Исследование детерминант уровня ВВП на душу населения

Отчёт по итогам компьютерной работы по курсу «Многомерные статистические методы»

ВВП на душу населения является одним из ключевых показателей при определении уровня жизни в стране. Однако на этот фактор влияет множество переменных. В проведенном нами исследовании было выбрано 17 показателей (впоследствии их число было сокращено до 11), которые предположительно играют важную роль при оценке качества жизни в стране.

Для анализа мы взяли 142 страны мира, по которым было возможно собрать необходимые данные за 2018 год, взятые с сайта Всемирного банка. Остальные ~80 стран было необходимо убрать ввиду излишне большого количества пропусков в данных, которые было бы сложно восстановить.

Перед нами стоит задача по имеющейся выборке предсказать показатель ВВП на душу населения и понять его детерминанты, в зависимости от следующих объективных факторов:

Обозначение переменной	Переменная
Export	Экспорт товаров и услуг (в % от ВВП)
Consumption	Расходы на конечное потребление (в % от ВВП)
Savings	Валовые сбережния домохозяйств (в % от ВВП)
National Expenditure	Государственные расходы
Import	Импорт товаров и услуг (в % от ВВП)
Labor among youth	Коэффициент участия в рабочей силе в возрасте 15-24 лет (%)
Employment in services	Доля занятых в сфере услуг (%)
Employment in industry	Доля занятых в промышленном секторе (%)
Mortality rate under 5	Уровень смертности детей в возрасте до 5 лет (%)
Population in largest city	Население крупнейшего города (в % от городского населения страны)
University enrollment	Доля вовлеченности в профессиональное образование (%)
Unemployment	Общий уровень безработицы (%)
Urban population	Городское население (% от общего населения)

Целевые переменные, выбранные для нашего анализа, взяты из принципа их условной однородности, дабы максимально избавиться от масштабирования значений, обусловленных большим населением ряда стран. Таким образом, мы можем хотя бы примерно сравнить такие страны, как, например, Конго и США, делая скидку на размеры располагаемых бюджетов соответствующих правительств. Объясним принцип выбора ключевых переменных, использованных в нашей работе:

1. Экспорт предположительно демонстрирует то, что международная торговля повышает уровень жизни в государстве, так как приносит дополнительный доход.
2. Доля расходов на конечное потребление предположительно обуславливает отрицательную взаимосвязь с ВВП на душу населения: чем выше расходы, тем ниже уровень частных сбережений, что является индикатором наличия возможности у граждан подумать о завтрашнем дне, не будучи обременёнными одними лишь проблемами текущего дня.
3. Коэффициент участия в рабочей силе в возрасте 15-24 лет - экономическая вовлеченность потенциально необразованного населения демонстрирует, что граждане, которые потенциально должны заниматься образованием, вынуждены (или хотят) работать. Может также являться индикатором момента наступления зрелости граждан страны.
4. Занятость в промышленности и секторе услуг (% от рабочей силы государства) - эти показатели отражают процесс перехода государства к новой модели мира, в рамках которой физический труд всё больше и больше отходит на второй план.
5. Смертность среди детей до 5 лет (на 1000 рождённых), на наш взгляд, хорошо отражает уровень развития медицины (как инфраструктуры, так и квалификации врачей) в стране.
6. Население крупнейшего города (в % от городского населения страны) демонстрирует обратную зависимость с равномерностью развития городов в государстве, влияющего на уровень жизни в провинциях и желании развивать человеческий капитал в миноритарных регионах.
7. Коэффициент вовлеченности в профессиональное образование был взят с целью показать, что профессиональное образование положительно влияет на уровень жизни в стране, повышая количество квалифицированных рабочих кадров.
8. Общий уровень безработицы очевидно, негативно влияет на уровень благосостояния граждан.
9. Городское население (% от общего населения) предполжительно является дополнительным индикатором уровня индустриализации страны и положительно влияет на уровень ВВП на душу населения.

