НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Департамент статистики и анализа данных
факультета экономических наук

Многомерные статистические методы
2020-21 гг. – 2 курс

Итоговый отчёт по компьютерным работам

Выполнили:
студенты групп БСТ195 и БСТ196
Гильфанова Камиля

Москва 2021

Содержание

1 Постановка задачи

В рамках исследования основная задача была сформулирована следующим образом: выявление наличия и характера влияния различных показателей на продолжительность жизни населения в различных странах мира. Имеющиеся показатели описывают благосостояние населения (в том числе в области образования и здравоохранения) и государственные меры, направленные на улучшение этого благосостояния.

Хотя на текущий момент уже было проведено множество исследований факторов, влияющих на продолжительность жизни, тем не менее, было обнаружено что влияние иммунизации населения не принималось во внимание в достаточной мере. Следовательно, есть смысл посмотреть, насколько велико влияние таких переменных на продолжительность жизни населения.

Для анализа был взят массив данных по продолжительности жизни с сайта kaggle[1], содержащий данные по 193 уникальным странам за период с 2000 до 2015 года. В этом массиве, содержащем данные, собранные Всемирной Организацией Здравоохранения (WHO), включающем различные показатели, которые предположительно влияют на продолжительность жизни, изначально находится 2938 наблюдений по 22 уникальным переменным, 20 из которых – числовые. В рамках работы мы взяли 9 переменных с условно непрерывными распределениями (обоснованность взятия именно этих переменных будет представлена далее). Таким образом, финальный набор данных имеет 183 взаимно независимых наблюдения. Что касается репрезентативности выборки, то данные были получены по 183 странам, в числе которых как развитые страны (США, Великобритания, Франция), так и развивающиеся (Россия, Казахстан, Чили) и наименее развитые (Гамбия, Мьянма, Судан) страны[2]. Следовательно, с достаточной степенью уверенности можно говорить, что взятые данные хорошо отражают генеральную совокупность, так как все категории стран представлены в близких к реалиям пропорциях, особенно если учесть, что ООН обычно публикует данные по 195 странам (с учётом частично признанных государств). По условию задания мы не могли взять временные ряды, поэтому мы зафиксировали для анализа тот год, который, во-первых, имеет статистическую значимость для текущего времени и, во-вторых, имеет наиболее "чистые" с точки зрения статистики данные. Таким годом оказался 2014 год. При отборе переменных также учитывалось то, что для многих статистических методов предпочтителен нормальный (или логнормальный, эффективно сводящийся к нормальному) закон распределения. Особое внимание мы также уделили пропускам в данных и нулям. Также отслеживалось то, чтобы все переменные имели более-менее равноценный вклад в зависимую переменную на уровне здравого смысла. В результате такого анализа показателей из 22 переменных были оставлены 9 (и ещё 1 дамми-переменная, показывающая название страны).

Теперь необходимо описать, что именно характеризуют отобранные переменные (полужирным указано то, как эти переменные обозначаются в коде, выводы которого используются в ходе отчёта):

1. *Life expectancy* (предположительно зависимая): **Life_exp**. Варьируется в пределах от 36.3 до 89.

2. *Adult mortality*: **Adult_mort**. Показатель смертности среди взрослых обоих полов: численность смертей среди населения в возрасте от 15 до 60 лет на 1000 человек.

3. *Alcohol*: **Alcohol**. Потребления алкоголя на душу населения в возрасте от 15 лет (в литрах чистого спирта).

[1] www.kaggle.com/kumarajarshi/life-expectancy-who

[2] по классификации МВФ, которая отличается от классификации ООН

4. *Total expenditure*: **Total_exp**. Общие государственные расходы на здравоохранение как процент от общих государственных расходов (%).

5. *GDP*: **GDP**. Значение ВВП на душу населения (в долларах США).

6. *Diphtheria*: **Dipht**. Охват иммунизации от столбняка, дифтерии и коклюша среди детей в возрасте 1 года – одни из наиболее смертельных болезней для новорождённых[3] (%).

7. *Thinness 5-9*: **Thinnes_5_9**. Истощение среди детей и подростков в возрасте от 5 до 9 лет (%).

8. *Thinness 10-19*: **Thinnes_10_19**. Истощение среди детей и подростков в возрасте от 10 до 19 лет (%).

9. *Schooling*: **Schooling**. Количество лет, затраченных на обучения (в годах).

Распределения выбранных переменных будут проанализированы и продемонстрированы в следующих этапах работы (см. секцию Описание переменных). Также в этой секции будут рассмотрены проблемы распределений некоторых не вошедших в итоговый набор переменных.

2 Выдвижение рабочих гипотез исследования

Для начала стоит сказать о некоторых предположениях, которые принимались в ходе написания компьютерных работ. Они касаются того, что некоторые распределения, которые формально (по тесту Шапиро-Уилка, например) слабо подчиняются нормальному закону, были, тем не менее, приняты за нормальные. Также в ряде разделов мы принимали, что переменные слабо коррелированны. В частности, к примеру, для проведения дискриминантного анализа рекомендуется сначала, помимо нормальности переменных, проверить их корреляционную зависимость и однородность дисперсий внутри переменных (м-статистика Бокса).

Основной гипотезой, рассматриваемой в работе, является предположение о том, что все выбранные нами переменные объясняют вариацию зависимой переменной *Life expectancy*, то есть существует тесная связь между зависимой переменной (продолжительность жизни) и объясняющими переменными. Также предполагается, что наибольший вклад в изменчивость зависимой переменной по объективным причинам будут вносить *GDP*, *Total expenditure* и *Adult mortality*: эти показатели наиболее явным образом характеризуют развитую страну. Очевидно, что чем более развитая страна, тем выше, по идее, должна быть продолжительность жизни. С другой стороны, в n-мерном признаковом пространстве методами статистического анализа можно прийти к неявным выводам, по которым, возможно, прочие переменные будут лучше объяснять вариацию зависимой переменной – впрочем, это возможно проверить только в ходе работы.

3 Описание переменных

Отбор компонент признакового пространства опирался, помимо интерпретационной силы переменных, на распределения переменных. Так, в ходе анализа имеющихся переменных были выявлены следующие проблемы:

- атипичность (например, бимодальность) и далёкость распределений по некоторым переменным от нормального закона распределения[4]. К таким переменным относятся, например, переменные, описывающие ожирение и заболеваемость СПИДом и корью;

[3] https://www.who.int/ru/news-room/fact-sheets/detail/the-top-10-causes-of-death

[4] Достаточно большое число методов анализа данных опираются на нормальность данных или близость к нормальному закону.

- наличие большого количества пропусков по отдельным переменным;
- некорректные значения по переменной, содержащей данные о ВВП на душу населения по странам.

Так, например, выглядит диаграмма вида ящик с усами и гистограммы распределений некоторых переменных:

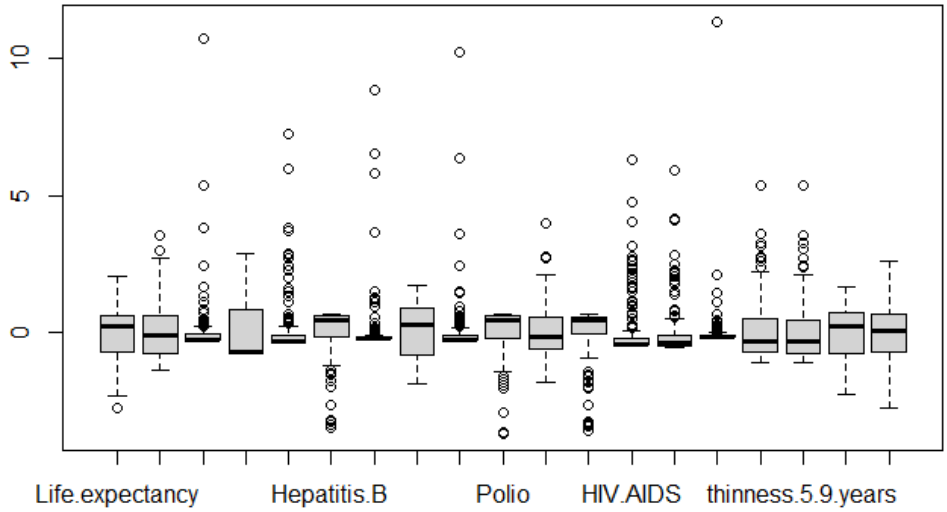

Рис. 1: Выбросы по всем исходным переменным

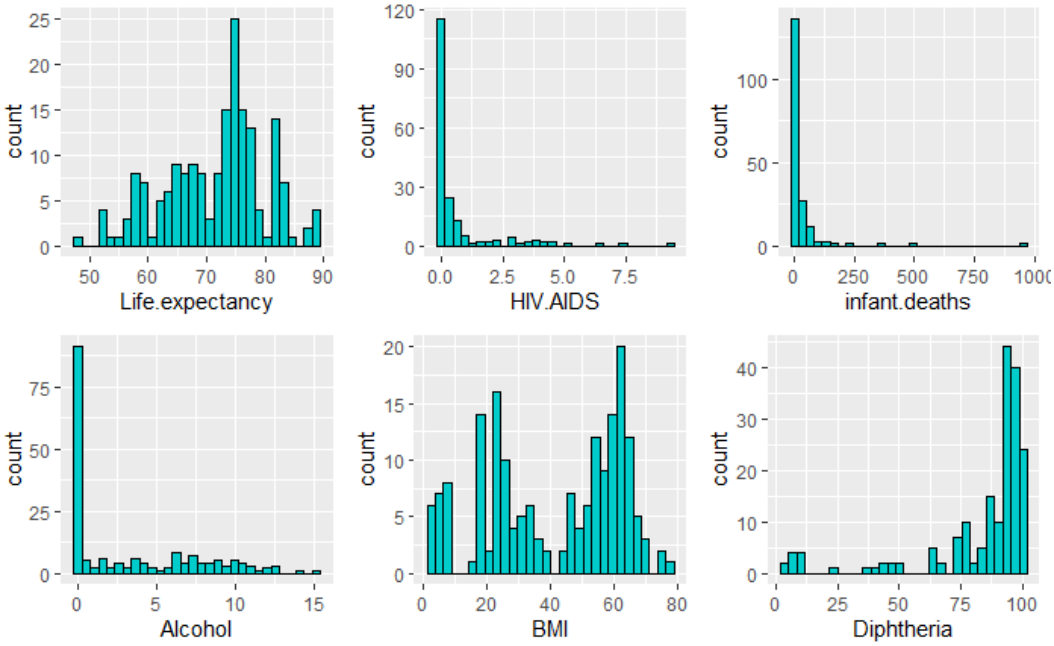

Рис. 2: Распределения исходных переменных

Для разрешения этих проблем производился более тщательный сортинг переменных. Также были найдены показатели ВВП на стороннем надёжном источнике[5] и обновлён соответствующий столбец в наборе данных. На этом же сайте (и сайтах ООН) были найдены отсутствующие данные по другим переменным по тем или иным странам, доступ к данным которых

[5] data.worldbank.org/indicator/NY.GDP.PCAP.CD

трудно осуществить (особенно это касается частично и недавно признанных государств).

В итоге был отобран набор данных, состоящий из 8 независимых переменных и одной результирующей с меньшим числом выбросов и более адекватными типами распределений. В результате применения логарифмирования, которое было уместным уже на начале исследования, также удалось стабилизировать дисперсию, и число выбросов сократилось ещё сильнее.

3.1 Описание случайной величины

В качестве случайной величины была выбрана переменная *Life expectancy*, показывающая продолжительность жизни в годах в той или иной стране. Визуальный анализ показал, что данная независимая переменная очень слабо напоминает нормальное распределение. Ядерная оценка плотности очень далека от идеального Гауссовского колокольчика.

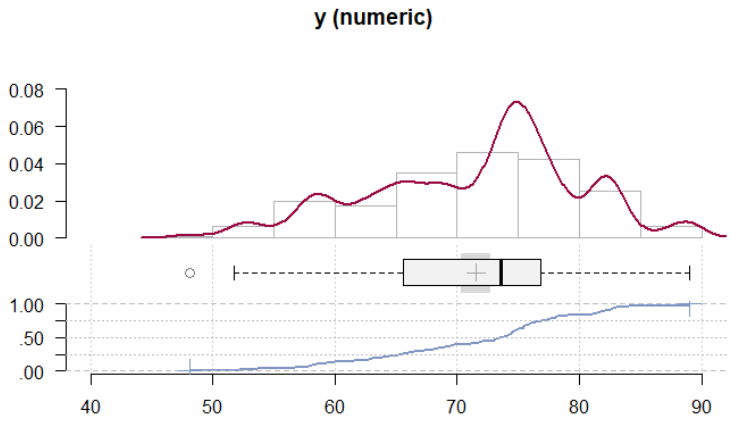

Рис. 3: Гистограмма распределения, ядерная оценка плотности, ящик с усами и огива СВ

Характеристики центра, рассчитанные для СВ, отличаются друг от друга, но не сильно. Мода и медиана: $Mo_y^* = 76.3$, $Me_y^* = 73.6$ – отличаются примерно на 3 единицы измерения, если ориентироваться на значение, полученное для моды вручную. Выборочное матожидание $\bar{y} = 71.54$ (относительно общего диапазона) отличается в меньшую сторону. Это, во-первых, говорит о том, что, поскольку $\bar{y} < Me_y$, выборка обладает **левосторонней асимметрией**, во-вторых, поскольку для нормального закона свойственно то, что характеристики центра – мода, медиана и матожидание – примерно равны друг другу, то имеющееся распределение данных, будучи асимметричным, не вписывается в эту парадигму.

Интерквартильный размах $IQR = 11.25$: в интервале такой ширины содержатся центральные 50% наблюдений выборки. Квартильное отклонение, устойчивый к выбросам аналог размаха, показывает значение $Q = 5.625$, причём значение разброса $R = 40.9$, что намного (в 8 раз) больше. То, что квартильное отклонение $Q = 5.625$, а $\frac{2}{3}\sigma =$ ‘$rsd(y) * (2/3)$‘, показывает, что распределение можно считать **умеренно асимметричным**, поскольку эти значения примерно одинаковы.

Выборочная (исправленная) дисперсия $s^2 = 73.3$, выборочное СКО $s = 8.56$ – довольно небольшое относительно выборки, в принципе, можно сказать – и визуальный анализ это подтверждает, – что значения сконцентрированы примерно около выборочного среднего. Неустойчивый к выбросам коэффициент вариации, показывающий меру рассеяния исследуемого признака, $V_s = 11.97\% < 33\%$, следовательно, выборку можно охарактеризовать как однородную.

Исследованные характеристики разброса позволяют сделать вывод, что либо в выборке **совсем не содержится выбросов**, либо их мало и они незначительно отличаются от общепринятых метрик, например, от $3IQR$. Коэффициент асимметрии $Ac^* = -0.37$ показывает, что имеет место левосторонняя асимметрия и, несмотря на то, что $|Ac^*| < 0.5$, коэффициент асимметрии довольно близок по своему значению к пороговому, следовательно, следует говорить о **достаточной асимметрии** гистограммы (и именно это показывало построение графических интерпретаций имеющихся данных). Коэффициент эксцесса $Ek^* = -0.38$, что показывает следующее: во-первых, так как $Ek^* < 0$ и $|Ek^*| < 0.5$ (хоть и близок к этому значению), график распределения имеет **более плоскую вершину** относительно графика плотности нормального распределения.

Суммируя все выкладки и проведённую в R диагностику выбросов, можно сделать следующее краткое заключение: выборка – и это связано со спецификой обрабатываемых данных, а именно – возраста человека, который ограничен по значению – **однородна**, **имеет мало** (или совсем не имеет) **выбросов**, значения внутри выборки **не сильно варьируются**. Более того, **логарифмирование** некоторых переменных позволяет стабилизировать распределение и дисперсию в этих переменных. Распределение имеет левостороннюю асимметрию, что означает, что средние значения продолжительности жизни (интервал примерно в 11 лет) сконцентрированы – об этом же свидетельствует значение моды и медианы – около примерно 70+ лет, и это хороший, довольно большой возраст. Значит, в целом человеческая цивилизация успешно добивается увеличения продолжительности жизни, и в дальнейшем будет проанализировано, какие именно факторы влияют на достижение такого результата.

Итак, в результате некоторых преобразований данных удалось улучшить их качество для дальнейших манипуляций, **не прибегая при этом к удалению пропусков в данных и даже выбросов** и сохранив исходную выборку практически в полном составе. Визуально оценить изменения, произошедшие с распределениями переменных, можно на следующих графиках:

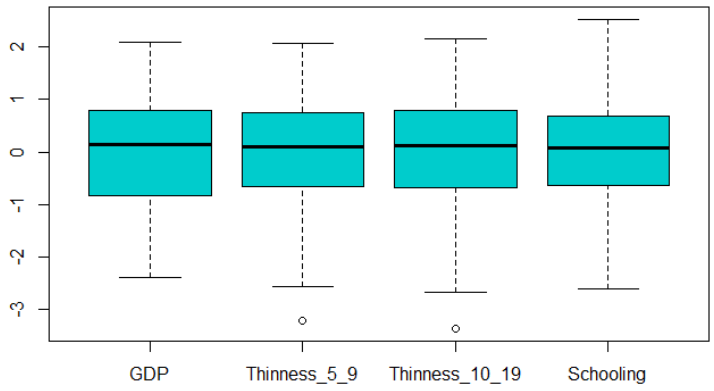

Рис. 4: Ящик с усами для некоторых независимых переменных

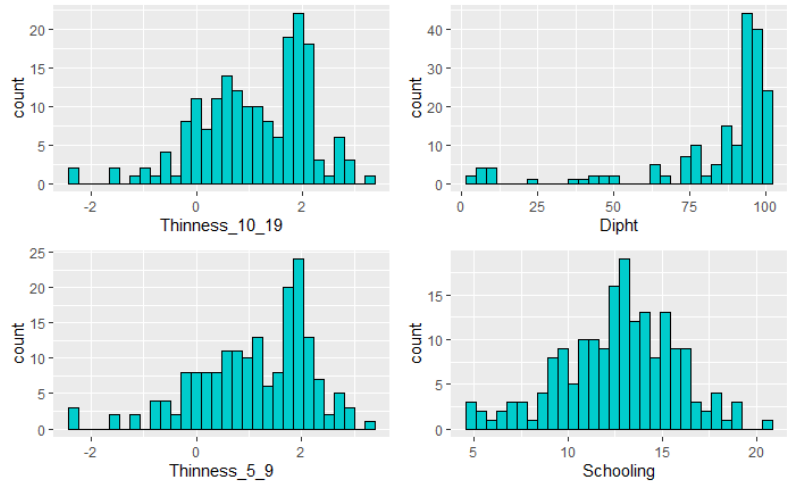

Рис. 5: Гистограммы некоторых независимых переменных

3.2 Проверка соответствия эмпирического распределения нормальному закону

Ранее формально было показано, что численные характеристики коэффициентов асимметрии и эксцесса для результирующей переменной свидетельствуют о несущественной левосторонней асимметрии и некоторой плосковершинности относительно кривой Гаусса, а примерного равенства выборочных моды, медианы и среднего, свойственного нормальному закону распределения, у исследуемой зависимой переменной нет. Так что есть основания сомневаться в нормальности распределения этой переменной. Графические методы показывают, например, такую картину:

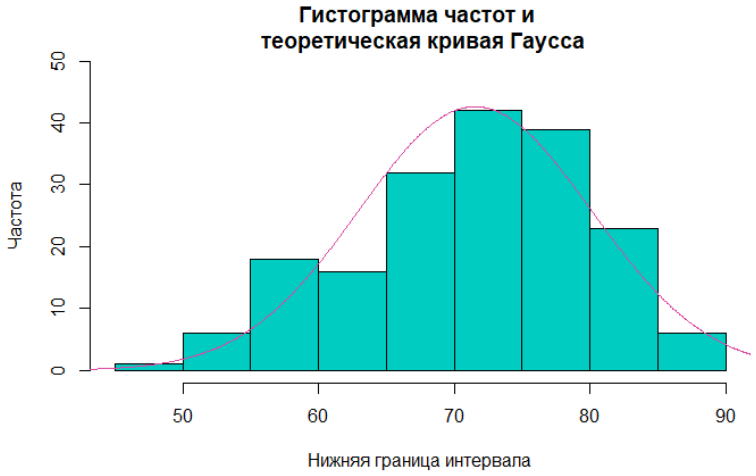

Рис. 6: Гистограмма частот с нанесённой поверх теоретической кривой Гаусса

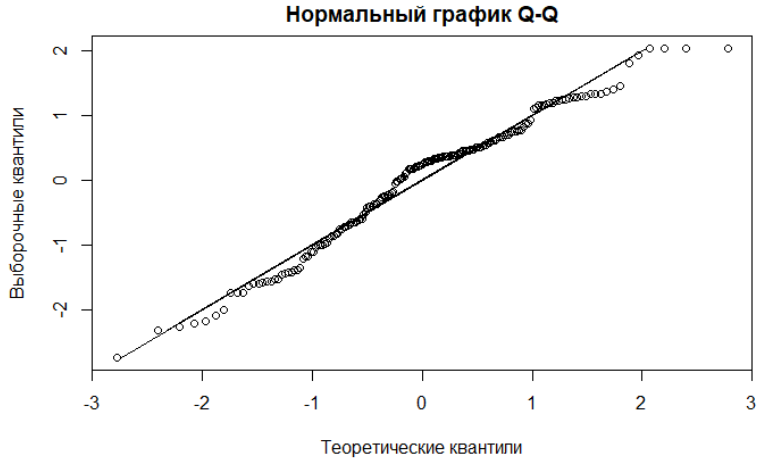

Рис. 7: График квантиль-квантиль

Видно всё то, что описывалось ранее: асимметрия и плосковершинность. Также заметно, что результирующая переменная неплохо приближается теоретической кривой Гаусса, хотя и не идеально.По графику квантиль-квантиль видно, что данные группируются вдоль прямой теоретического квантиля не очень хорошо, имеется много отклонений. Местами точки эмпирического распределения близки к прямой под углом 45 градусов на графике, что говорило бы в пользу нормальности распределения, однако есть области, где скачки и расхождения серьёзны.

