных от остальных и в случае, если VIF() = >10, то существует мультиколлинеарность и переменную следует исключить из уравнения регрессии. Значение VIF > 10 было обнаружено для переменных fertility, health\_expend, comp\_education, consumption, country\_type\_developing. Но при удалении из уравнения регрессии данных переменных, а также переменных “unemployment” и “gdp\_per\_cap”, которые не были значимы на уровне 5%, нормированный R^2 оказался ниже, чем у исходной модели (0,24).

Однако при последовательном удалении переменных с самым высоким значением VIF ( consumption и comp\_education - значения 48,5 и 21,66 соответственно) и построении модели было получено, что модель без двух указанных переменных, во-первых, имеет нормированный R^2 равный 0,313, что выше исходной модели и показывает ее более высокое качество, а во-вторых, среди переменных значение VIF не было больше 10, то есть проблема мультиколлинеарности была решена удалением двух переменных. Поэтому мы посчитали целесообразным использовать данную модель для дальнейшего исследования. Оценки ее коэффициентов представлены в таблице 2.

| Переменная | коэффициент | стандартная ошибка | t-статистика | p-value |
| --- | --- | --- | --- | --- |
| const | 0.8439 | 1.117 | 0.755 | 0.452 |
| gdp\_growth | 0.0558 | 0.032 | 1.758 | 0.081 |
| nat\_res\_rent | 0.0253 | 0.017 | 1.510 | 0.134 |
| fertility | 0.4189 | 0.169 | 2.475 | 0.015 |
| health\_expend | 0.1317 | 0.066 | 2.000 | 0.048 |
| reg\_quality | -0.0754 | 0.307 | -0.245 | 0.807 |
| unemployment | -0.0140 | 0.035 | -0.396 | 0.693 |
| rural pop | -0.0278 | 0.010 | -2.869 | 0.005 |
| gdp\_per\_cap | -1.07e-05 | 1.25e-05 | -0.857 | 0.393 |
| country\_type\_developing | 1.9102 | 0.612 | 3.121 | 0.002 |

**Таблица 2** - Результат построения регрессионной модели без мультиколлинеарности

Таким образом, на уровне значимости в 5 процентов значимыми переменными оказались: fertility, health\_expend, rural\_pop, country\_type\_developing, на уровне 10% значима еще переменная gdp\_growth, но дальнейшее удаление переменных с высоким значением p-value приводило к снижению нормированного R^2, на основании этого было принято решение остановиться на использовании данной модели.

Так как R^2 модели составил 0,361, то включенные в модель переменные объясняют вариацию 36,1% дисперсии целевой переменной. Интерпретация коэффициентов будет проведена только для переменных значимых на уровне значимости 5%:

1. fertility - коэффициент перед переменной составил 0,4189, то есть при росте уровня фертильности в стране на 1, отношение дохода 10% наиболее обеспеченных к 50% наименее обеспеченных растет на 0,4189. Это может быть объяснено тем фактом, что такой рост рождаемости чаще всего происходит в наименее обеспеченных слоях населения, что приводит к росту населения с низким доходом и суммарная доля дохода для них снижается.
2. health\_expend - коэффициент перед переменной составил 0.132, то есть при росте доли расходов на здравоохранение на 1 п.п. происходит рост отношения дохода наиболее обеспеченных к наименее обеспеченным на 0,132. Данный результат выглядит контринтуитивным и противоречит изначальному предположению. Возможное объяснение может быть связано с неравномерным доступом населения к здравоохранению, то есть если большая часть расходов на медицину была направлена в районы, где живут наиболее обеспеченные слои населения, то качество их жизни улучшится и они смогут больше зарабатывать, что увеличит разрыв между богатыми и бедными
3. rural pop - коэффициент перед переменной равен -0.028, то есть при сокращении доли сельского населения на 1 процентный пункт отношение между доходами 10% наиболее обеспеченных к 50% наименее обеспеченных снижается на 0.028. Это подтверждает изначальную идею о более низком уровне дохода в селах, поэтому переезд в города сельских жителей может приводить к росту дохода наименее обеспеченного слоя населения
4. country\_type\_developing - коэффициент перед переменной равен 1,91, то есть если страна переходит из статуса развивающейся в статус развитой (так как 1- это развивающаяся страна,а 0 - развитая), то отношение дохода 10% наиболее обеспеченных к 50% наименее обеспеченным снижается на 1,9. Полностью соответствует изначальному предположению о более высоком уровне неравенства в развивающихся странах, что было показано и в рамках дескриптивного анализа