Формальные методы – а именно, тесты на нормальность с помощью критериев Шапиро-Уилка, Шапиро-Франчиа, Лиллиефорса (обоснованность применения того или иного критерия показана в компьютерной работе) – показали, что имеющуюся результирующую переменную можно с большими оговорками считать похожей на нормальную: для самого мощного теста Шапиро-Уилка $p - value = 0.005046$. При пессимистичном анализе, разумеется, по многим причинам (как графическим, так и аналитическим) гипотезу о нормальности независимой переменной нужно отвернуть. Тем не менее, p-value получился не самым маленьким из возможных.

Также были рассмотрены распределения независимых переменных. Среди них наиболее близким к нормальному распределением обладает переменная $Schooling$ ($p - value = 0.21$). Относительно неплохие результаты у тех переменных, что были прологарифмированы. Очевидно сильное отклонение от нормального закона имеют переменные $Alcohol$, $Dipht$ и $Adult_mort$ – то, что мы и предполагали ранее.

4 Корреляционный анализ

На этом этапе работы был проведен корреляционный анализ, перед которым было осуществлено выявление выбросов согласно правилу $3IQR$ для исключения искажений рассчитываемых показателей. Выбросы, в результате предобработки данных (логарифмирования некоторых переменных), были обнаружены только по переменной *Dipht*. Далее проводится сравнение полей корреляции по выборкам до и после удаления выбросов. В обоих случаях построены графики, демонстрирующие положительную, обратную и отсутствующую зависимости между переменными, чтобы проиллюстрировать все возможные варианты взаимосвязи. Также было построено облако корреляции по переменной *Dipht*, чтобы отобразить изменение на диаграмме рассеивания в случае до удаления выбросов и после (так как выбросы были удалены только по переменной *Dipht*):

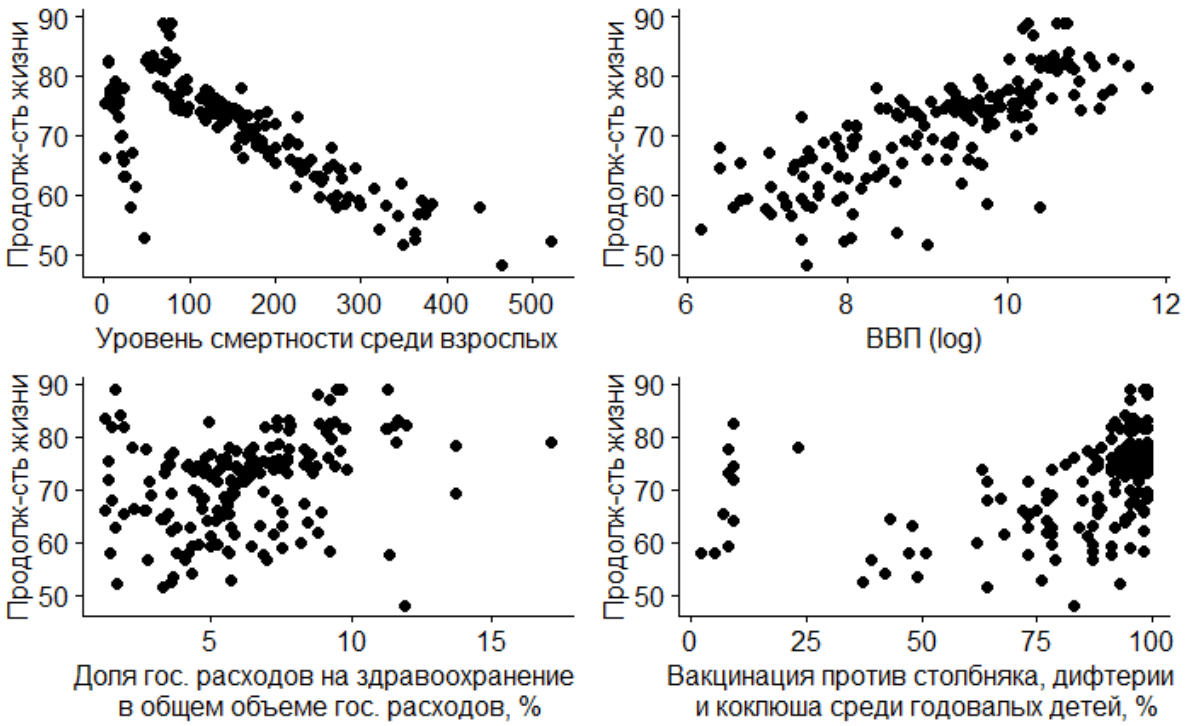

Рис. 8: Облака корреляции по переменным *Adult_ mort*, *GDP*, *Total_ exp*, *Dipht* до удаления выбросов

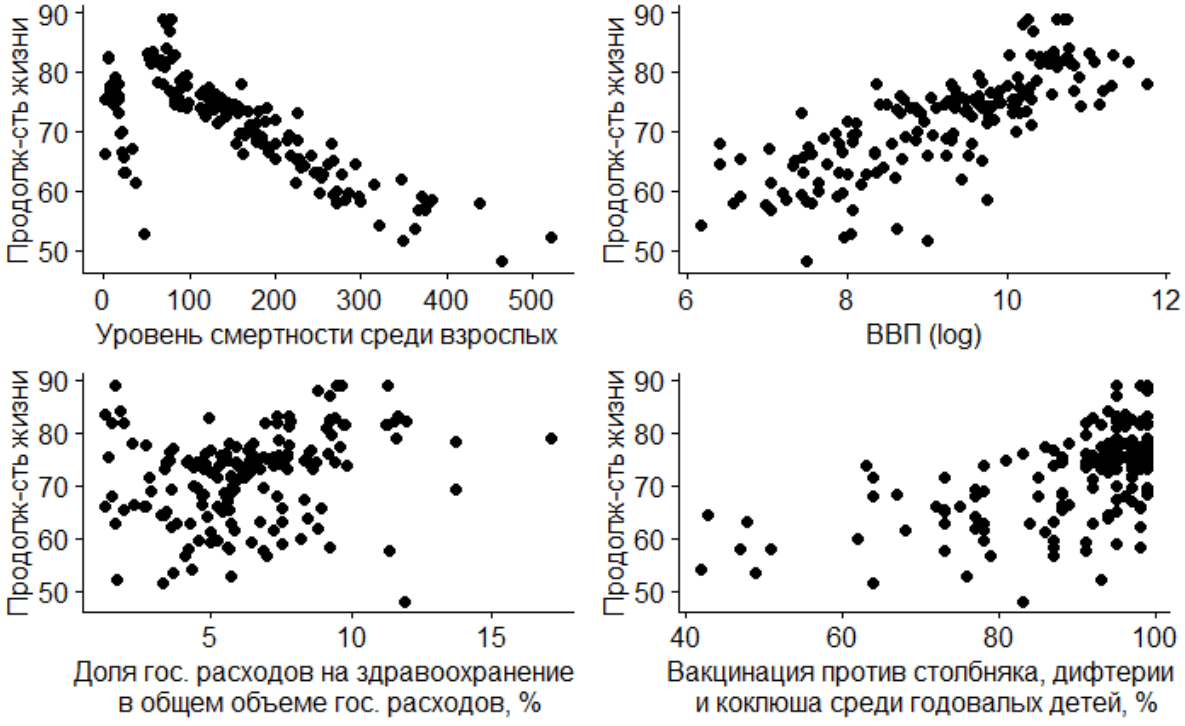

Рис. 9: Облака корреляции по переменным *Adult_ mort*, *GDP*, *Total_ exp*, *Dipht* до удаления выбросов

В результате сравнения диаграмм рассеяния отчетливо видно, что первые три диаграммы для данных после удаления выбросов идентичны тем, что были построены в случае до удаления выбросов, и они отображают обратную, прямую и отсутствующую зависимость соответственно. Видим, что последняя диаграмма имеет немного другой вид по сравнению со случаем до удаления выбросов из-за ставшего куда менее заметным кластера точек, находящегося ближе к началу координат – из-за этого угол наклона облака точек стал более

пологим, а само облако – более рассеянным.

Следующим этапом проведения корреляционного анализа стало построение и интерпретация матрицы парных коэффициентов корреляции. Сравнение матриц для выборок до и после удаления выбросов продемонстрирует, насколько на результаты влияет удаление выбросов.

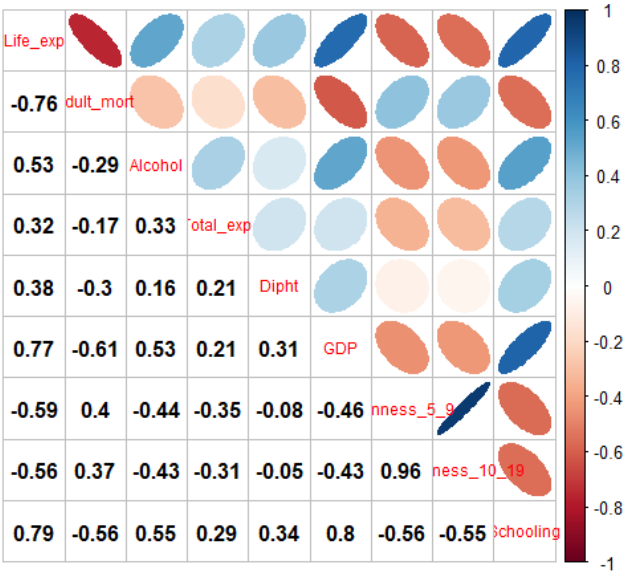

Рис. 10: Корреляционная матрица до удаления выбросов

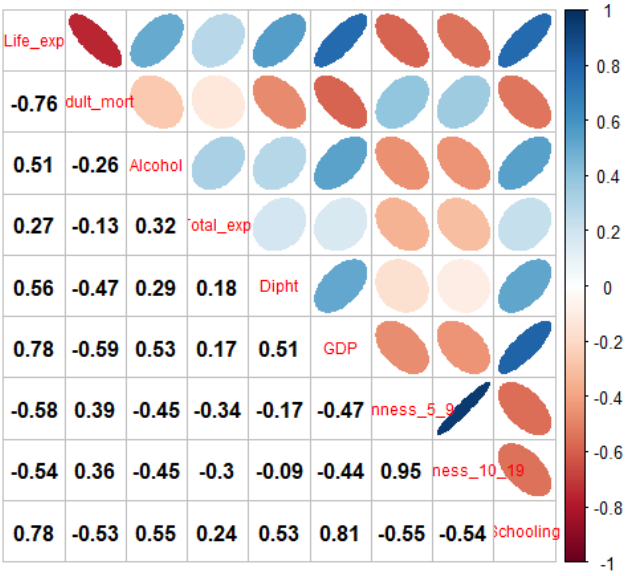

Рис. 11: Корреляционная матрица после удаления выбросов

В обоих случаях наиболее сильная положительная связь наблюдается между переменными *Thinness_5_9* и *Thinness_10_19* (0,96 до удаления выбросов и 0,95 - после). Это вполне логично, учитывая, что признаки одинаковы по своему смыслу и лишь применяются для разных возрастных категорий (5-9 лет и 10-19 лет соответственно). Также совпадают коэффициенты корреляции, отражающие наиболее сильную обратную связь: это коэффициенты для переменных *Life_exp* и *Adult_mort* ($\hat{\rho} = -0{,}76$), что также разумно, при том что продолжительность жизни до определенного момента тем меньше, чем больше вероятность смерти в возрасте от 5 до 60 лет. Кроме того, необходимо заметить, что до и после удаления выбросов связь между зависимой переменной *Life_exp* и объясняющими переменными в большинстве случаев достаточно сильная. Такая высокая степень сходства результатов для выборок до и

после удаления выбросов может объясняться тем, что выбросы были обнаружены лишь по одной переменной (*Dipht*). Поэтому все изменения после удаления выбросов наиболее сильно отражаются на коэффициентах, связанных с *Dipht*. Таким образом, в большинстве случаев положительная и обратная связь между переменными стала незначительно слабее после удаления выбросов, однако в нескольких случаях связь незначительно возросла по модулю. До удаления выбросов связь между *Life_ exp* и *Total_ exp*, а также *Life_ exp* и *Dipht* является умеренной, тогда как после удаления выбросов связь между *Life_ exp* и *Total_ exp* стала слабой, а связь между *Life_ exp* и *Dipht* стала средней. И положительная, и обратная связь между переменной *Dipht* и остальными переменными стала сильнее.

Расчет коэффициентов корреляции важен, однако не менее важна проверка их значимости, что и было осуществлено при помощи комбинации оценки корреляционной матрицы с результатами тестов на значимость:

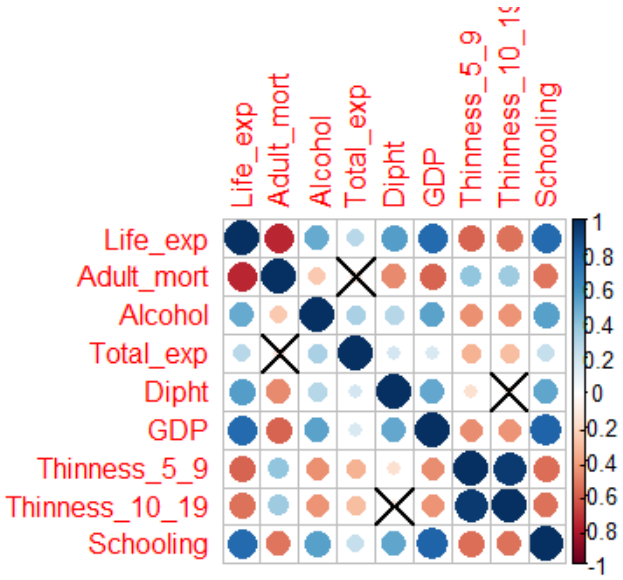

Рис. 12: Корреляционная матрица со статистически незначимыми коэффициентами

Статистически незначимыми определены связи между *Thinness_ 10_ 19* и *Dipht*, а также между *Total_ exp* и *Adult_ mort* являются статистически незначимыми.

Далее было проведено построение доверительных интервалов для парных коэффициентов корреляции. Если такой интервал включает в себя 0, то коэффициент является статистически незначимым. Полученные результаты совпадают с проведенной ранее проверкой на значимость коэффициентов: коэффициент корреляции между *Total_ exp* и *Adult_ mort* $\hat{\rho} \in (-0.27; 0.0228)$, коэффициент корреляции между *Dipht* и *Thinness_ 10_ 19* $\hat{\rho} \in (-0.24; 0.06)$.

Следующим шагом в проведении корреляционного анализа стало рассмотрение частных коэффициентов корреляции, показывающих тесноту линейной зависимости между двумя переменными при исключении влияния всех остальных показателей, входящих в модель. Матрица частных коэффициентов корреляции позволяет увидеть силу и направление связи между двумя переменными при фиксированном влиянии остальных переменных. Наиболее сильная положительная связь наблюдается между переменными *Thinness_ 10_ 19* и *Thinness_ 5_ 9* (значение коэффициента $\hat{\rho} = 0{,}92$). Наиболее сильная обратная зависимость наблюдается между переменными *Life_ exp* и *Adult_ mort* (коэффициент $\hat{\rho} = -0.57$: средняя связь). В большом числе случаев значения частных коэффициентов корреляции меньше, чем значения парных, следовательно, **прочие переменные, которые фиксировались при вычислении частного коэффициента корреляции, усиливают связь между переменными**. Например, это особенно хорошо иллюстрируется между *GDP* и *Schooling*.

Парный коэффициент корреляции между *GDP* и *Schooling* составлял $\hat{\rho} = 0.81$, но частный коэффициент корреляции между этими переменными меньше практически в 2 раза ($\hat{\rho} = 0.46$). Следовательно, на связь между этими переменными оказывают влияние и прочие факторы.

Как и в случае с парными коэффициентами корреляции, были построены доверительные интервалы, позволяющие увидеть, в каком диапазоне находится значение коэффициента, вероятность ошибки I рода которого составляет 5%, а также проведена проверка на значимость частных коэффициентов корреляции. Если p-value больше 0.05, то коэфициент является незначимым на уровне значимости 0.95. Многие коэффициенты оказались статистически незначимыми, гораздо больше, чем в случае с парными коэффициентами. И это многое говорит о выборке в целом: значит, **связь между переменными гораздо сильнее на фоне остальных переменных**, с учётом их влияния.

Проверка на значимость (результат вывода рассчитывающей значимость функции – довольно большой массив чисел – содержится в соответствующем разделе компьютерной работы) показала, что в большинстве случаев парные коэффициенты корреляции превышают значения частных, а влияние остальных переменных усиливает связь между двумя переменными.

Наконец, был рассчитан множественный коэффициент корреляции, отражающий тесноту линейной связи между переменной *Life_ exp* и массивом остальных переменных. Коэффициент составляет примерно $\hat{\rho} = 0.9$, что показывает довольно сильную связь, а также является статистически значимым в результате проведения F-теста.

5 Регрессионный анализ

Ранее был проведен предварительный анализ данных, также к данным были применены методы корреляционного анализа с целью выявления направления и силы линейных связей между переменными. Теперь стало возможным перейти к регрессионному анализу непосредственно. Отметим, что при построении линейных моделей регрессии работа велась с массивом данных, в котором на этапе корреляционного анализа уже были обработаны выбросы. Это позволит сделать наиболее правдоподобные выводы о построенных моделях регрессии.

5.1 Построение двумерных моделей линейной регрессии

В качестве анализа значимых переменных, которые будет иметь смысл включать в двумерные и, впоследствии, в множественную модель линейной регрессии, нами была построена множественная модель, включающая в себя все переменные. Выяснилось, что значимыми являются только переменные *Adult_ mort*, *Dipht*, *GDP* и *Schooling*. Соответственно, было построено 4 двумерные модели линейной регрессии с каждой из этих переменных.