Был посчитан тест Чоу для модели при разделении выборки на две подвыборки - тех, которые показали высокий темп роста ВВП в 2021 году, и тех, которые показали низкий (тест Чоу может не быть показательным при разделении выборки по уровню развития стран из-за недостаточного количества развитых стран в датасете, поэтому было принято решение разделять по другой переменной). В качестве разделителя послужит медиана по данной переменной 5,44.

общая модель корректна

отдельные модели корректны

F-stat = F(k+1,n-2(k+1))

p-value для теста составила 0,99, то есть на любом разумном уровне значимости нулевая гипотеза о корректности общей модели не отвергается

Для проверки наличия в исходной модели гетероскедастичности был проведен тест Уайта:

: в модели отсутствует гетероскедастичность

: в модели имеет место гетероскедастичность. Было получено, что p-value теста составляет 0.31, то есть на любом разумном уровне значимости недостаточно оснований для отвержения нулевой гипотезы об отсутствии гетероскедастичности в пользу альтернативной.

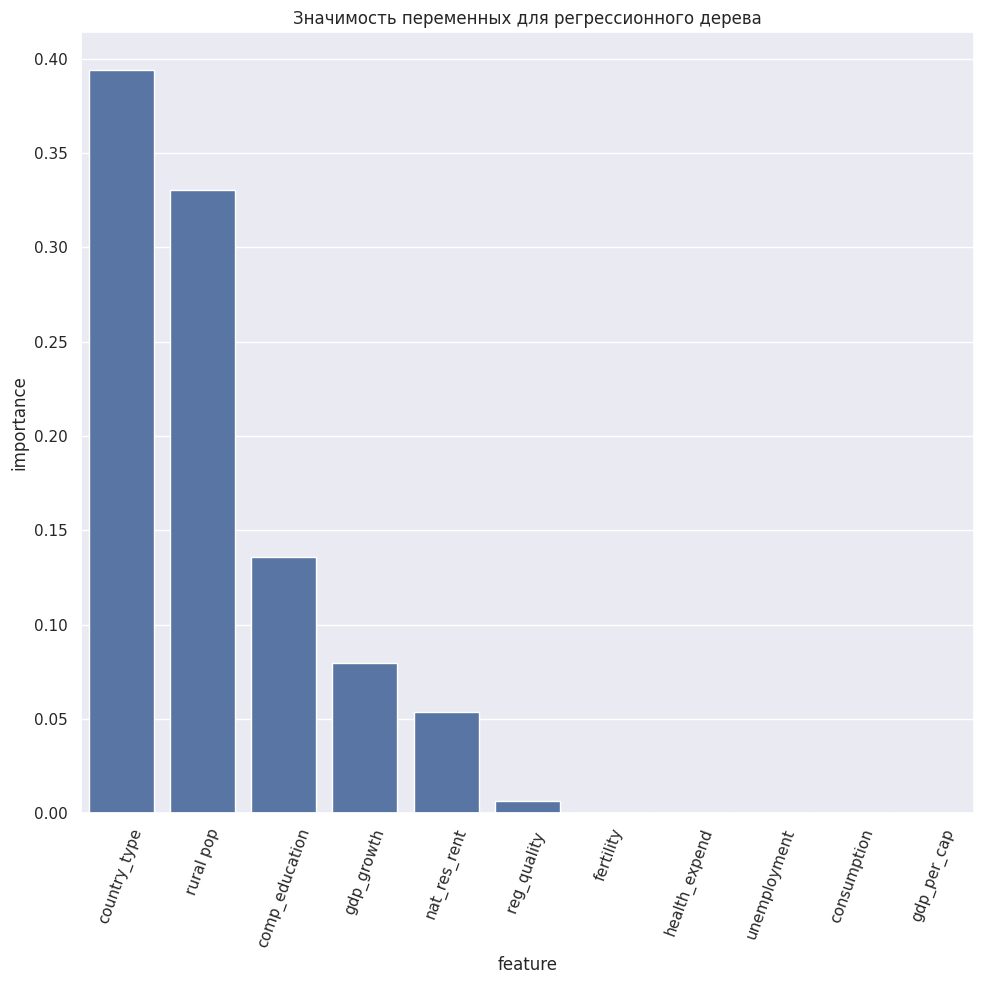
Проблем в модели может быть несколько. Во-первых, относительно малое количество наблюдений среди развитых стран не позволяет разделить построение модели отдельно по группам стран, чтобы выявить возможное существование других взаимосвязей для развитых стран. Решением могло бы быть добавление наблюдений по развитым странам за другие года, не только за 2021, однако это приводит к переходу к панельным данным, в рамках cross-section решение данного затруднения не представляется возможным. Во-вторых, количество рассматриваемых факторов при относительно небольшом размере датасета не позволяет проводить безболезненное включение в модель какого-либо значительного числа переменных в других степенях или переменных взаимодействия, так как это может привести к некорректному соотношению между количество наблюдений и переменных. Решением этого может быть переход к панельным данным или использование регрессионного дерева, благодаря которым можно отследить нелинейные взаимосвязь. Моделирование с использованием регрессионного дерева будет описано в следующей части.