Для первой модели ($M(y|x) = \hat{y} = 80.71 - 0.06x$) коэффициент детерминации был равен 0,5723, следовательно, смертность взрослых (*Adult_ mort*) объясняет 57.23% (что не очень много) вариации продолжительности жизни. Значение p-value $<$ 2.2e-16, что гораздо меньше всякого разумного уровня значимости, следовательно, нулевая гипотеза $H_0 : \beta_1 = 0$ (т.е. гипотеза о том, что независимая переменная не объясняет зависимую), отвергается. Был построен график для оцененных уравнений регрессии с регрессором (с аргументом в виде оценки модели с этим регрессором), на котором видно, что модель не учитывает все особенности зависимости (также было выявлено то, что остатки отклоняются от нормального распределения):

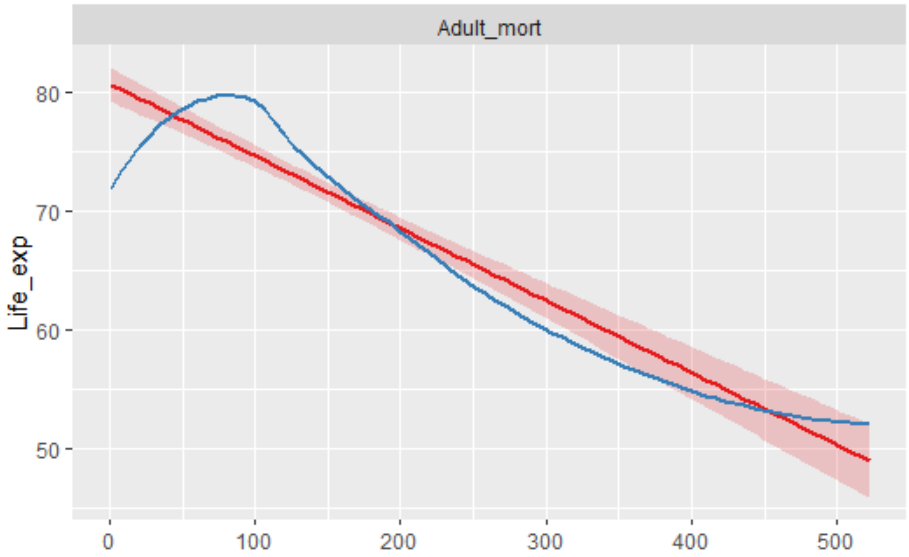

Рис. 13: График оцененного уравнения регрессии с регрессором *Adult_ mort*

Коэффициент детерминации второй модели ($M(y|x) = \hat{y} = 37.35 + 0.39x$) с регрессором *Dipht* оказался меньше, а именно, $R^2 = 0{,}2596$, что говорит о том, что меньшая доля вариации зависимой переменной *Life_ exp* была объяснена данной моделью. Нулевая гипотеза, указанная выше, так же была отвергнута. График данной модели, как видно, все же учитывает направление истинной связи, однако были выявлены некоторые особенности, которые модель не учитывает (в частности, линия истинной зависимости выбивается из доверительного интервала). Остатки очень похожи на нормальное распределение.

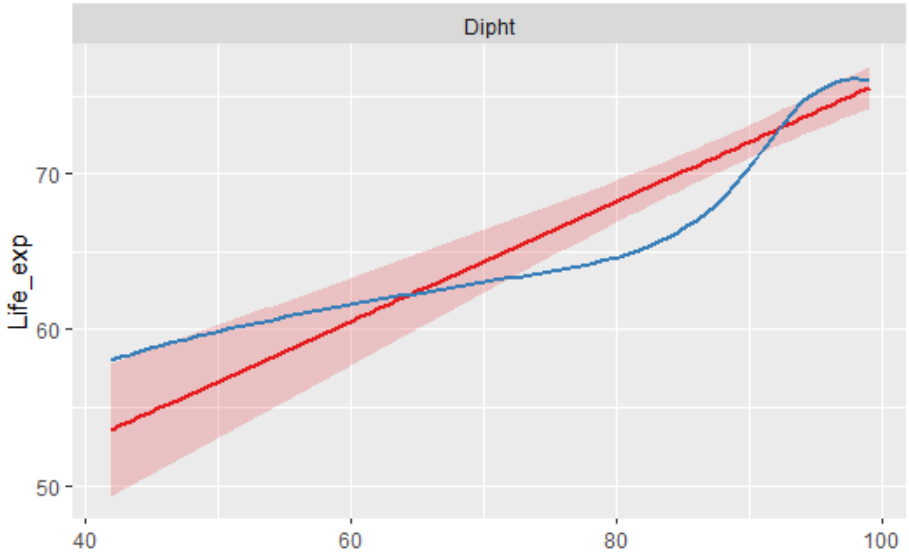

Рис. 14: График оцененного уравнения регрессии с регрессором *Dipht*

Что касается третьей модели ($M(y|x) = \hat{y} = 23.55 + 5.26x$) с регрессором *GDP*, отметим, что ее коэффициент детерминации больше, чем у двух моделей, рассмотренных выше. Данная модель объясняет уже 60.67% вариации зависимой переменной, что делает ее более выгодной для исследования (и заодно подтверждает поставленную изначально гипотезу о том, что **ВВП на душу населения, будучи одним из основных факторов, определяющих развитость страны, транзитивно через развитость обосновывает высокий уровень продолжительности жизни**). Нулевая гипотеза о незначимости коэффициента регрессии точно так же, как и у двух других моделей, отвергается. График третьей модели показывает хорошее приближение истинной зависимости и приблизительную нормальность остатков (с

некоторой асимметрией):

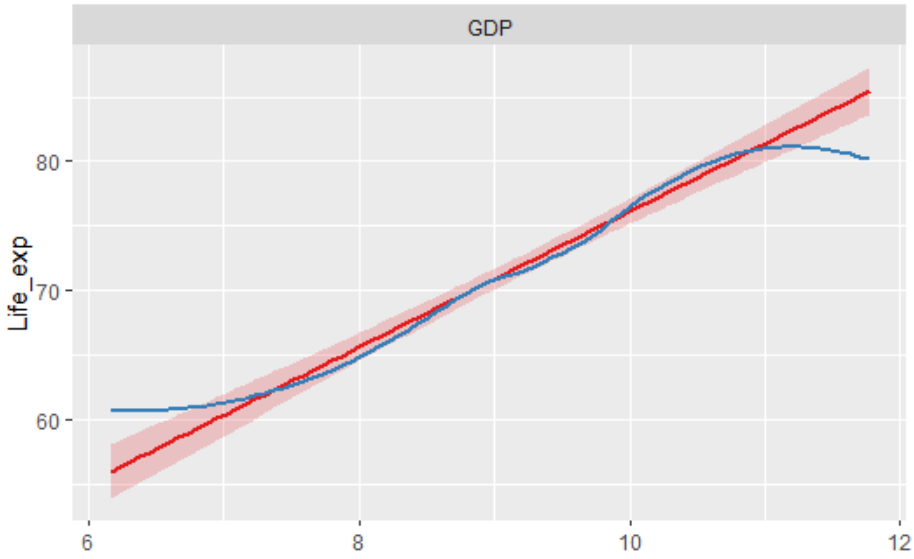

Рис. 15: График оцененного уравнения регрессии с регрессором *GDP*

Наконец, последняя двумерная модель обладает коэффициентом детерминации $R^2 = 0.6074$. Это хороший показатель, самый высокий из всех четырех рассматриваемых моделей (что удивительно). Нулевая гипотеза о незначимости коэффициента регрессии отвергается. График данной модели обладает практически теми же свойствами, что и график третьей модели: наблюдается хорошее приближение истинной зависимости и приблизительно нормальное распределение остатков с некоторой асимметрией:

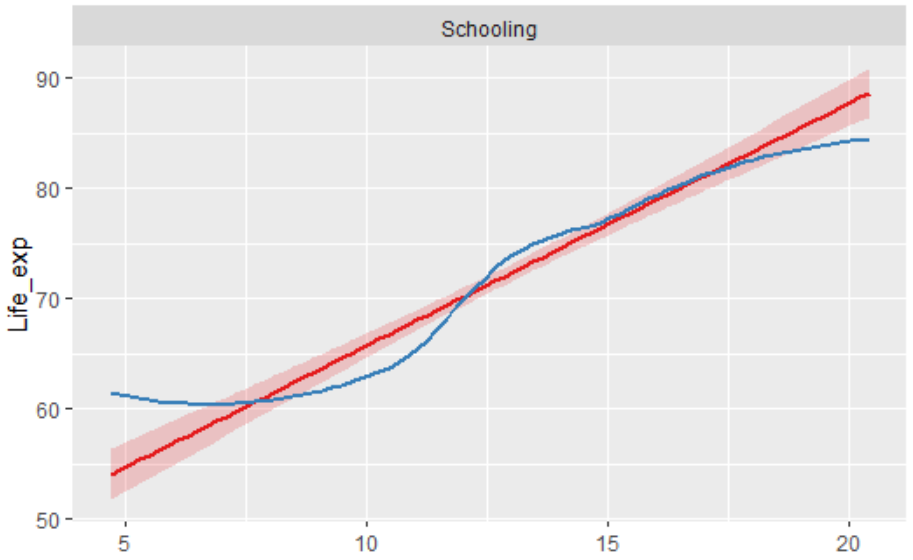

Рис. 16: График оцененного уравнения регрессии с регрессором *Schooling*

Таким образом, среди двумерных моделей лучшими стали модели зависимости между *Life_ exp* и *Schooling* и между *Life_ exp* и *GDP*. Однако этот вывод стоит назвать предварительным, так как моделям еще предстоит пройти сравнение по коэффициенту Акаике, для окончательного выявления лучшей.

5.2 Построение модели множественной линейной регрессии

Перейдем к последней модели линейной регрессии. В данной модели было принято решение учитывать лишь те 4 переменные, которые являются значимыми. Напомним, таковыми яв-

ляются: *Adult_ mort*, *Dipht*, *GDP* и *Schooling*. Было выдвинуто предположение, что именно эта модель будет показывать хороший результат.

Для данной модели коэффициент детерминации равен $R^2 = 0.791$, следовательно, модель объясняет 79.1% вариации результирующей переменной *Life_ exp*. Отметим, что уже на первом этапе анализа модель множественной линейной регрессии оправдывает ожидания и показывает лучший результат из всех рассмотренных моделей. Значение p-value $<$ 2.2e-16, что гораздо меньше всякого разумного уровня значимости, следовательно, нулевая гипотеза $H_0 : \beta_1 = \ldots = \beta_4 = 0$ (т.е. гипотеза о том, что независимые переменные не объясняют зависимую) отвергается.

Также построим график наблюдаемых и модельных значений зависимой переменной. Было проведено тестирование на 4 значимых переменных, но в качестве демонстрации была оставлена переменная *Adult_ mort*, наблюдения которой имеют необычную форму 2 кластеров. Как видно, модель очень неплохо приближает истинную зависимость: замечательно то, что модель "увидела", что данные по форме представляют собой два кластера точек, и в правой части первого квадранта модель также демонстрирует очень хорошую обобщающую способность:

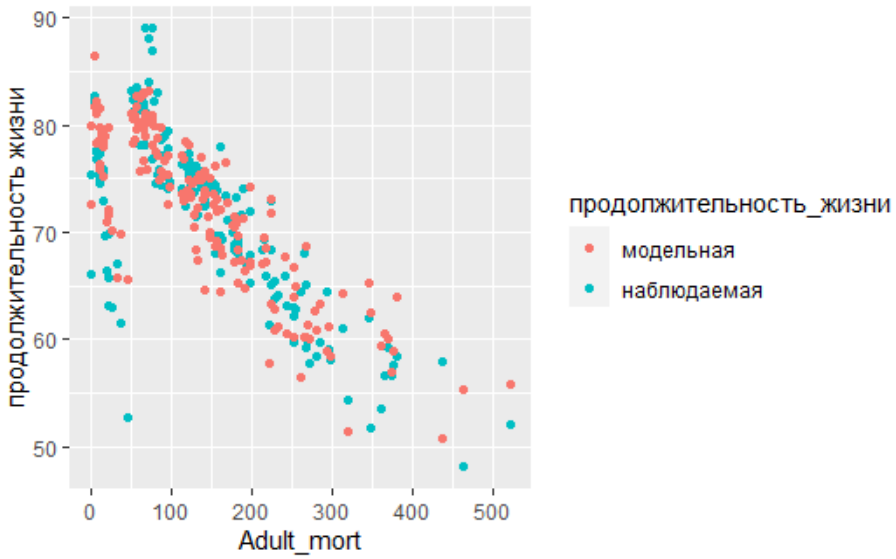

Рис. 17: График наблюдаемых и модельных значений зависимой переменной

Далее была выявлена лучшая регрессионная модель по критерию Акаике:

1. $AIC(lm1) = 1066.565$;
2. $AIC(lm2) = 1148.132$;
3. $AIC(lm3) = 1052.305$;
4. $AIC(lm4) = 1052.002$;
5. $AIC(lm5) = 951.1917$;

Как и предполагалось, лучшей моделью стала множественная модель линейной регрессии ($lm5$). Именно ее значение AIC является наименьшим. Можно сделать вывод о том, что продолжительность жизни наиболее оптимальным образом объясняется с учетом всех значимых характеристик, а не лишь одной конкретной. Для нас это стало очевидным и ожидаемым выводом, ведь продолжительность жизни человека определяется множеством факторов. Это

суждение подтверждает и статистика, и банальный жизненный опыт человека. Однако теперь становится ясно, что переменная *Life_ exp* наилучшим образом объясняется группой переменных *Adult_ mort*, *Dipht*, *GDP* и *Schooling*.

5.3 Построение нелинейных регрессионных моделей

Для построения нелинейных регрессионных моделей важно, чтобы среди наблюдений не было отрицательных значений, потому что в ходе построения моделей некоторые (а то и все) переменные могут быть прологарифмированы, а от отрицательных значений, как известно, логарифм брать нельзя. Поэтому были удалены отрицательные значения по переменным *Thinness_ 5_ 9* и *Thinness_ 10_ 19*. Также мы умножили значения переменной *Alcohol* на 1000 (перевели литры в миллилитры), чтобы избежать отрицательных значений после логарифмирования.

По построенной в предыдущем пункте линейной модели множественной регрессии видно, что наименьшее влияние на зависимую переменную оказывают переменные *Total_ exp*, *Thinness_ 5_ 9* и *Thinness_ 10_ 19*. Возможно, это обусловливается тем, что взаимосвязь между зависимой переменной и этими тремя переменными отлична от линейной. В связи с этим были построены графики зависимости переменной *Life_ exp* (по оси ординат) от этих трех переменных (по оси абсцисс), чтобы попробовать подобрать оптимальную нелинейную модель.

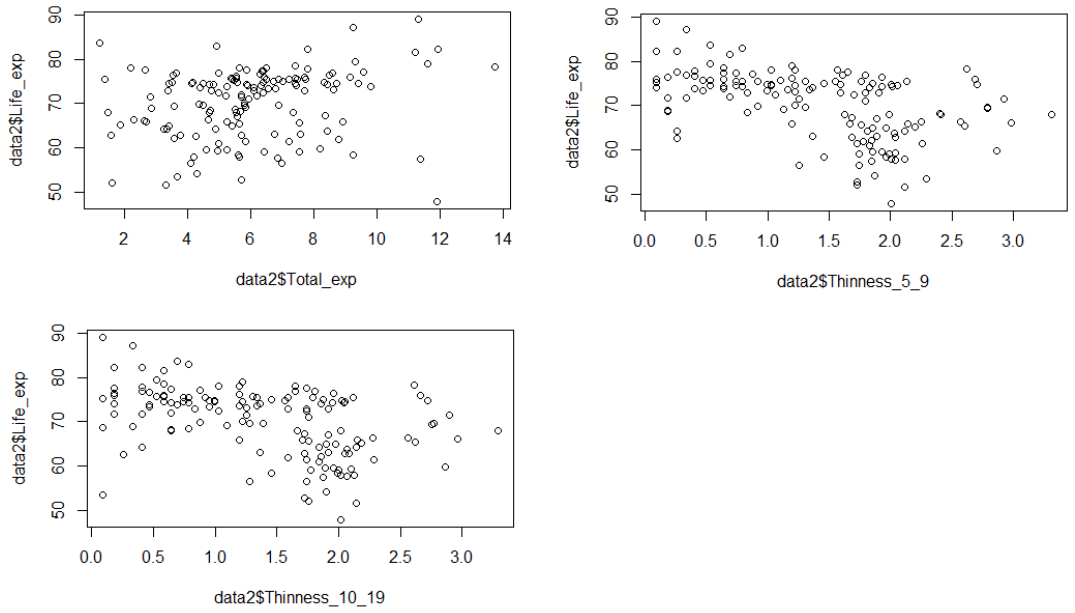

Рис. 18: Диаграмма рассеяния для зависимой переменной и переменных *Total_ exp*, *Thinness_ 5_ 9* и *Thinness_ 10_ 19*

Видно, что связь продолжительности жизни с *Thinness_ 5_ 9* и *Thinness_ 10_ 19* отдаленно напоминает линейную. В совокупности с тем фактом, что эти переменные оказывают малое влияние на зависимую, можно предположить, что вклад этих переменных уже учтен другими. Что касается зависимости продолжительности жизни от *Total_ exp*, то она кажется совсем хаотичной и не похожей ни на одну из известных нам элементарных функций.

Получается, в этой ситуации мы не можем однозначно быть уверены в том, что можно улучшить регрессионную модель за счёт нелинейных зависимостей. Тем не менее, попробуем применить некоторые базовые нелинейные преобразования: возможно, удастся в большей степени, чем в линейном случае, учесть вариацию зависимой переменной.

5.4 Полиномиальная модель

Мы попробовали рассмотреть квадратическую зависисмость продолжительности жизни от переменной *Total_exp*. У полученной модели коэффициент детерминации равен 0.793, это означает, что 79.3% вариации продолжительности жизни может быть объяснено всеми остальными переменными в этой модели. Значение p-value $<$ 2.2e-16, что меньше любого разумного уровня значимости, значит гипотеза о том, что независимые переменные не объясняют зависимую, отвергается. Стандартная ошибка остатков составляет 3.619, что прилично отличается от нуля, значит, данная регрессионная модель не имеет сильной прогнозирующей способности.

5.5 Экспоненциальная модель

Данная модель имеет вид:

$$y = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i \cdot x_i + \varepsilon).$$

В нашем случае она была построена с помощью логарифмирования зависимой переменной и дальнейшего построения многомерной линейной модели. У полученной модели коэффициент детерминации равен 0.78, это означает, что 78% вариации логарифма продолжительности жизни может быть объяснено всеми остальными переменными в этой модели (этот результат лучше всех линейных моделей, кроме множественной, но если рассмотреть экспоненциальную модель с учетом только статистически значимых переменных, то в этом случае экспоненциальная модель имеет лучший результат, хотя всё равно множественный R^2 у такой модели меньше, чем у множественной линейной модели регрессии). Значение p-value $<$ 2.2e-16, что меньше любого разумного уровня значимости, значит, гипотеза о том, что независимые переменные не объясняют зависимую, отвергается. Стандартная ошибка остатков 0.055, что достаточно мало отличается от нуля. По этому параметру описываемая регрессионная модель имеет наибольшую прогнозирующую способность из всех построенных в этом разделе моделей.

5.6 Степенная модель

Также была построена степенная модель, противоположность экспоненциальной:

$$y = \beta_0 \cdot \prod_{i=1}^{k} x_i^{\beta_i} \cdot \varepsilon.$$

Для этого были логарифмированы все переменные, кроме зависимой (и тех, что уже логарифмировались).

У этой модели коэффициент детерминации равен 0.69, это означает, что 69,36% вариации продолжительности жизни может быть объяснено всеми остальными логарифмированными переменными в этой модели. Значение p-value $<$ 2.2e-16, что меньше любого разумного уровня значимости, значит гипотеза о том, что независимые переменные не объясняют зависимую, отвергается. Стандартная ошибка остатков 4.43, что прилично отличается от нуля, значит, данная регрессионная модель не имеет сильной прогнозирующей способности.

5.7 Сравнение моделей. Графический анализ лучшей модели

По коэффициентам детерминации (R^2) и RSE (стандартной ошибке остатков) наилучшей является вторая модель - экспоненциальная. Она имеет наибольший коэффициент детерминации (0.78) при наименьшей ошибке (0.055). Также мы сравнили модели с помощью информационного критерия Акаике - AIC. Наименьший $AIC = -413$ оказался у экспоненциальной

модели, значит, по этому критерию она также считается наилучшей (лучше неё работает разве что экспоненциальная модель на статистически значимые переменные).

Далее будет рассмотрена только наилучшая (экспоненциальная) модель. Нормальный график квантиль-квантиль остатков выглядит следующим образом:

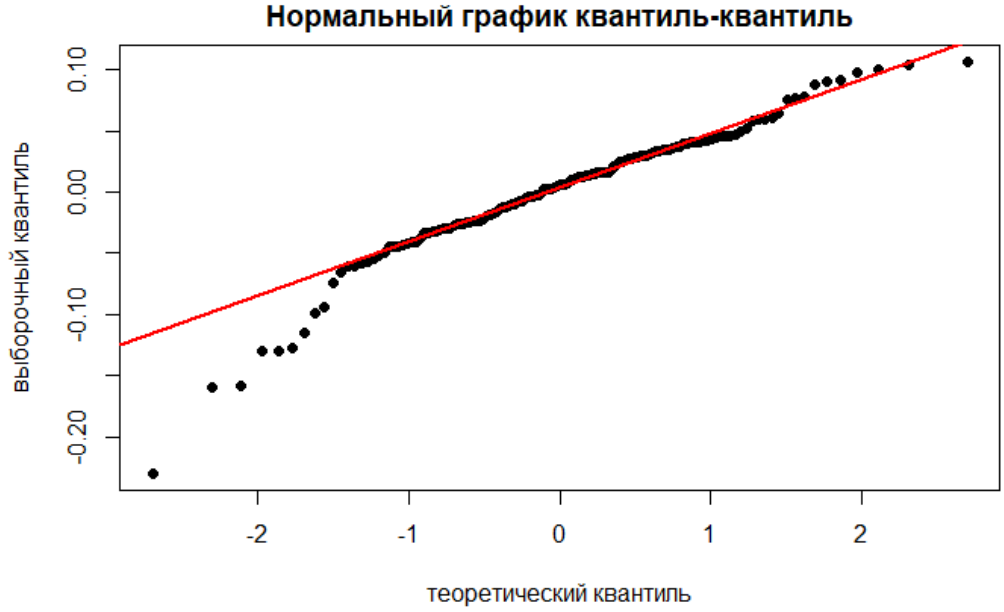

Рис. 19: График квантиль-квантиль

Как и обсуждалось, в хвостах выборочного квантиля заметно серьёзное отклонение от теоретического, следовательно, нормальность остатков сомнительна (что и подтверждают весьма низкие значения p-value по тесту Шапиро-Уилка).

Далее был построен график модельных и фактических значений для зависимости двух переменных. В качестве рассматриваемой переменных была взята переменная *Adult_mort*, так как во всех построенных моделях она вносит достаточно большой вклад в зависимую переменную:

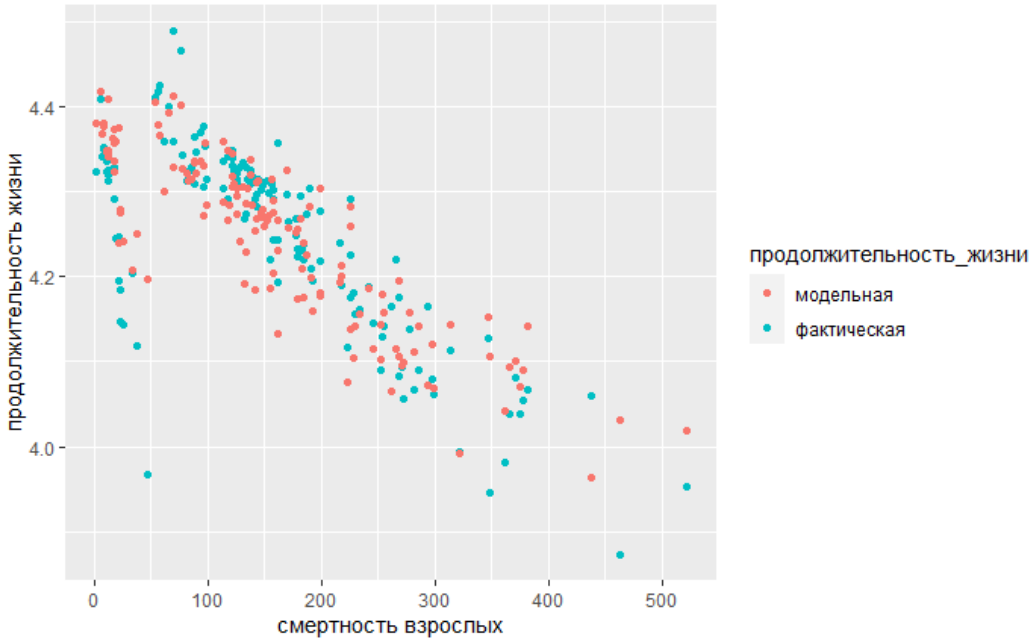

Рис. 20: Диаграмма рассеяния модельных и истинных значений зависимой переменной

Можно заметить, что фактическое распределение поделено на некоторые 2 кластера, и модельные значения даже повторяют это разделение на кластеры. Это, безусловно, хороший показатель, значит, наша модель достаточно точная, причём способна учитывать такое сложно распределение данных.

5.8 Корректная запись уравнения регрессии. Выводы

На этом этапе компьютерной работы была выведена следующая лучшая модель (лучшая как среди нелинейных, так и среди линейных моделей, по совокупности различных критериев):

$$\hat{y} = \exp\left(b_0 + \sum_{i=1}^{8} b_i \cdot x_i + \varepsilon\right) = \exp(4.026 - 0.0005 \cdot x_1 + 0.0000008554 \cdot x_2 + 0.001554 \cdot x_3 +$$

$$+0.0008 \cdot x_4 + 0.0126 \cdot x_5 - 0.027 \cdot x_6 + 0.0056 \cdot x_7 + 0.011 \cdot x_8 + \varepsilon)$$

Но нам также известна лучшая экспоненциальная модель со статистически значимыми переменными, построенная по всему датасету и упущенная на этом этапе, запишем также и её уравнение:

$$\hat{y} = \exp(b_0 + b_1 \cdot x_1 + b_5 \cdot x_5 + b_8 \cdot x_8 + \varepsilon) = \exp(3.983 - 0.0005213 \cdot x_1 + 0.0162 \cdot x_5 + 0.0165 \cdot x_8 + \varepsilon).$$

Далее для второй экспоненциальной модели проинтерпретируем значения $\hat{\beta}_i$. Поскольку рассматривается экспоненциальная модель, то значения $\hat{\beta}_i$, домноженные на 100% показывают, **на сколько процентов изменится результирующая переменная при изменении независимой переменной x_i на 1 единицу в абсолютном значении**. То есть:

- уменьшение значения смертности в возрасте 15-60 лет на 1000 человек на 1 единицу своего измерения ведёт к увеличению продолжительности жизни в среднем на 0.052%;
- увеличение значения ВВП на душу населения на 1 единицу своего измерения (то есть на 1 доллар США) ведёт к увеличению продолжительности жизни в среднем на 1.62%. Вот что доллар животворящий делает! Видно, что не очень большое изменение ВВП на душу населения увеличивает продолжительность жизни сильнее, чем единичное измерение смертности среди взрослых;

- увеличение количества лет, потраченных на образование, на 1 год ведёт к увеличению продолжительности жизни в среднем на 1.651%. Это интересно: получается, вопреки нашим предположениям, базирующимся на наивной логике и здравом смысле, образование существенно влияет на продолжительность жизни. Хотя тут, наверное, имеет место **корреляция**: скорее всего, в странах с высоким уровнем развития, обладающих высоким уровнем жизни и образования, выше и продолжительность жизни.

Анализ коэффициентов эластичности показывает нам следующее:

- уменьшение значения смертности в возрасте 15-60 лет на 1000 человек на 1% ведёт к увеличению продолжительности жизни в среднем на 0.08%;
- увеличение ВВП на душу населения на 1% ведёт к увеличению продолжительности жизни в среднем на 0.15%;
- увеличение количества лет, потраченных на образование, на 1% ведёт к увеличению продолжительности жизни в среднем на 0.21%.

6 Выводы по первой компьютерной работе

Подводя итог всей первой компьютерной работе в целом и выполнения регрессионного анализа в частности, можно сказать, что многие выдвинутые гипотезы подтвердились, но были и не совсем корректные предположения.

Исследование показало, что переменные, которые были обозначены как наиболее значимые, были выбраны не совсем точно: в частности, *Total_ expenditure* показала себя не так уверенно, как предполагалось, также была недооценена переменная *Schooling*, которая, как оказалось, вносит существенный вклад (хотя по-прежнему разумным кажется объяснение о наличии скорее транзитивной корреляции посредством других переменных, характеризующих страны с низким и высоким уровнем продолжительсти жизни).

Заметим также, что **иммунизация**, которая, как предполагалось, может оказать существенное влияние на продолжительность жизни, не попала в число регрессоров лучшей модели. С другой стороны, нельзя сказать, что применение исследованных нелинейностей смогло сделать модель идеальной: как было показано, у модели имеются некоторые неточности и неучтённые зависимости. Возможно, это связано с большой гетероскедастичностью данных. В следующих частях компьютерной работы будет рассмотрено разделение пространства наблюдений на гипотетически более значимые части, в которых, возможно, построение регрессии будет более качественным – эта гипотеза впоследствии не оправдалась, но провести исследование стоило.

7 PCA

Для выполнения задачи выделения главных компонент было необходимо предварительно преобразовать датасет, убрав дамми-переменную и зависимую из числа факторов при построении главных компонент.

7.1 Проверка применимости PCA

Перед выделением главных компонент необходимо провести проверку применимости метода главных компонент для данной выборки с помощью трех способов: построения корреляционной матрицы, расчета значения критерия Кайзера-Майера-Олкина (KMO) и теста сферичности Бартлетта.

Согласно всем трем способам применимость метода главных компонент была подтверждена. Признаки, описывающие совокупность, оказались достаточно коррелированы между собой, судя по построенное ранее корреляционной матрице, значение критерия KMO близко к единице (0,77), как и должно быть для применимости МГК, а нулевая гипотеза о том, что теоретическая корреляционная матрица многомерного распределения вектора случайных величин, представляет собой единичную матрицу, по тесту Бартлетта была отвергнута, что тоже свидетельствует о возможности использования МГК.

7.2 Отбор главных компонент

Далее необходимо определить число главных компонент, оставленных для дальнейшего анализа. Сначала был использован критерий Кайзера. В результате отобраны две главные компоненты, чьи собственные значения $\lambda_i > 1$.

Затем был применён метод определения числа главных компонент по долям суммарной вариации. По данному критерию выделяем три главные компоненты, так как доля суммарной вариации составляет 76%, что достаточно для сохранения. Более того, собственное значение третьей главной компоненты $\lambda_i = 0.93$, что ненамного меньше 1.

Наконец, построим график каменистой осыпи:

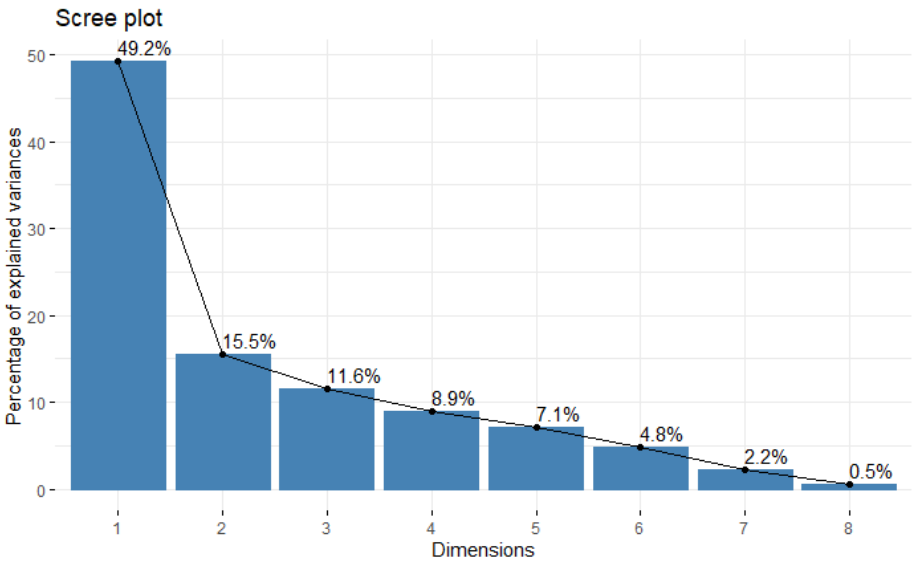

Рис. 21: График каменистой осыпи

Резкое падение происходит после первой компоненты, значит, в соответствии с данным критерием, нужно выделить одну главную компоненту, что, конечно, недостаточно, поскольку одна главная компонента не объясняет и половины вариации результирующей переменной. В итоге было решено выделить **три главные компоненты**, так как на долю этих компонентов приходится более 70% вариации зависимой переменной и эти главные компоненты будут наиболее точны и **информативны**, что будет особенно заметно по матрице факторных нагрузок далее.

Теперь проведём анализ суммарного вклада первых главных компонент. На первую главную компоненту приходится практически половина (49.2%) сохраненной дисперсии. На вторую - значительно меньше – 15.5%), на третью - 11.6%. Доля суммарной вариации для трех главных компонент составляет 76.3%. Величина стандартного отклонения для трех главных компонент составляет соответственно 1.98, 1.11 и 0.96.

Оценим вклад в главные компоненты различных переменных:

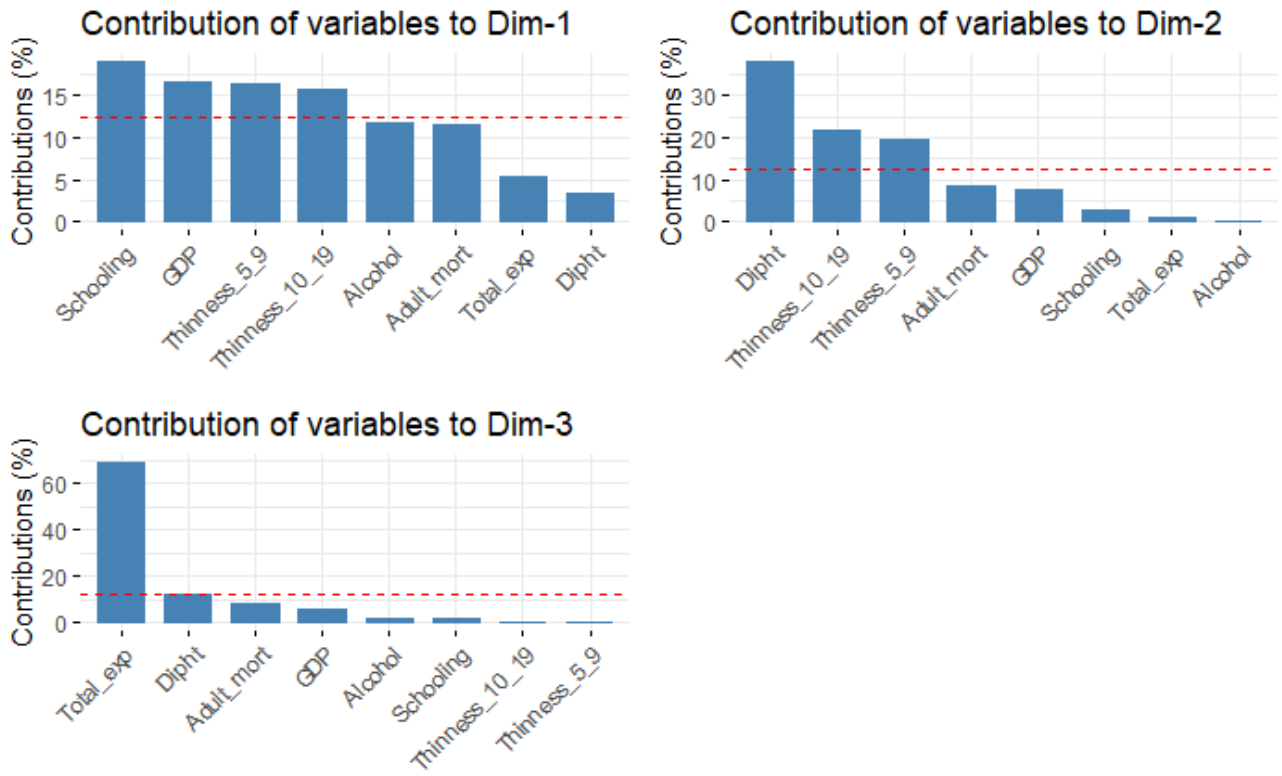

Рис. 22: Вклад в главные компоненты различных переменных

По построенным графикам четко видно, что максимальный вклад в первую главную компоненту вносит переменные *Schooling*, *GDP*, *Thinness_5_9* и *Thinness_10_19*. Во вторую главную компоненту наибольший вклад вносят переменные *Dipht*, *Thinness_5_9* и *Thinness_10_19*, а наименьший - *Alcohol*. В третью главную компоненту основной вклад (более 50%) вносит переменная *Total_exp*.

7.3 Интерпретация главных компонент

Матрица факторных нагрузок позволяет проинтерпретировать главные компоненты и дать им название. По данной матрице определяется корреляция каждой переменной с каждой главной компонентой. Но сама по себе матрица факторных нагрузок не позволяет в достаточной степени проанализировать главные компоненты и дать им название, поэтому было применено варимакс-вращение для ясной видимости факторов, отмеченных высокими нагрузками для одних переменных и низкими – для других:

	PC1	PC2	PC3
Adult_mort	-0.77	0.21	0.02
Alcohol	0.42	-0.45	0.33
Total_exp	0	-0.29	0.88
Dipht	0.58	0.33	0.51
GDP	0.84	-0.31	0.07
Thinness_5_9	-0.27	0.9	-0.15
Thinness_10_19	-0.25	0.91	-0.11
Schooling	0.77	-0.42	0.17

Рис. 23: Матрица факторных нагрузок с варимакс-вращением

С помощью варимакс-вращения получена более понятную и удобная матрица факторных нагрузок. Первая главная компонента сильно коррелирует с переменными *GDP* (ВВП на душу населения), *Schooling* (количество лет обучения в школе), *Adult_mort* (уровень смертности среди взрослого населения) (с этой переменной у главной компоненты обратная связь) и *Dipht* (количество привитых от столбняка детей в возрасте до одного года в процентах). Учитывая эти знания, первая главная компонента может быть названа **Уровень жизни населения**, так как чем он выше, тем выше ВВП, ниже смертность среди взрослого населения и выше число привитых детей, что говорит о высоком уровне экономического развития страны и высоком внимании государства к сфере образования и здравоохранения. Вторая главная компонента имеет сильную корреляцию с переменными *Thinness_5_9* (показатель истощения среди детей в возрасте от 5 до 9 лет в процентах), *Thinness_10_19* (показатель истощения среди детей и подростков в возрасте от 10 до 19 лет в процентах), *Alcohol* (потребление алкоголя на душу населения) и *Schooling* (количество лет обучения в школе). С последними двумя признаками связь обратная. Следовательно, вторая главная компонента будет носить название **Уровень бедности населения**. Третья главная компонента по большей части коррелирует с признаками *Total_exp* (доля расходов государства на здравоохранение от общих расходов) и *Dipht* (количество привитых от столбняка детей в возрасте до одного года в процентах). Эта главная компонента названа **Уровень развития системы здравоохранения**.