# **Результаты второго выбранного метода**

В качестве второго метода анализа связи объясняющих переменных с отношением дохода 10% наиболее обеспеченного населения к 50% наименее обеспеченного было выбрано использование регрессионного дерева. Суть алгоритма заключается в итеративном разбиении пространства признаков на две части так, что снижение неопределенности благодаря разбиению будет наибольшим. То есть на каждом шаге алгоритм рассматривает всевозможные разбиения каждого из 11 признаков и осуществляет разбиение на две дочерние вершины так, чтобы был достигнуто максимальное снижение среднеквадратичной ошибки по сравнению с предыдущим шагом. Данная процедура будет повторяться до тех пор, пока в каждом листе не окажется только одна вершина или не будет выполнен критерий останова. Для избежания переобучения было принято решение использовать следующие критерии останова: максимальная глубина, минимальное количество наблюдений в листе.

Для оценивания качества моделирования исходная выборка была разделена на тренировочную и тестовую (70% всей выборки использовались для обучения, 30 - для оценки результата), категориальная переменная “country\_type” была бинарно закодирована. Для нахождения оптимального критерия останова использовался подбор параметров по сетке, включавшей в себя значения от 1 до 19 для максимальной глубины дерева и от 1 до 19 для минимального количества наблюдений в листе. Таким образом, для каждой пары параметров было составлено дерево и посчитана среднеквадратичная ошибка между предсказанным и реальным значением на тестовой выборке. . В результате были получены оптимальные значения критерий останова, при которых достигалось минимальное значение MSE на тестовой выборке: максимальная глубина дерева - 3, минимальное число наблюдений в листе - 2, при этом MSE на всей выборке составила 1.87, что ниже показателя для линейной регрессии.

Решающие деревья позволяют оценить важность признака для разбиения по тому, как часто разбивалась выборка по ним, и насколько рано тот или иной признак использовался для разбиения. Значимость переменных для объяснения целевой переменной представлена на рисунке 9.