Если построить матрицу факторных нагрузок с вари-макс вращением для двух главных компонент, то заметна неточность интерпретации и размытость определения главных компонент, так как первая главная компонента представляет собой некий позитивный показатель, а вторая - отрицательный, что, конечно, является чрезмерным упрощением имеющейся системы показателей. Поэтому принятое решение о выделении трех главных компонент имеет смысл и дает больше пространства для анализирования.

	PC1	PC2	PC3
Adult_mort	-0.77	0.21	0.02
Alcohol	0.42	-0.45	0.33
Total_exp	0	-0.29	0.88
Dipht	0.58	0.33	0.51
GDP	0.84	-0.31	0.07
Thinness_5_9	-0.27	0.9	-0.15
Thinness_10_19	-0.25	0.91	-0.11
Schooling	0.77	-0.42	0.17

Рис. 24: Матрица факторных нагрузок с варимакс-вращением для 2 главных компонент

7.4 Построение уравнения регрессии на главные компоненты

Итак, после выделения главных компонент и их всевозможного объяснения и описания, можно перейти к построению линейного уравнения регрессии на ГК. В результате получилась модель, в которой все три используемые главные компоненты являются статистически значимыми, а значение p-value $<$ 2.2e-16, что говорит о том, что независимые переменные объясняют зависимую, а гипотеза $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ отвергается. Коэффициент детерминации = 0,7931, связь между переменными очень тесная: модель объясняет 79.3% вариации результирующей переменной (что является **самым высоким показателем среди всех построенных в обеих компьютерных работах уравнений регрессии**).

7.5 Сопоставление свойств ранее полученных уравнений регрессии с уравнением регрессии на ГК

Для лаконичности анализа уточним, что лучшей **линейной** моделью была выбрана модель множественной линейной регрессии, учитывающая следующие значимые переменные: *Adult_mort*, *Dipht*, *GDP* и *Schooling*. Среди моделей нелинейной регрессии, как было решено ранее, лучшей считается экспоненциальная модель. Данные выводы позволяют нам перейти к непосредственному сравнению упомянутых выше моделей по информационному критерию Акаике, как мы и делали в разделе регрессионного анализа. Суть модели регрессии, построенной на главные компоненты, фактически не меняется, что дает нам основание использовать именно AIC в своем анализе. Разумеется, в этом случае лучшей моделью оказывается **экспоненциальная**, но есть основания полагать, что если бы мы попробовали применить какие-то преобразования к главным компонентам, удалось бы получить новую лучшую модель регрессии. Тем не менее, следует учитывать, что всякое следующее преобразование снижает интерпретационную силу модели, поскольку так или иначе подразумевают некоторое обобщение, сглаживающее особенности переменных.

8 Кластерный анализ

8.1 Построение и анализ дендрограмм

Рассмотрим несколько вариантов разбиения объектов на кластеры. Но для начала необходимо рассмотреть коррелированность признаков: если мы собираемся воспользоваться евклидовым расстоянием, то предпочтительна слабая коррелированность. Для этого рассмотрим матрицу корреляций до удаления выбросов, чтобы примерно оценить, насколько качественным будет дальнейшая кластеризация. Как видно, присутствует достаточно число независимых переменных, которые довольно сильно коррелируют. Также размер выборки ($>$100) говорит нам о том, что к дальнейшему кластерному анализу следует относиться осторожно, поскольку некоторые предпосылки, необходимые для качественного построения кластеров нижеследующими методами, не выполняются в полной мере.

Нами были использованы следующие методы разбиения объектов на кластеры:

1. метод ближнего соседа;
2. метод дальнего соседа;
3. метод центра тяжести;
4. метод средней связи;
5. метод Уорда.

Все дендрограммы, построенные на основе данных методов, показали, что, судя по такой метрике, как расстояние между местами объединения данных в кластеры, оптимальным будет разбиение на 2 кластера. Например, это показал метод дальнего соседа:

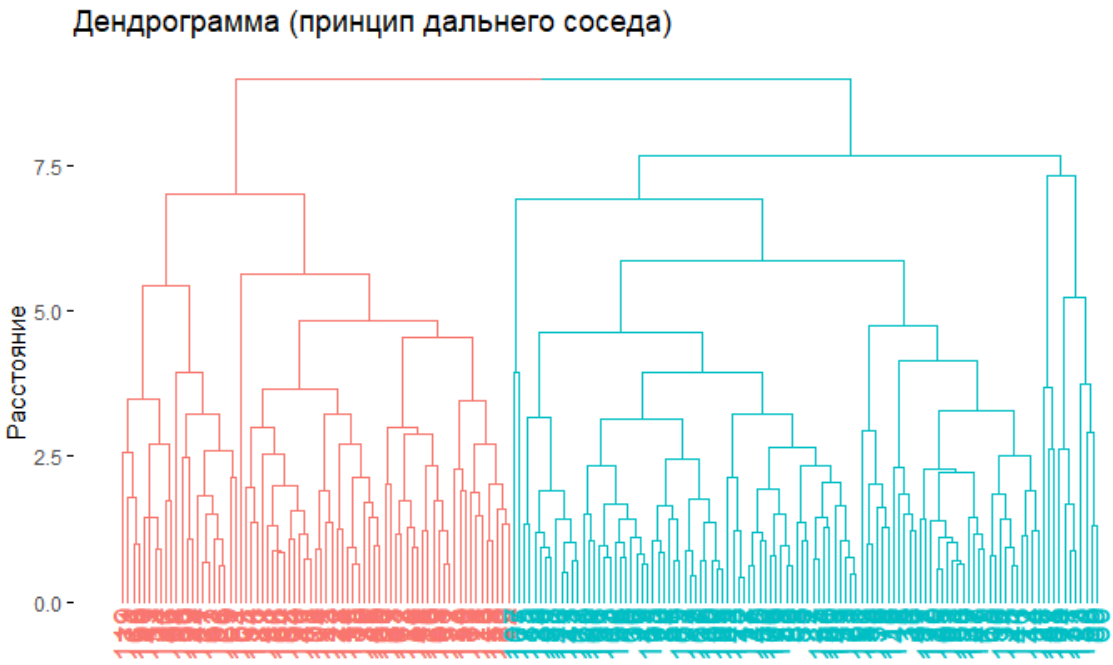

Рис. 25: Дендрограмма (принцип дальнего соседа)

И ещё более наглядно – метод Уорда:

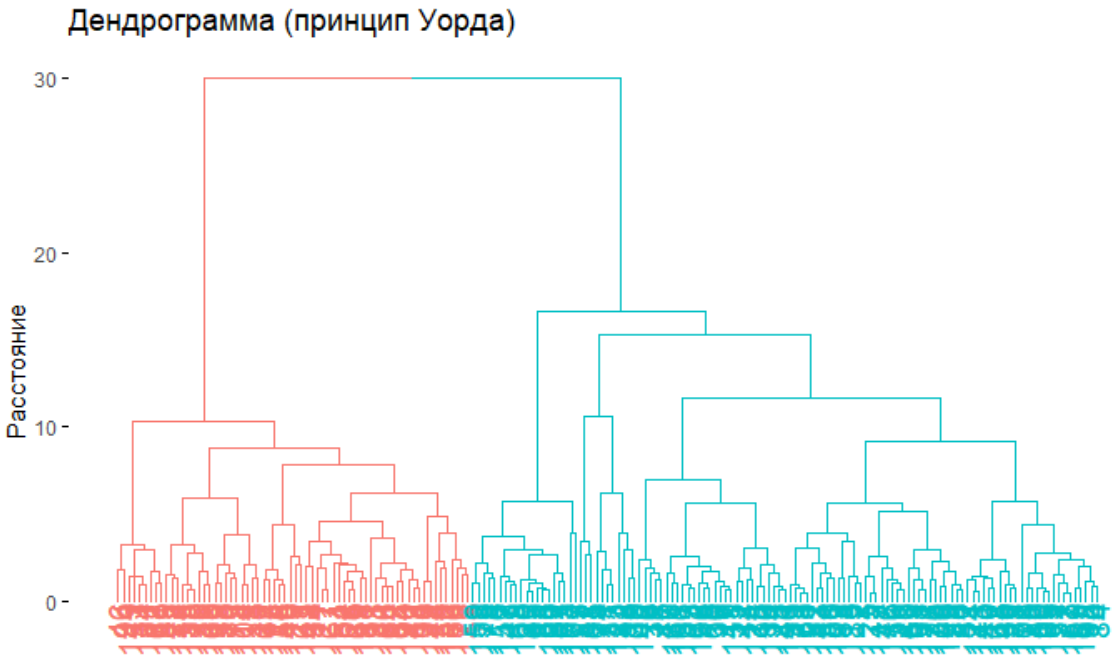

Рис. 26: Дендрограмма (принцип Уорда)

На этих графиках видно, как четко и аккуратно объекты разбиты на кластеры. Также можно сделать вывод о том, что объекты разбиты приблизительно равномерно, не наблюдается сильное количественное преобладание одного кластера над другим.

К сожалению, не все дендрограммы были такими же информативными. Приведем пример такого “неинформативного” графика:

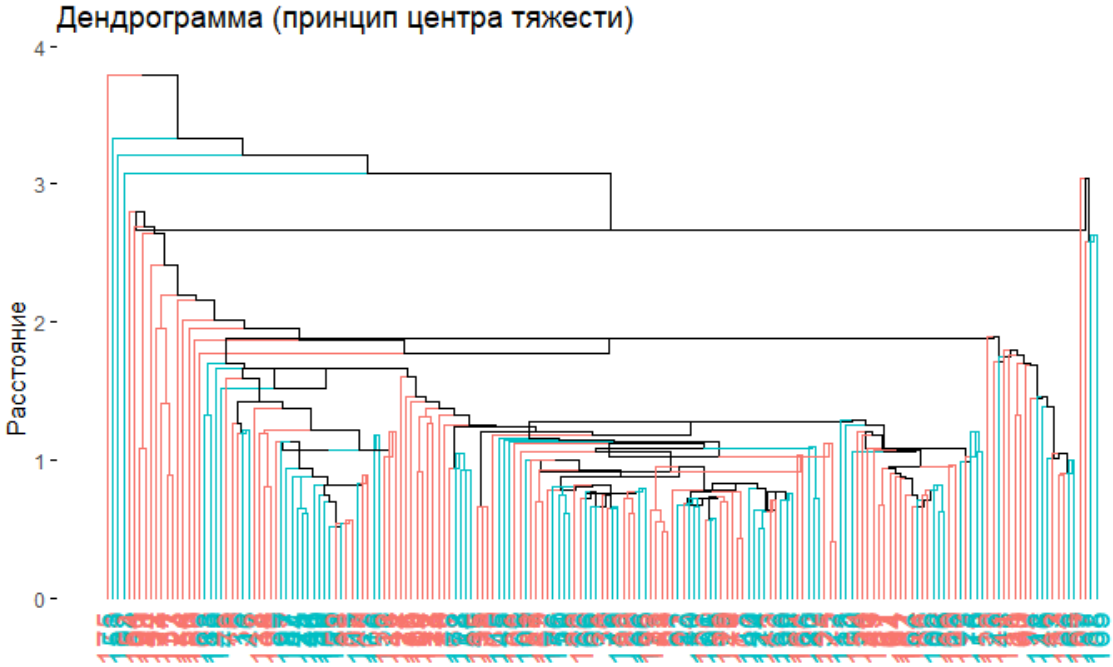

Рис. 27: Дендрограмма (принцип центра тяжести)

Как можно видеть, дендрограмма, построенная согласно методу центра тяжести, не показывает четкой иерархии кластеров, следовательно, не является информативной и может ввести в заблуждение своей нагроможденностью и нечеткостью.

8.2 Деление на кластеры методом k-means. Построение графика средних значений показателей в кластерах.

Перед делением на кластеры мы стандартизовали данные и создали новый набор данных, содержащий все стандартизованные переменные, кроме зависимой.

Для начало было рассмотрено деление на оптимальное по **методам локтя и силуэтов** число кластеров – 2. Мы получили 2 кластера по 77 и 106 наблюдений соответственно. Значение $\frac{BSS}{TSS}$ оказалось достаточно низким - 34.1%. В хороших моделях оно должно стремиться к 100%.
График средних значений выглядит так:

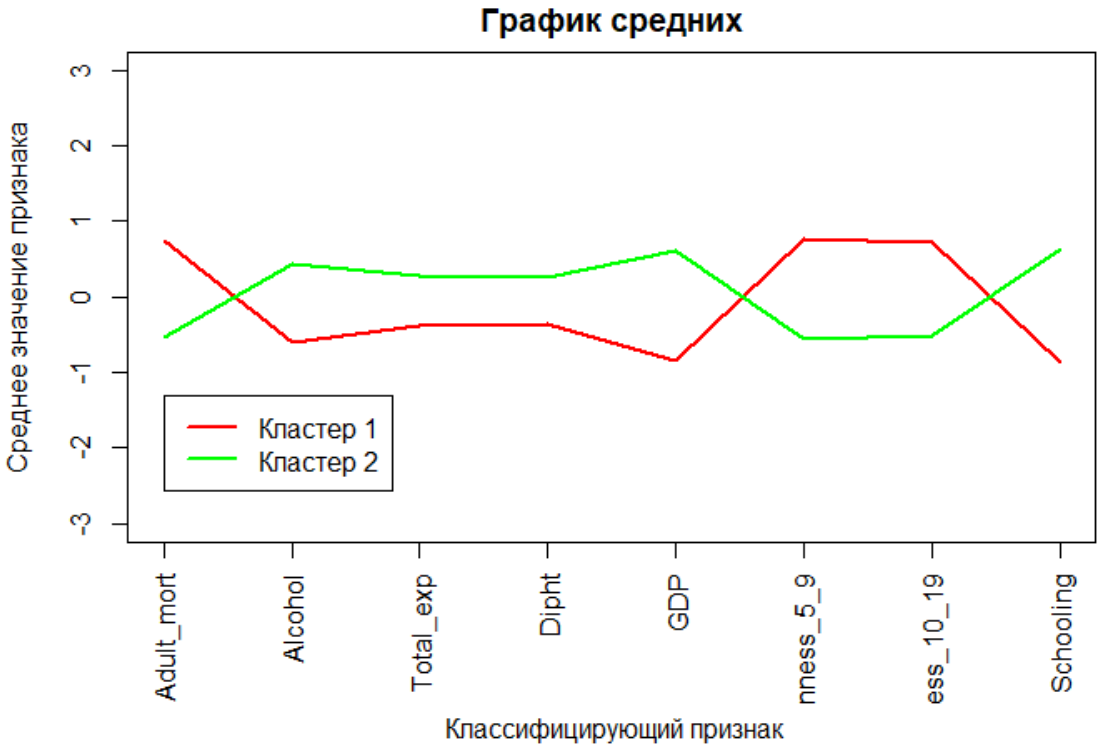

Рис. 28: График средних ($k = 2$)

Заметно, что кластеры различаются по всем переменным. Второй кластер, однако, получился достаточно большой. Поэтому мы решили, было бы логично рассмотреть возможность деления на большее количество кластеров.

Поэтому было проведено деление на 3 кластера. Мы получили кластеры по 79, 43 и 61 наблюдений, что кажется лучше, чем в прошлом случае. Статистика $\frac{BSS}{TSS}$ заметно увеличилась до 45.5%. Сравним график средних с предыдущим случаем:

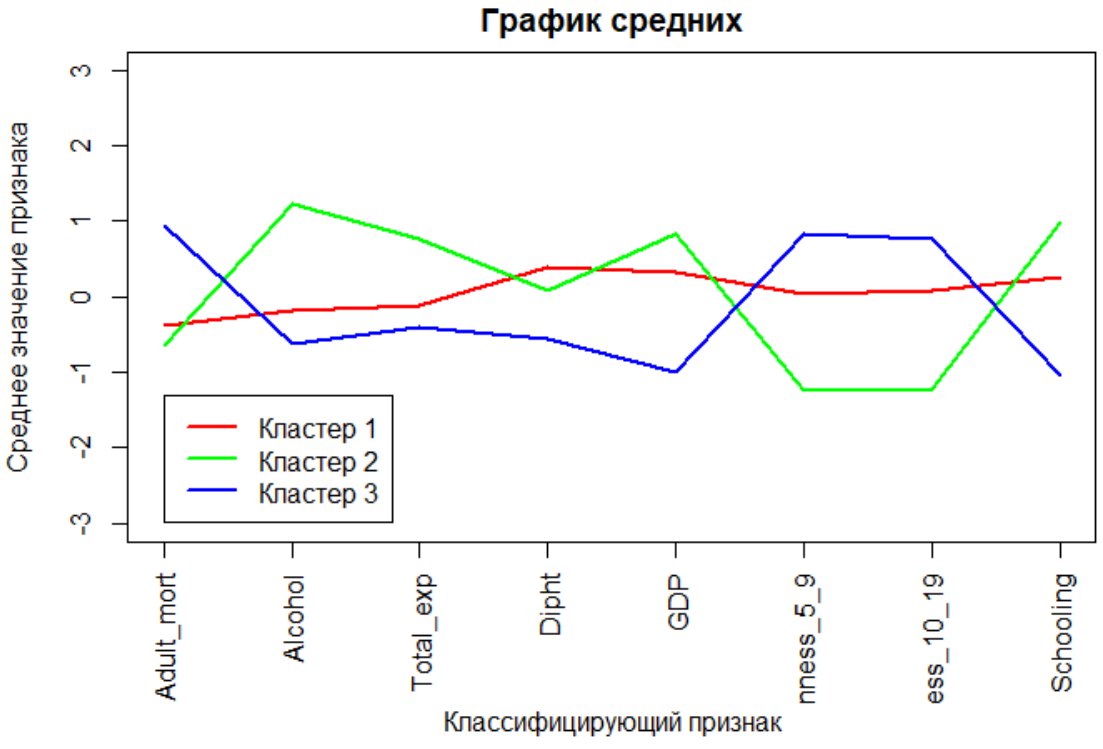

Рис. 29: График средних ($k = 3$)

По графику ситуация не сильно отличается от случая с двумя кластерами - так же заметные различия по всем переменным. Учитывая все вышеупомянутые факторы, мы считаем более грамотным в случае метода k-means рассматривать деление на 3 кластера, а не на 2. Таким образом нам удастся отойти от той классификации, что есть в исходном датасете (подразделение стран на развитые и развивающиеся), и получить более интересные кластеры.

8.3 Проверка гипотезы о равенстве средних в кластерах

На всякий случай проверим гипотезу о равенстве средних значений в кластерах. Для этого сначала проверим гипотезу о равенстве дисперсий по каждому признаку в каждой паре кластеров (1 и 2, 2 и 3, 1 и 3). Гипотеза отвергается лишь в 3 случаях из 24, что можно считать, что она не отвергается и что дисперсии равны. Затем воспользовались t-тестом Уэлча для проверки гипотезы о равенстве средних. Эта гипотеза отвергается в 21 случае из 24. Следовательно, гипотеза о равенстве средних отвергается, и средние значения кластеров можно считать неравными. Это хороший результат, потому что качественное деление на кластеры должно давать различные значения средних в кластерах.

8.4 Интерпретация полученных результатов

Видно, что во втором кластере (43 наблюдения) наименьшие средние показатели у переменной *Adult_ mort* (смертности взрослых), *Thinness_ 5_ 9* и *Thinness_ 10_ 19* (худоба среди детей разных возрастов), наибольшие по *Alcohol* (потребление алкоголя), *Total_ exp* (траты государства на здравоохранение), *GDP* (ВВП страны) и *Schooling* (среднее количество лет образования по стране), и средние (относительно двух других кластеров) по переменной *Dipht* - иммунизация от дифтерии и проч. среди детей возраста менее 1 года.

В третьем кластере (61 наблюдение) ровно наоборот - наибольшие средние показатели у переменной *Adult_ mort*, *Thinness_ 5_ 9* и *Thinness_ 10_ 19*, наименьшие - по всем остальным переменным.
Средние первого кластера (79 наблюдений) находятся посередине между средними второго и третьего кластера по всем переменным, кроме *Dipht* - иммунитет к дифтерии среди детей возрастом 1 год.