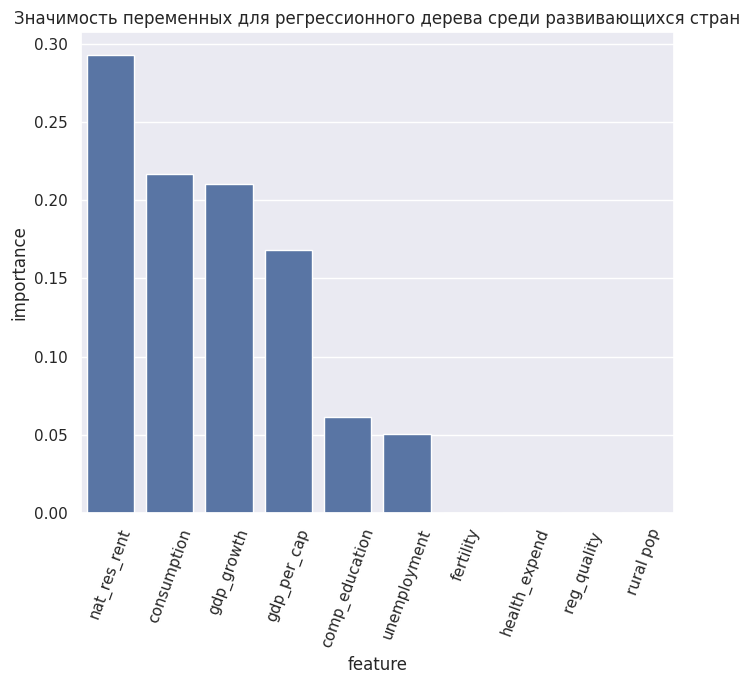


**Рисунок 9 - Значимость переменных для регрессионного дерева**

Таким образом, наибольшее значение для объяснения неравенства в стране играли уровень развития страны и доля сельского населения. Существует менее серьезная важность продолжительности обучения в стране, темп роста ВВП и доля ренты от полезных ископаемых в ВВП, при этом качество правительства практически не играло роль. Уровень фертильности, расходы на здравоохранение, уровень безработицы, потребления, ВВП на душу населения никак не влияли на разбиение.

Стоит отметить, что переменные “comp\_education”,”nat\_res\_rent” и reg\_quality не были значимы при проведении регрессионного анализа.

Это стало причиной решения провести дополнительный анализ для развивающихся стран, значимость переменных после которого представлена на рисунке 10.



**Рисунок 10 - Значимость переменных для регрессионного дерева среди развивающихся стран**

В результате было получено, что наибольшую значимость для объяснения неравенства для данной группы стран имеет доля природных ресурсов в ВВП, также значимы переменные, отражающие экономическое развитие (ВВП на душу населения, темп роста ВВП, потребление, безработица), меньшую значимость имеет фактор развития человеческого капитала - количество лет обязательного образования. Другие социальные и политические факторы влияния на результат не оказывали. Отдельная линейная регрессионная модель для группы развивающихся стран не показала значимость части переменных, указанных для регрессионного дерева. Это может объясняться тем фактом, что решающие деревья способны отслеживать нелинейные взаимосвязи, в то время как линейные модели без добавления переменных пересечения - нет, при этом данные переменные могут быть и высокого порядка, что затрудняет возможность их обнаружения.

# **Заключение**

В рамках исследования факторов, влияющих на уровень неравенства по доходу в стране, на основе данных Всемирного Банка и World Inequality Database было выявлено, какие из показателей являются значимыми для объяснения отношения дохода 10% наиболее обеспеченного к 50% наименее обеспеченного населения. Регрессионный анализ - использование метода наименьших квадратов для построения множественной линейной регрессии, и использование Variance Inflation Factor показали высокий уровень мультиколлинеарности между переменными, но после удаления части из них получилось достигнуть более высокой объясняющей способности модели. Значимыми оказались расходы на здравоохранение (при этом изначальное предположение о влиянии роста расходов на него на снижение неравенства не подтвердилось), доля сельского населения и уровень фертильности. Но наибольшее влияние на неравенство оказывает принадлежность к стране к определенной группе по уровню развития, это может быть объяснено более высоким качеством показателей, не включенных в данный анализ, которые характеризуют развитые страны, что может вести к значительному снижению неравенства.

Использование регрессионного дерева - итерационное разбиение выборки по переменной, ведущей к наибольшему снижению хаотичности, показал значимость в том числе и других факторов. Это может быть объяснено более высокой способностью регрессионного дерева улавливать нелинейные закономерности, в то время как регрессионный анализ проводился без добавления дополнительных переменных из-за ограниченности выборки. Изначальная гипотеза о равноценном влиянии группы социальных, экономических и политических факторов, не подтвердилась. Для выборки из всех стран качество политических институтов является практически незначимым, а из группы социальных факторов учитывается только продолжительность обязательного образования. Наибольшее воздействие на уровень неравенства по доходу в стране оказывают экономические факторы и, в первую очередь, уровень экономического развития страны, а в рамках одной группы стран - продолжительность обязательного образования и уровень безработицы менее значимы, чем доля природных ресурсов в ВВП и темп экономического роста.

Исследование при расширении набора данных может быть теоретически применено для разработки мер и выявлению основных факторов, которые могут привести к снижению неравенства при изменении экономической активности государства.

# **Список литературы:**

1. Afonso, António and Schuknecht, Ludger and Tanzi, Vito, Income Distribution Determinants and Public Spending Efficiency (May 26, 2010). Journal of Economic Inequality, Vol. 6, No. 3, pp. 367-389, 2010, Available at SSRN: <https://ssrn.com/abstract=1615962> (дата обращения 10.12.2023)
2. Malerba, G., & Spreafico, M. (2014). STRUCTURAL DETERMINANTS OF INCOME INEQUALITY IN THE EUROPEAN UNION: EVIDENCE FROM A PANEL ANALYSIS. *Rivista Internazionale Di Scienze Sociali*, *122*(1), 37–83. <http://www.jstor.org/stable/43830200> (дата обращения 10.12.2023)
3. World Development Indicators, Compulsory education duration (years): <https://data.worldbank.org/indicator/SE.COM.DURS?view=chart> (дата обращения 07.09.2023)
4. World Development Indicators, Current health expenditure (% of GDP): <https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS?view=chart> (дата обращения 07.09.2023)
5. World Development Indicators, Fertility rate, total (births per woman): <https://data.worldbank.org/indicator/SP.DYN.TFRT.IN?view=chart> (дата обращения 07.09.2023)
6. World Development Indicators, Final consumption expenditure (% of GDP): <https://data.worldbank.org/indicator/NE.CON.TOTL.ZS?view=chart> (дата обращения 07.09.2023)
7. World Development Indicators, GDP growth (annual %): <https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?view=chart> (дата обращения 07.09.2023)
8. World Development Indicators, GDP per capita (current US$): <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?view=chart> (дата обращения 07.09.2023)
9. World Development Indicators, Regulatory Quality: Estimate: <https://data.worldbank.org/indicator/RQ.EST?view=chart> (дата обращения 07.09.2023)
10. World Development Indicators, Rural population (% of total population): <https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?view=chart> (дата обращения 07.09.2023)
11. World Development Indicators, Total natural resources rent (% of GDP) <https://data.worldbank.org/indicator/NY.GDP.TOTL.RT.ZS?view=chart> (дата обращения 07.09.2023)
12. World Development Indicators, Unemployment, total (% of working force) (modeled ILO estimate): <https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?view=chart> (дата обращения 07.09.2023)
13. World Inequality Database: [Data - WID - World Inequality Database](https://wid.world/data/) (дата обращения 07.09.2023)