Таким образом, мы назвали второй кластер **Хорошо развитые страны**, потому что уровень медицины и образования, ВВП там наибольший, уровень заболеваний наименьший. Третий кластер назвали **Неразвитые страны**. По тому же принципу первый кластер назвали **Страны со средним уровнем развития**. Получилось довольно любопытно: от классификации на развитые и развивающиеся страны, принятой МВФ, мы перешли к кластеризации на 3 группы. Было бы интересно посмотреть, насколько сильно пересекаются полученные кластеры с той классификацей стран, что принята в ООН (в ООН страны так же делятся на развитые, развивающиеся и наименее развитые).

В пространстве первых двух главных компонент кластеры выглядят так:

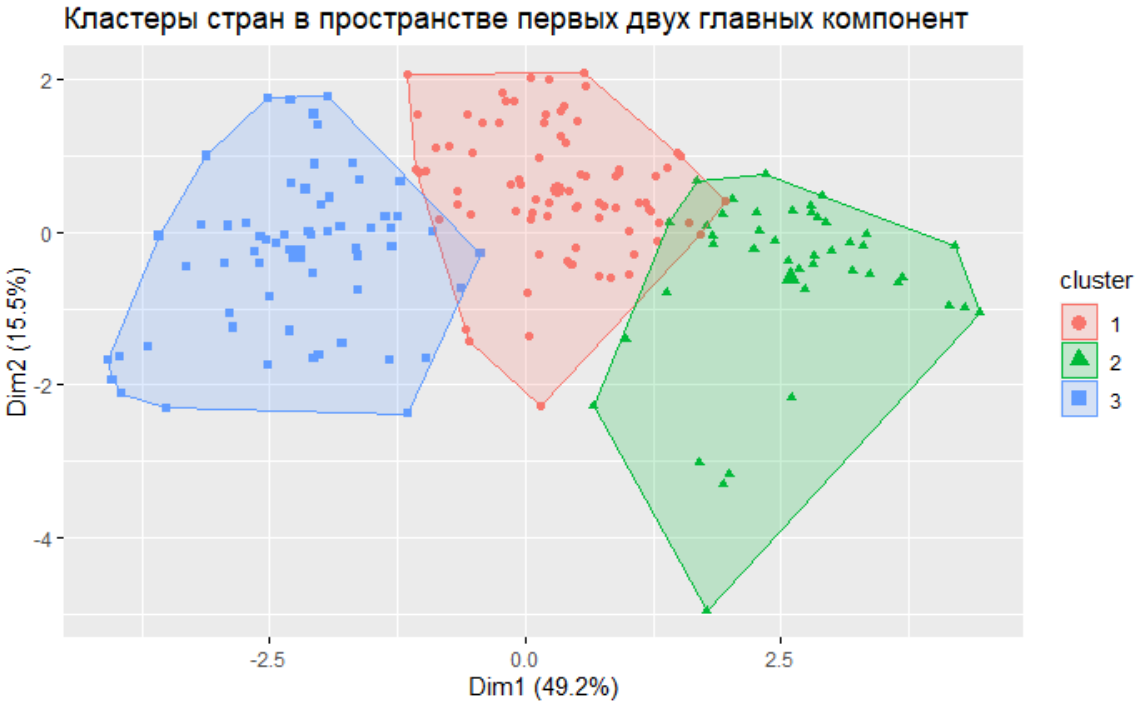

Рис. 30: Кластеры стран в пространстве первых двух главных компонент

Этот график подтверждает целесообразность нашего деления выборки на 3 кластера, потому что здесь видно, что они вполне четко выделяются. По второй компоненте разница в кластерах неочевидна, но вот по первой четко видно разделение между кластерами. Также можно сказать, что 2 главные компоненты объясняют около 65

Затем рассмотрим кластеры в пространстве переменных *Adult_ mort* и *GDP*:

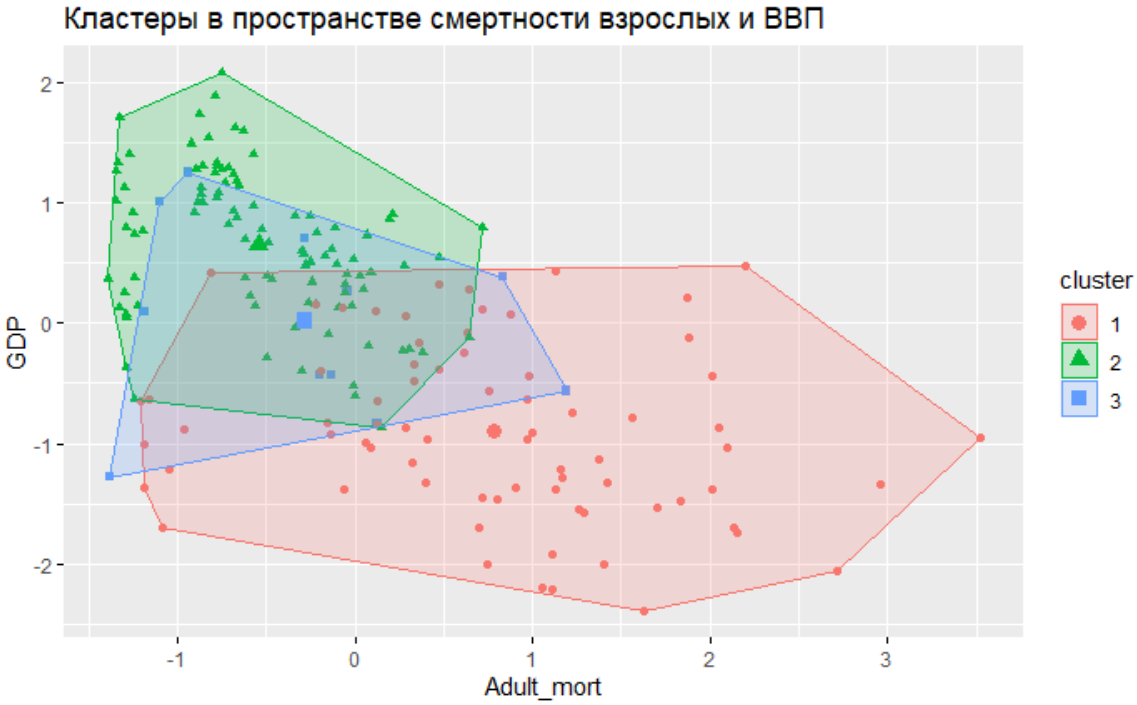

Рис. 31: Кластеры стран в пространстве *Adult_ mort* и *GDP*

В пространстве этих переменных кластеры, конечно, сильно пересекаются, но можно отметить, что для красного кластера (Неразвитые страны) характерны наблюдения с наибольшей смертностью взрослых и наименьшим ВВП, а для зеленого кластера (Хорошо развитые страны), наоборот, - наименьшая смертность и наибольший ВВП. Соответственно, синий кластер

(Страны со средним уровнем развития) находится посередине.

В пространстве других двух переменных ситуация такая:

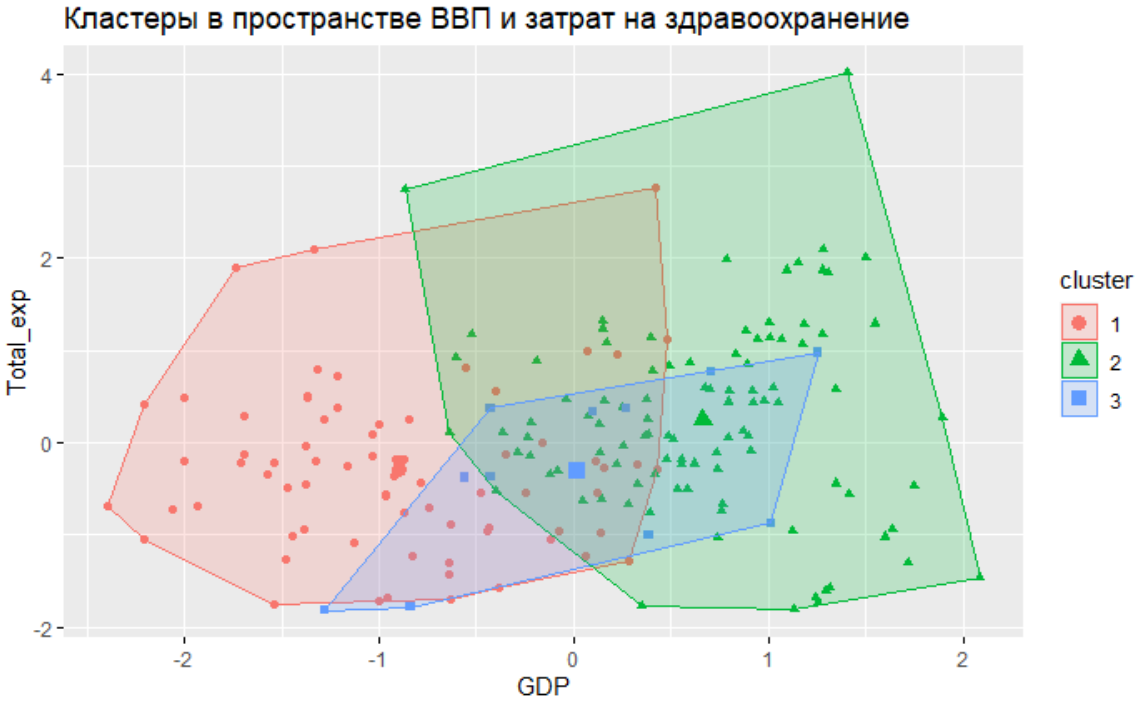

Рис. 32: Кластеры стран в пространстве *Total_ exp* и *GDP*

Хотя в этом случае кластеры очень сильно пересекаются, всё-таки по этому графику видно, что страны с наибольшими ВВП и затратами на здравоохранение отнесены к одному кластеру (зеленый) - Хорошо развитые страны. Страны с наименьшим ВВП и низкими затратами на здравоохранение относятся к красному кластеру - Неразвитые страны. Синий же кластер содержит страны со средним уровнем развития, которым характерны средние показатели ВВП и затрат на здравоохранение. Однако можно заметить, что средние (более крупные точки соответствующей формы) по переменной *Total_ exp* достаточно близки (близость между средними кластеров Неразвитые страны и Страны со средним уровнем развития видна и по графику средних).

Проводя кластеризацию методом k-means, мы решили взять 3 кластера, а не 2 (оптимальное количество), чтобы уменьшить различия в размерах кластеров, а также увеличить статистику $\frac{BSS}{TSS}$. Ни логика, ни уникальность кластеров от этого не пострадали, так что данное решение было вполне обдуманным. По графику средних кластеры выглядели уникально, средние не совпали ни по одной переменной. Более того, гипотеза о равенстве средних значений отверглась, что подтверждает грамотность проведенной кластеризации. По получившимся средним значениям переменных было решено назвать получившиеся кластеры стран так: **Хорошо развитые страны**, **Неразвитые страны**, **Страны со средним уровнем развития**.

9 Типологическая регрессия

На этом этапе работы в тех кластерах, что были получены методом k-means, были построены различные виды регрессионных моделей, среди которых затем с помощью Байесовского информационного критерия были выбраны лучшие образцы, и было проведено сопоставление качества отобранных регрессионных моделей с той моделью, что были построены для

совокупности в целом в первой компьютерной работе.
Есть все основания полагать (об этом говорит тот факт, что данные по странам даже на уровне гипотезы компактности должны группироваться по разным кластерам в зависимости от уровня благополучия, и это же показал результат проведения кластеризации), что регрессионные модели, построенные в кластерах, будут более адекватны и качественны. Проверка этой гипотезы и будет проведена в ходе выполнения этого пункта.

Для всех трёх классов рассматривались три основных вида уравнения регрессии (с учётом модификации):

- линейная модель регрессии:
$$y = \beta_0 + \sum_{i=1}^{k} \beta_i \cdot x_i + \varepsilon;$$
- степенная модель регрессии:
$$y = \beta_0 \cdot \prod_{i=1}^{k} x_i^{\beta_i} \cdot \varepsilon;$$
- экспоненциальная модель регрессии:
$$y = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i \cdot x_i + \varepsilon).$$

Далее для построенных моделей проверялись критерий Акаике, Байесовский информационный критерий, были просмотрены проценты объяснённой дисперсии и качество предсказаний с использованием различных графических пакетов.

9.1 Регрессия в кластере 1: страны со средним уровнем развития (развивающиеся страны)

Лучшей моделью оказалась **экспоненциальная модель**, построенная на регрессоры *Adult_mort*, *Total_exp* и *GDP*. Эта модель объясняет лишь 45% вариации результирующей переменной. Означает это лишь одно: истинная зависимость не приближается ни к линейной, ни к степенной. Это утверждение можно проиллюстрировать с помощью графиков для оцененных уравнений регрессии с каждым регрессором в отдельности, указав в качестве аргумента оценку модели с этими регрессорами:

Рис. 33: Графики для оцененных уравнений регрессии с каждым регрессором

Видно, что истинные зависимости довольно сложные и нелинейные. Графический анализ остатков показывает, что остатки вполне гетероскедастичны, визуально приблизительно нормальны (могло быть и хуже), хотя выборочный квантиль едва приближает теоретический (особенно в хвостах): заметно также и осциллирование посередине. Гистограмма остатков имеет заметную островершинность.

В частности, если говорить о качестве предсказаний по такой модели, видно, что раньше показывалось аналитически: неплохо в общем уловленное истинное распределение данных (примерно такая же картина и для переменной *GDP* и *Total_exp*), хотя некоторые особенности были чрезмерно обобщены. С другой стороны, видно, что модель даже учитывает небольшой кластер точек слева, что положительно её характеризует.
Что неприятно удивило – так это то, что в результате кластеризации не произошло отделение правого кластера точек (предполагалось, что этот кусок отделится). Значит, в многомерном признаковом пространстве этот кластер не выделяется так явно, как в двумерном:

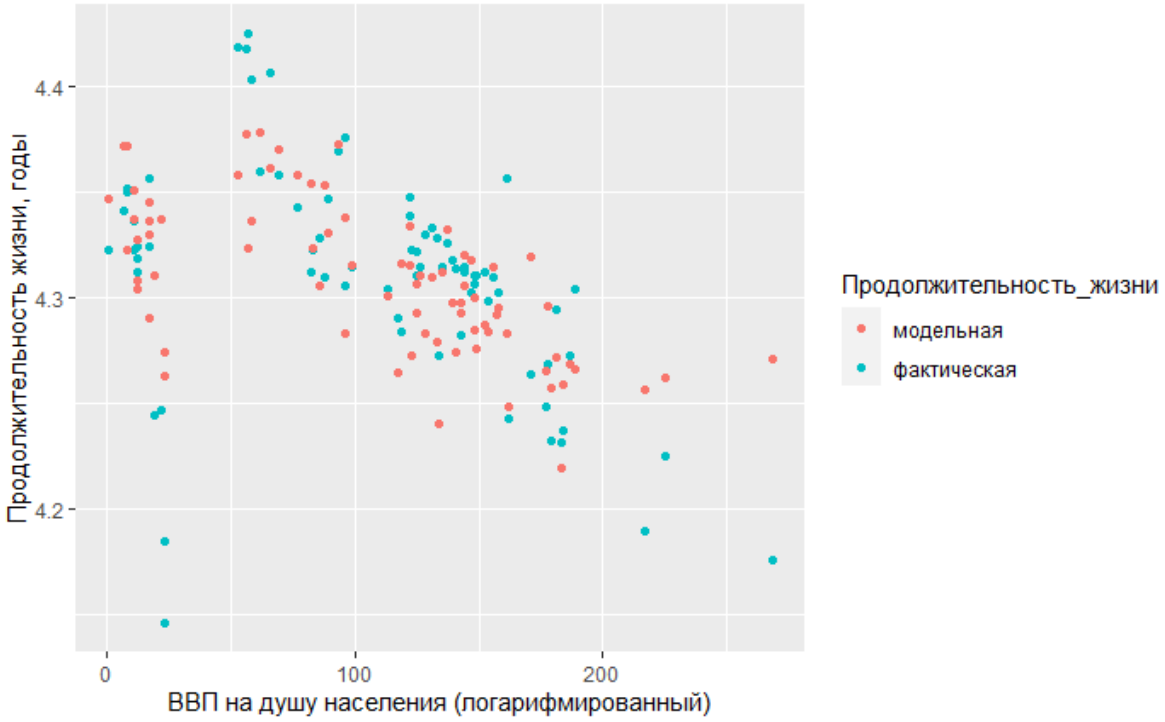

Рис. 34: Диаграмма рассеяния модельных и истинных значений зависимой переменной

Что касается интепретации коэффициентов эластичности:

- уменьшение значения смертности в возрасте 15-60 лет на 1000 человек на 1% ведёт к увеличению продолжительности жизни **в среднем** на 0.026% (мизерное изменение, объясняющееся, впрочем, мизерным изменением объясняющей переменной); .

- увеличение значения процента государственных расходов, выделяемого на здравоохранение (по отношению к общей сумме расходов) на 1% ведёт к увеличению продолжительности жизни **в среднем** на 0.04% – также мизерное изменение, объясняющееся малым варьированием независимой переменной;

- увеличение ВВП на душу населения на 1% ведёт к увеличению продолжительности жизни **в среднем** на 0.35% – в этом случае изменение более существенное, хотя всё равно составляет доли процента (впрочем, если учесть, что такое изменение вызвано процентным изменением объясняющей переменной, можно размышлять о значительном влиянии изменения ВВП на продолжительность жизни, и это логично).

9.2 Регрессия в кластере 2: хорошо развитые страны

Согласно Байесовскому информационному критерию, наилучшая регрессионная модель (со значением Байесовского критерия -133.3) – это степенная модель, построенная на регрессоры *GDP* и *Schooling*. Эта модель объясняет 59% вариации результирующей переменной, что, конечно, лучше, чем было в первом кластере. Для сравнения: регрессия, построенная на все объясняющие переменные, в линейном и нелинейном случае даёт $R^2 = 61\%$, что, разумеется, тоже не особо много. Означает это лишь одно: истинная зависимость, как и в прошлом случае, не аппроксимируется достаточно хорошо ни линейной, ни степенной зависимостями.

Графический анализ оцененных уравнений регрессии с каждым регрессором в отдельности и остатков модели показал, что модель регрессии слабо аппроксимирует истинную зависимость – по крайней мере, направление истинной зависимости (хотя игнорирует сложную нелинейную траекторию с имеющимися у неё подъёмами и спадами: истинная зависимость выбивается из коридора возможных значений линейной регрессии). Качество модели хорошо визуализируют графики истинной зависимости с остатками:

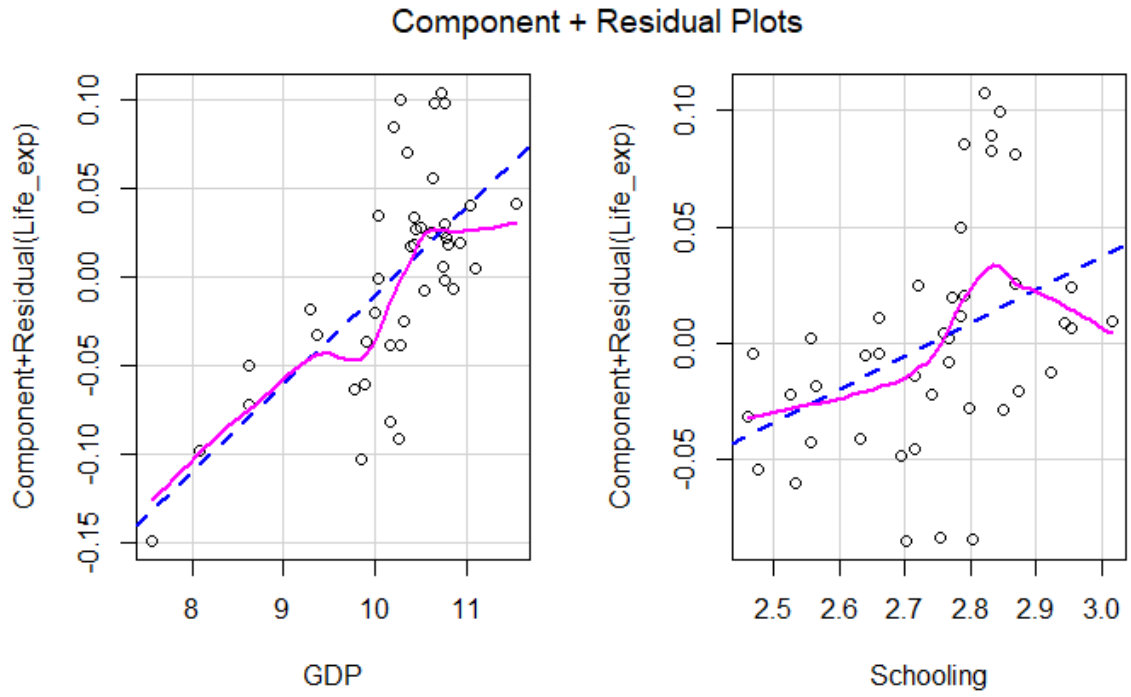

Рис. 35: Графики истинной зависимости с остатками

В частности, если говорить о качестве предсказаний по такой модели, видно, что раньше показывалось аналитически: довольно хороший процент объяснённой дисперсии (примерно такая же картина и для переменной *Schooling*), уловленное моделью истинное распределение (например, в хорошая аппроксимация особенности расположения данных в левом нижнем углу)

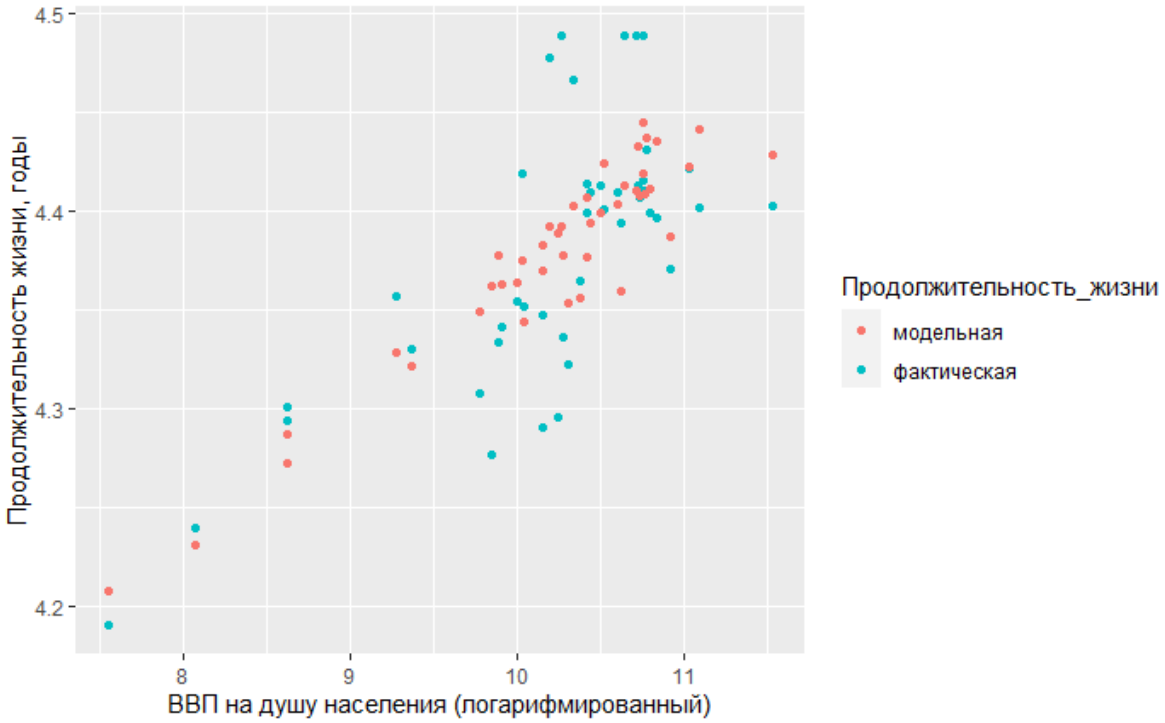

Рис. 36: Диаграмма рассеяния модельных и истинных значений зависимой переменной

Что касается интепретации коэффициентов эластичности:

- увеличение ВВП на душу населения на 1% ведёт к увеличению продолжительности жизни **в среднем** на 0.05% – маленькое изменение, оно аж в 6.8 раз меньше, чем аналогичное изменение в случае регрессионной модели, проведённой в первом кластере. Значит, для стран, находящихся во втором кластере, при построении регрессии продолжительности жизни на выбранный набор переменных независимая переменная *GDP* имеет существенно меньшее влияние на **процентное среднее** изменение продолжительности жизни, чем для стран в первом кластере, и это хорошо объяснимо с экономико-социологической точки зрения: как известно, рост ВВП в хорошо развитых странах имеет не такой высокий темп прироста, как в развающихся странах, поэтому процентное изменение ВВП на душу населения оказывает меньший эффект на продолжительность жизни;

- увеличение количества лет, потраченных на образование, на 1% ведёт к увеличению продолжительности жизни **в среднем** на 0.14%. Также повторим, что здесь, наверное, уместно обратная интерпретация: в странах, где высокая продолжительности жизни, население больше времени тратит на образование, поэтому нельзя сказать наверняка, что при увеличении продолжительности образования вдруг возрастёт продолжительность жизни: вполне имеет место не причинно-следственная связь, а корреляция.

9.3 Регрессия в кластере 3: неразвитые страны

Как и ранее, согласно Байесовскому информационному критерию наилучшая регрессионная модель (со значением Байесовского критерия -140.22) – это экспоненциальная модель, построенная на регрессоры *Adult_ mort* и *Schooling*. Эта модель объясняет 43.7% вариации результирующей переменной, что, конечно, провал по сравнению с предыдущими кластерами (хотя тоже не самый худший возможный показатель). Означает это лишь одно: истинная зависимость, как и в прошлом случае, не аппроксимируется в превосходной степени ни линейной, ни степенной зависимостями.

Графический анализ оцененных уравнений регрессии с каждым регрессором в отдельности и остатков модели показал, что модель регрессии слабо аппроксимирует истинную зависимость – по крайней мере, направление истинной зависимости (хотя игнорирует сложную нелинейную траекторию с имеющимися у неё подъёмами и спадами: истинная зависимость выбивается из коридора возможных значений линейной регрессии). Качество модели хорошо визуализируют графики истинной зависимости с остатками:

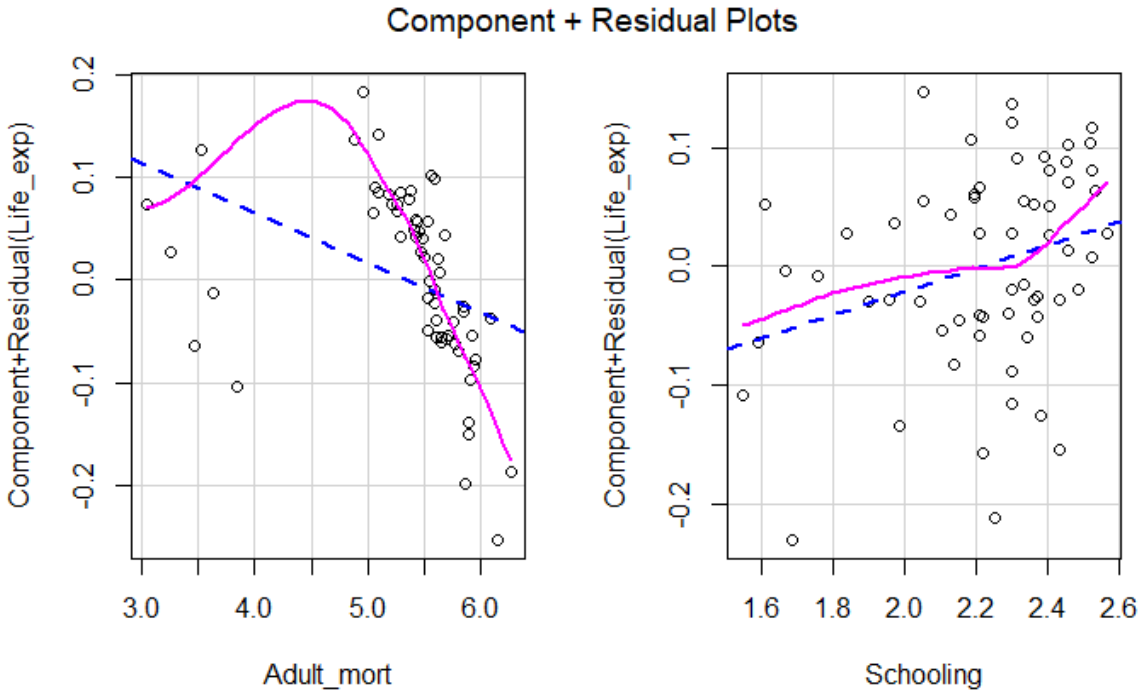

Рис. 37: Графики истинной зависимости с остатками

График предсказанных и истинных значений на диаграмме рассеяния показывает среднее приближение (модель примерно угадывает облако точек):

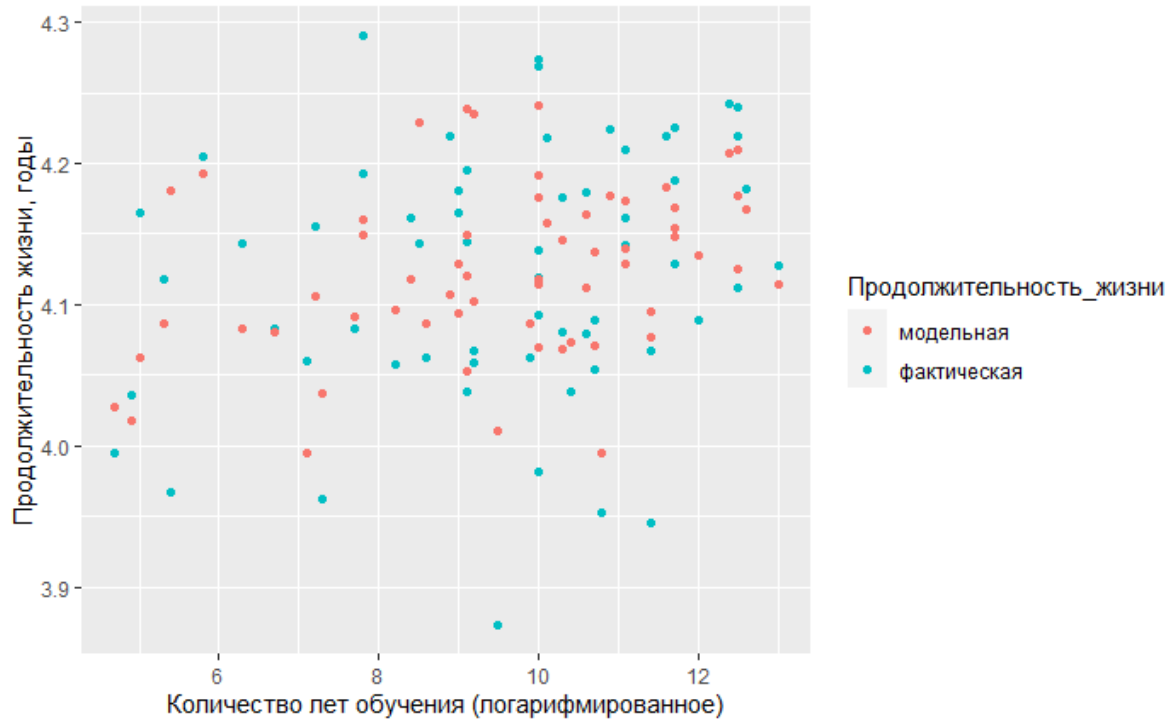

Рис. 38: Диаграмма рассеяния модельных и истинных значений зависимой переменной

Анализ коэффициентов эластичности показывает следующее:

- уменьшение значения смертности в возрасте 15-60 лет на 1000 человек на 1% ведёт к увеличению продолжительности жизни **в среднем** на 0.13%. В этом кластере стран, как видно, изменение смертности сильнее (**в 5 раз** сильнее, если быть точнее) влияет на продолжительность жизни, чем в первом кластере, и это тоже объяснимо: грубо говоря, это показывает, что в неразвитых странах гораздо острее стоит вопрос именно выживания, и именно высокая смертность людей в возрасте 15-60 лет не позволяет (что логично) доживать людям до почтенного возраста;

- увеличение количества лет, потраченных на образование, на 1% ведёт к увеличению продолжительности жизни **в среднем** на 0.12% (чуть менее сильно влияние, чем во втором кластере). Обратим внимание на экономико-социологический смысл полученного нами значения: аналогичный регрессор присутствует в кластере для развитых стран, но там он имеет большее влияние. Есть основания предполагать, что это связано с тем, что уровень образования в развитых странах коррелирует как с общим кругозором, так и с уровнем дохода, который оказывает существенное влияние на продолжительность жизни (почти во всех построенных моделях регрессии именно *GDP* был одним из самых статистически значимых переменных). В случае же неразвитых стран образование в меньшей степени обусловливает доход, который и так имеет довольно низкий уровень. Так что здесь имеет место, надо полагать, транзитивная связь "уровень образования - доход - продолжительность жизни хотя это лишь гипотеза.

9.4 Регрессия по всему набору данных

Уже было выяснено, что лучшей моделью является экспоненциальная модель. Более того, она обгоняет по значению критерия Акаике и Байесовского информационного критерия все предыдущие модели, и при этом объясняет больший процент вариации (R^2 больше, чем у всякой другой построенной нами модели, кроме разве что построенной на все переменные по всему набору данных модель множественной линейной регрессии). Более того, в этой части была построена экспоненциальная модель исключительно на статистически значимые переменные *Adult_mort*, *GDP* и *Schooling* с учётом выбросов по переменной *Dipht*, которые, как выяснилось, **отнюдь не дестабилизируют модель, а, напротив, улучшают её качество**: видимо, при удалении выбросов удаляется некоторая статистически важная часть данных). По компромиссу критерия Акаике, Байесовского критерия и процента объяснённой дисперсии эта модель обходит все рассмотренные, в том числе те, что были рассмотрены в первой компьютерной работе

Анализ коэффициентов показывает нам следующее:

- уменьшение значения смертности в возрасте 15-60 лет на 1000 человек на 1% ведёт к увеличению продолжительности жизни в среднем на 0.08%: видно, что это значение находится в интервале между таковым для кластера развивающихся стран;

- увеличение ВВП на душу населения на 1% ведёт к увеличению продолжительности жизни в среднем на 0.15% – и вновь видно, что это значение располагается в промежутке между значениями для развивающихся и развитых стран;

- увеличение количества лет, потраченных на образование, на 1% ведёт к увеличению продолжительности жизни в среднем на 0.21%. Это значение, напротив, превышает те, что были получены для отдельных кластеров: тем самым для всей совокупности стран получилось большее влияние образования на продолжительность жизни.

9.5 Сопоставление качества построенных моделей для кластеров и всей совокупности

Было построено множество регрессионных моделей, отбор лучших моделей производился в несколько этапов, были использованы как аналитические, так и графические средства проверки качества моделей. И, несмотря на то, что исходной гипотезой было то, что в кластерах регрессионная модель будет вести себя лучше, реальность оказалась иной: и по критерию Акаике (остался за кадром), и по Байесовскому критерию информативности, и по уровню объяснённой дисперсии лидирует – причём с существенным отрывом – модель, построенная на всей совокупности данных. Есть несколько гипотез, которые обосновывают, почему это так:

1. Истинные функциональные зависимости между результирующей переменной и регрессорами не были найдены верно. Действительно, графики явно показывали на достаточный пласт неучтённых зависимостей (особенно речь идёт о графиках для оцененных уравнений регрессии с каждым регрессором в отдельности), а нами были рассмотрены лишь несколько базовых преобразований, позволяющих найти нелинейные зависимости.

2. Исходное признаковое пространство таково, что гораздо более информативным является построение регрессии для всей совокупности, нежели для отдельных её частей. Об этом свидетельствует, например, то, что, несмотря на то, что в регрессионную модель по всей совокупности обобщённо вошли те же регрессоры, что присутствовали в уравнениях регрессии для отдельных кластеров, причём коэффициенты эластичности находятся в промежутках между значениями коэффициентов эластичности для отдельных кластеров. Возможно, имели место некоторые неучтённые зависимости между переменными, которые разрушились при делении на кластеры.

3. Деление на кластеры не было идеальным. Хотя мы предположили, что то распределение данных по кластерам, которые мы получили в результате применения алгоритма k-means вполне логично и оправданно, вполне может быть так, что пространство наблюдений формирует нечёткие и размытые между собой кластеры, которые не меняют радикально ни наклон, ни intercept обобщённой регрессионной модели. Также вполне возможен сценарий, что форма кластеров в n-мерном признаковом пространстве имеет какую-то нелинейную форму. Более того, если взглянуть на кластеры, то видно, что они несколько накладываются друг на друга, что тоже могло сыграть свою роль при качестве построенных в кластерах регрессионных моделей. Об этом свидетельствует также и то, что исходный набор данных был разделён **на 2 класса**, но нами была сделана попытка попробовать другой способ кластеризации на большее количество групп. Возможно, что в том признаковом пространстве, в каком мы находимся, это было не так уместно, как кажется.

Тем не менее, даже анализируя коэффициенты эластичности в таких кластерах в неидеальных моделях, можно сделать интересные и подтверждающиеся здравым смыслом и экономической, социологической парадигмой выводы касательно того, какой аспект человеческого развития и в какой мере влияет на продолжительность жизни в тех или иных странах.

10 LDA

В данном разделе речь пойдет про дискриминантный анализ. Он используется для отнесения новых элементов генеральной совокупности к какому-то из классов при условии, что новый элемент будет относиться к какому-то из известных классов. Для решения задачи используются два основных метода ДА: линейный и вероятностный, по факту, в нашей работе будет

задействовано оба метода, к примеру, построение дискриминантной функции относится к линейному подходу в то время как таблица соответствия предсказанных классов исходным относится к вероятностному подходу.

10.1 Построение дискриминантных функций. Выводы о качестве модели

Так как наши данные уже были логарифмированы, почищены на предмет нулей, там где это требовалось, и стандартизированы для применения евклидовой метрики, то можем приступать к ДА.

Для построения дискриминантной функции воспользуемся результатами кластерного анализа, прибавим вектор значений кластеров к нашему массиву, затем поделим выборку на $\frac{2}{3}$ и $\frac{1}{3}$ и построим саму дискриминантную функцию.

В результате построения дискриминантных функций были получены априорные вероятности: вероятность быть в 1 группе составляет 0.4754098, во второй – 0.2377049, в 3 – 0.2868852. Также был получен вклад построенных дискриминантных функций: LD1 – 0.9387, LD2 – 0.0613 - это та доля, сколько наблюдений каждая из функций помогает определить без построения иных дискриминантных функций.

10.2 Вывод о качестве модели

На первый взгляд, так как вклад первой дискриминатной функции составляет аж 0.9387, что крайне много, а линейных дискриминант у нас всего две, то это говорит о том, что модель крайне неплохо построена, так как она строится так, чтобы сначала максимизировался вклад первой дискриминатной функции (он вышел аж 0.9387), затем – вклад ld2 (понятно, что, так как модели всего две, то на вторую дискриминанту придется оставшаяся доля наблюдений).

10.3 Отнесение новых объектов к выделенным и описанным кластерам различными способами с использованием ДФ

Так как по условию требуется взять 3-4 наблюдения, то прошлые выборки не годятся. Создадим новые выборки. Так как у нас 183 исходных наблюдения, а нам надо отобрать всего лишь 4 для тестовой выборки, то возьмем для тренировочной выборки шаг 1.022346, так как мы берем в нее $\frac{183-4}{183}$ часть от общего числа, затем перевернем полученную дробь, будет $\frac{183}{179}$, что и равно 1.022346.

В итоге применения дискриминантной функции получим, что **из 4 наблюдений 1 попадает в 1 класс, одно в - во второй и 2 попадают в третий класс**.

10.4 Линейный метод

Проведем саначал анализ попадания наших наблюдений с помощью линейного метода:

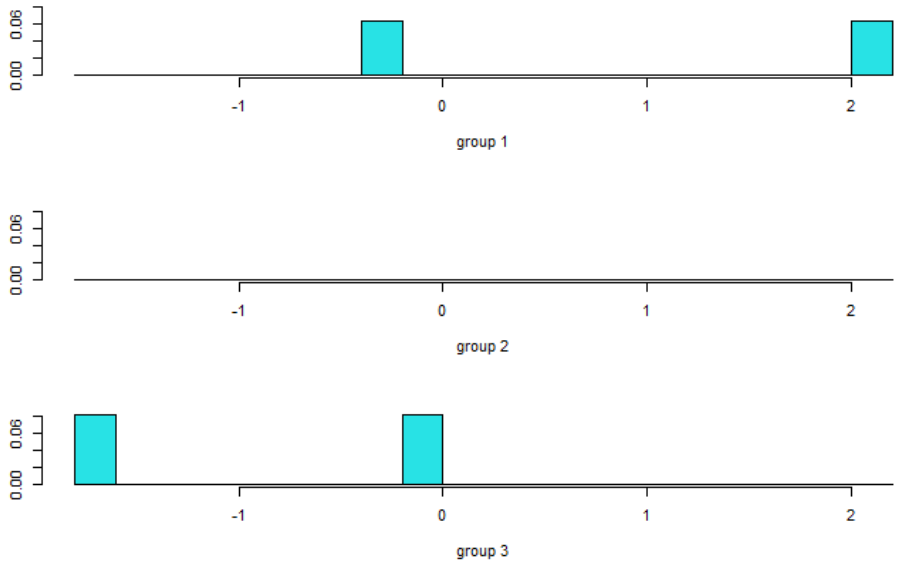

Рис. 39: Диаграмма рассеяния модельных и истинных значений зависимой переменной

Итак, у нас в 1 группу прогнозируются 2 попадания, во 2 ноль попаданий и в 3 группу два попадания. Данный метод работает некорректно, так как один из элементов вместо попадания во 2 группу попал в 1 группу, таким обращом в 1 группе у нас предсказывается 2 наблюдения.

10.5 Вероятностный метод

Теперь перейдем к вероятностному подходу:

```
Classification table:
     obs
pred 1 2 3
   1 0 0 0
   2 0 1 0
   3 1 0 2
Misclassification errors:
  1   2   3
100   0   0
[1] 33.33
```

Рис. 40: Ошибка классификации

Этот метод же показывает, что наша модель работает не идеально на 4 тестовых наблюдениях, так как ошибка определения, как и в предыдущем методе, составляет 0.33. Надежной моделью считается такая, у которой погрешность меньше 0,1. Тем не мене не будем забывать, что по условию просили взять 4 наблюдения, что крайне мало для точного понимания того, насколько классно работает обученная модель, надо брать куда больше значений, чем мы и займемся в одном из следующих пунктов.

10.6 Уточнение результатов классификации, выполненной с помощью метода k-средних, с помощью аппарата дискриминантного анализа

Если мы взглянем на средние значения в k-means и в ДА, то поймем, что в целом модели очень похожи по тому, как они делят объекты на кластеры, что неудивительно:

```
Call:
lda(data.train[, -c(9)], data.train$cl)

Prior probabilities of groups:
        1         2         3
0.4754098 0.2377049 0.2868852

Group means:
  Adult_mort    Alcohol  Total_exp      Dipht        GDP Thinness_5_9 Thinness_10_19  Schooling
1 -0.3439208 -0.1396238 -0.1759366  0.3280507  0.3794418  -0.02272828    0.004845249  0.2581736
2 -0.6933812  1.2915132  0.8778588  0.3332906  0.9781599  -1.21457732   -1.170104430  0.9983736
3  0.7658268 -0.6931238 -0.5051313 -0.6448769 -1.0200167   0.86345826    0.756561420 -1.0950608

Coefficients of linear discriminants:
                      LD1        LD2
Adult_mort      0.3958526 -0.2512907
Alcohol        -0.7711265 -0.8500487
Total_exp      -0.3634558 -0.3602686
Dipht          -0.1111810  0.5032175
GDP            -0.5039811  0.5202433
Thinness_5_9    1.1892464 -0.3117090
Thinness_10_19 -0.2774017  0.5640446
Schooling      -0.4325377  0.3038750

Proportion of trace:
   LD1    LD2
0.9387 0.0613
```

Рис. 41: Результат проведения ЛДА в R

На вопрос о неверной классификации ответим в следующем пункте, потому как пока что была проделана работа только по выборке в 4 наблюдения, а работа по тестовой выборке (которая 1/3 от общей) проделана еще не была.

10.7 Анализ классификационной матрицы (classification matrix). Вывод о качестве разбиения объектов на кластеры

Найдем ошибку предсказания:

```
classification table:
     obs
pred  1  2  3
   1 21  1  2
   2  0 13  0
   3  0  0 24
Misclassification errors:
   1    2    3
0.00 7.14 7.69
[1] 4.95
```

Рис. 42: Ошибка классификации на тестовом множестве

Как мы видим, модель ДА прекрасно справилась с заданием, ведь ошибка составила всего 0.0495 по сравнению с результатами кластерного анализа. Вот только теперь мы можем

точно сказать (в предыдущем пункте было рано сравнивать ДА с КА, так как еще не была построена матрица для выборки $\frac{1}{3}$), что отличий почти нет, так как 0.0495 более чем в 2 раза менее 0.1 что считается максимальной допустимой погрешностью модели.

10.8 Построение графика принадлежности тестовой и тренировочной выборок к кластерам по результатам проведенного анализа

Построим сначала гистограммы, чтобы показать, как разбиваются тренировочные наблюдения на классы:

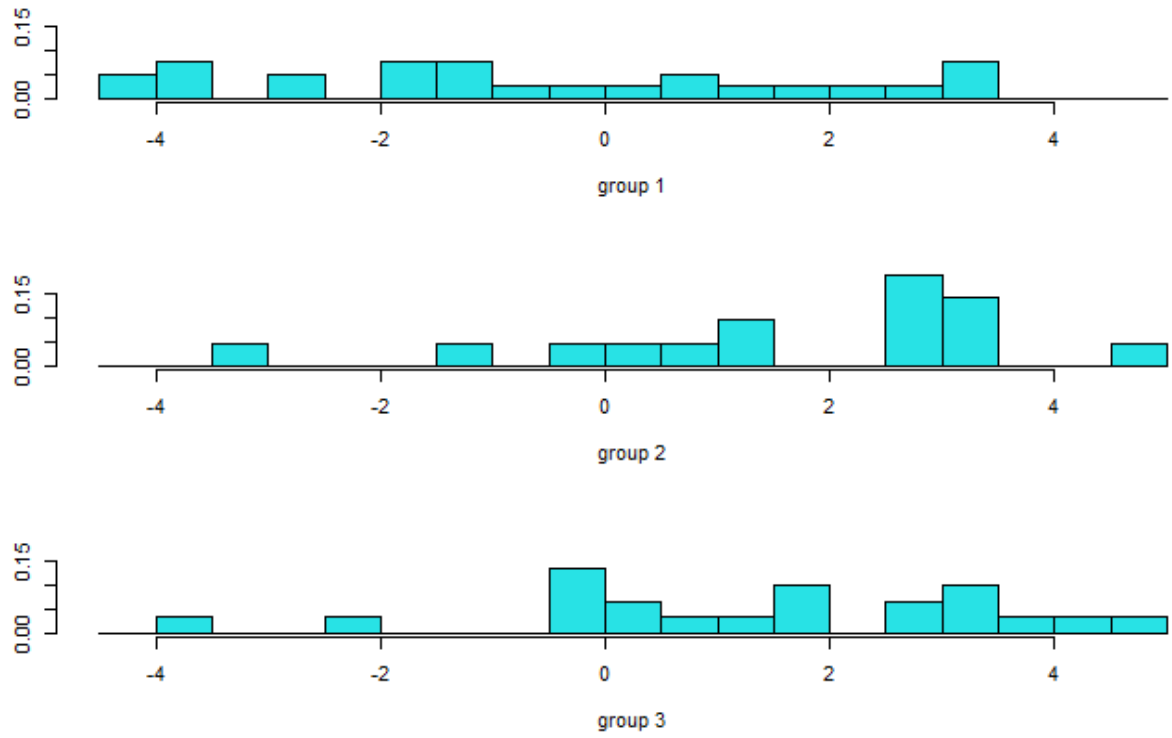

Рис. 43: Разбиение тренировочных наблюдений на классы

Видим, что у разбиений гистограммы отличаются по показателям (мода, смещение и т.д.), это говорит о том, что все приемлемо, кластеры отличны друг от друга.

На гистограмме же второй дискриминантной функции заметно пересечение классов, что не очень хорошо.

Само распределение данных тренировочной выборки в пространстве LD1-LD2 выглядит так:

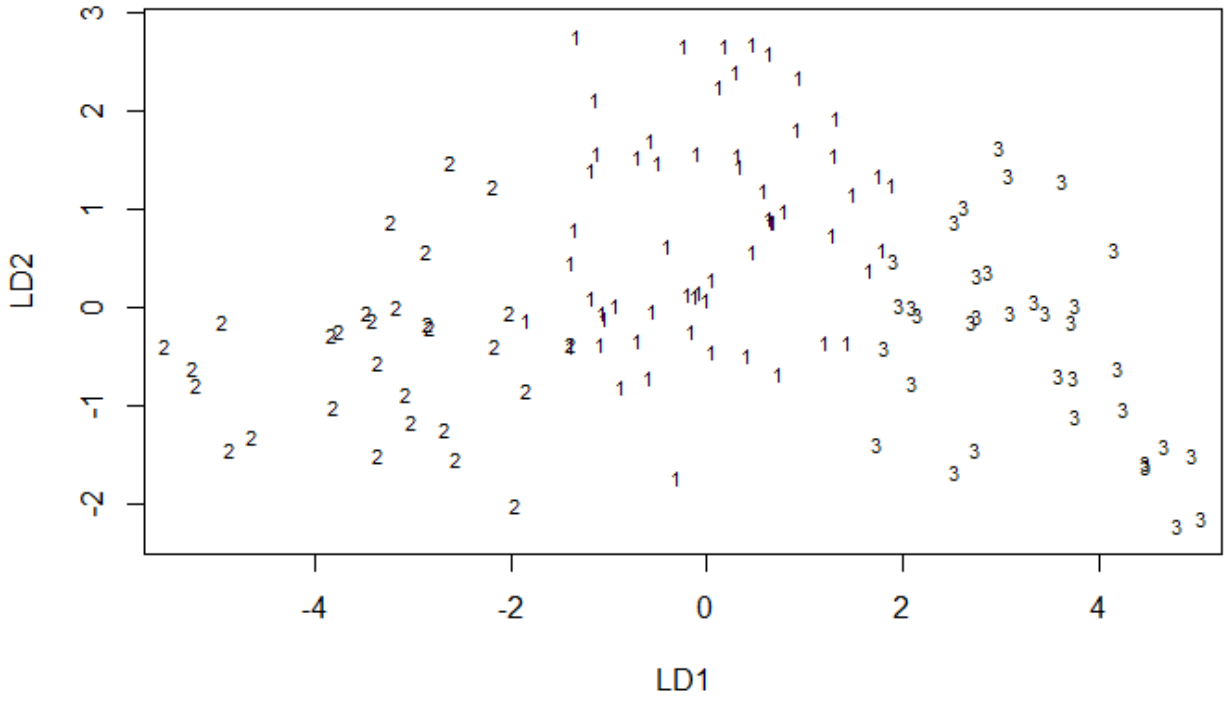

Рис. 44: Распределение данных обучающей выборки в пространстве LD1-LD2

Просматривается вполне четкое деление на классы в обучающей выборке.

Теперь посмотрим разбиение наблюдений в тестовой выборке по первой дискриминанте:

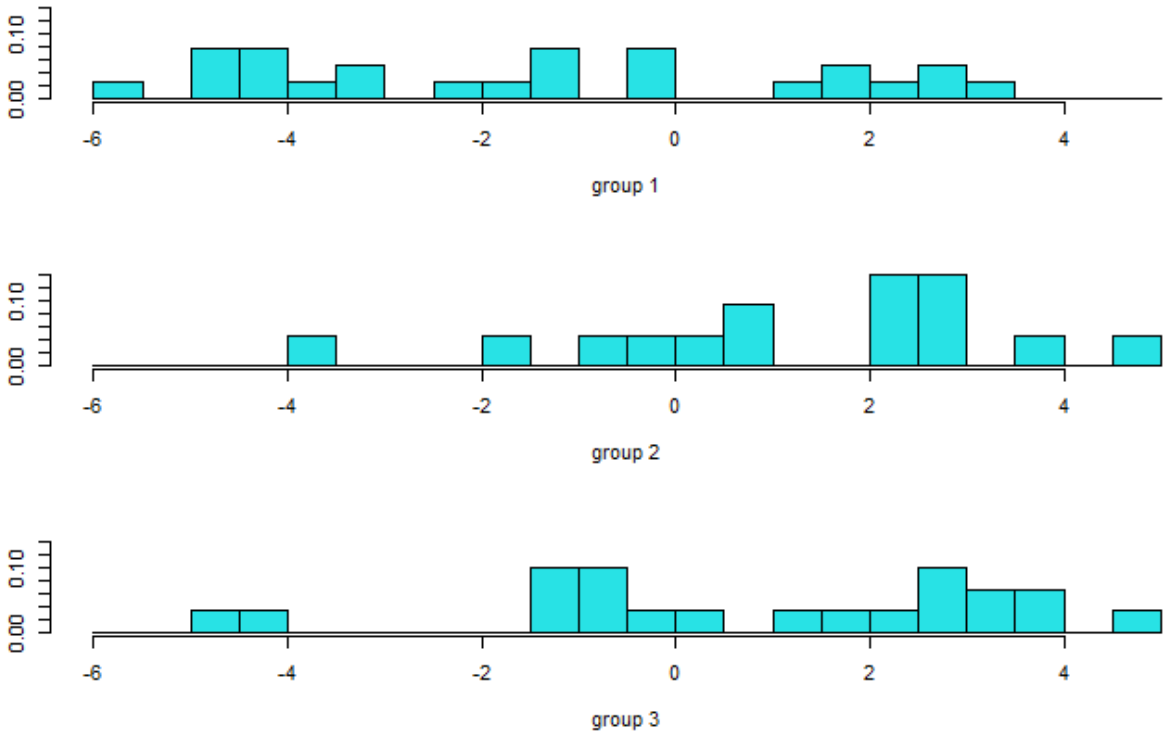

Рис. 45: Распределение данных обучающей выборки в пространстве LD1-LD2

Наблюдаем кардинально отличающиеся друг от друга разбиения, что говорит о хорошем различии между классами, а это, в свою очередь, говорит о качестве нашей модели. По второй

дискриминанте ситуация аналогична, но имеется некоторое смешение классов. Разбиения отличны друг от друга.

Если взглянуть на график рассеяния для тестовой выборки, то можно заметить крайне четкое деление на классы:

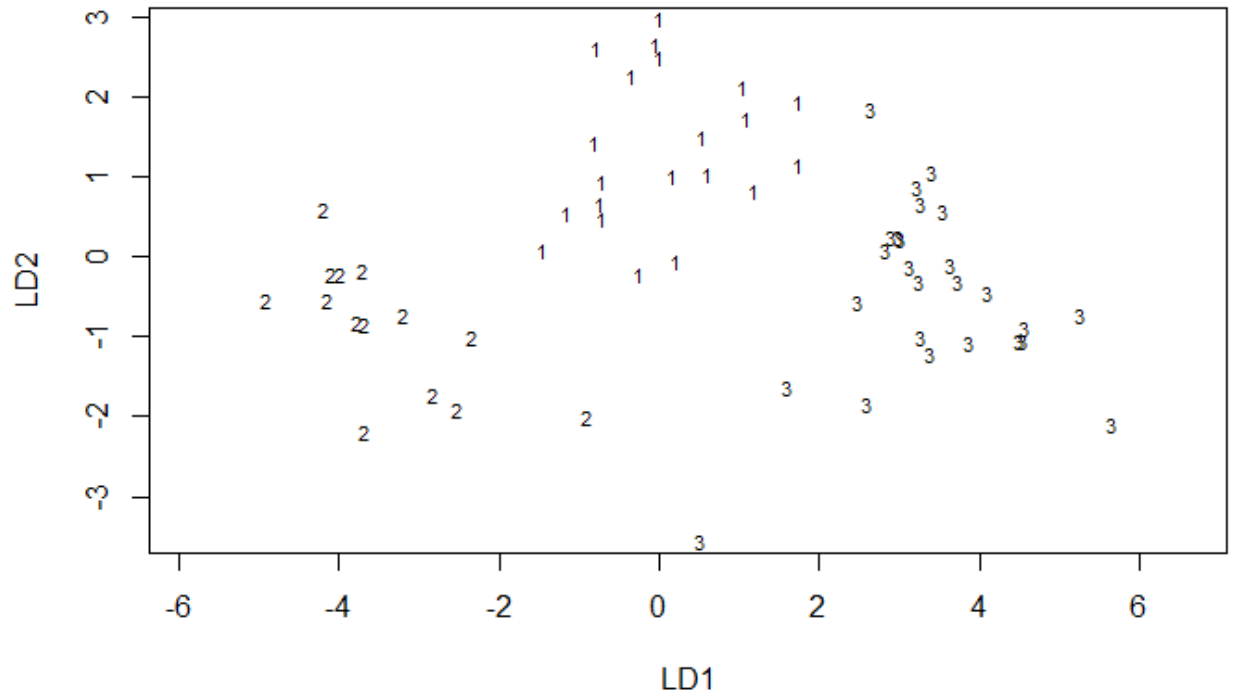

Рис. 46: Распределение данных тестовой выборки в пространстве LD1-LD2

10.9 Выводы по ДА

Дискриминантная функция вышла у нас вполне неплохая, что подтвердилось как высоким значением вклада ld1, так и низкой ошибкой в матрице и распределениями с гистограммами по переменным. Что касается характеристики 4 объектов, то имеющаяся у нас модель по нашей выборке справилась также вполне сносно, учитывая, что наблюдений в тестовой выборке бралось всего 4 (что крайне мало при общем наблюдении в 183), так что тут тоже можно было бы поставить плюс. Что же касается сравнения результатов ДА с результатами кластерного анализа, то и тут я бы сказал, что мы справились уверенно, так как оба статистических подхода к выделению в выборке групп дали почти одинаковые значения средних показателей по переменным в каждом из классов(в ДА) или кластеров (в КА), как таковых значимых отличий в результатах двух подходов не выявлено.

11 Итоги

На этом – завершающем – этапе выполнения компьютерной работы была проведена большая работа с признаковым пространством исходного массива данных: были применены как методы извлечения наиболее информативных переменных с помощью PCA, так и разделение объектов на отдельные кластеры по их признаковому описанию. Были уточнены результаты построения регрессионных моделей в отдельных кластерах и проведён линейный дискриминантый анализ, выясняющий, действительно ли деление на кластеры было качественным. Судя по регрессионным моделям, деление на 2 кластера, от которого в итоге отказались, было бы не менее (а возможно, даже более) репрезентативным. С другой стороны, деление на 3 кластера показывает большую изменчивость стран по продолжительности жизни, а задача любой модели заключается именно в обобщении, а не в чрезмерном упрощении или усложнении реальности.