Изучим характеристики формы распределений переменных:

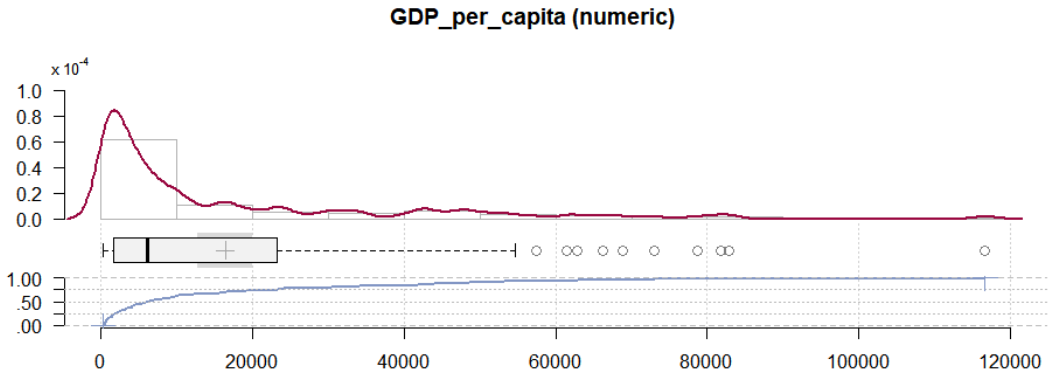
Переменные	Среднее	Медиана	Стандартное отклонение	Асимметрия	Эксцесс
Экспорт	44,27	36,03	33,53	2,42	7,43
Потребление	78,72	78,07	15,55	0,64	3,58
Участие молодежи в рабочей силе	43,65	42,15	13,68	0,41	-0,55

Занятость в секторе услуг	55,94	57,61	17,73	-0,41	-0,53
Занятость в секторе промышленности	20,16	19,77	8,23	0,40	1,10
Доля населения в крупнейшем городе от всего городского населения	34,26	31,95	18,42	1,03	1,37
Смертность среди детей до 5 лет	27,98	14,30	30,53	1,30	0,64
Общий уровень безработицы	6,66	5,10	5,06	1,62	2,80
Коэффициент вовлеченности в профессиональное образование	43,67	43,69	29,20	0,18	-1,36
Доля городского населения от всего населения страны	62,62	64,52	21,93	-0,32	-0,76

Изучая асимметрию, у всех признаков наблюдается правосторонняя асимметрия, за исключением сбережений, занятости в сфере услуг, городского населения и роста ВВП. Также хочется подметить, что у всех признаков показатель асимметрии значителен, кроме доли поступивших в высшие учебные заведения.

Коэффициент эксцесса показывает отрицательные значения у показателей городского населения, занятости подростков в возрасте 17-24 лет, занятости в сфере услуг, что говорит об их островершинном распределении. В свою очередь, крайне выделяется на фоне остальных значений плосковершинность экспорта, свидетельствующий о почти равномерном распределении этого значения среди остальных показателей.

Диагностика выбросов в целевой переменной

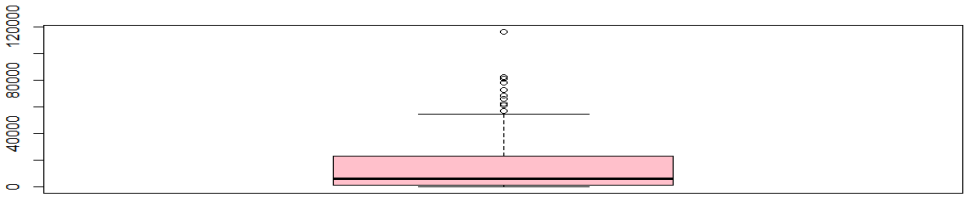


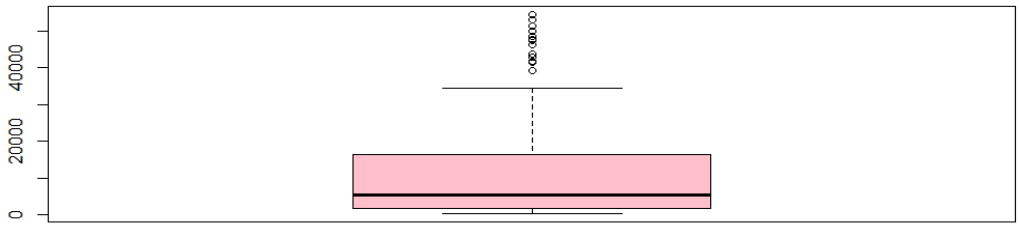
Разброс ВВП на душу населения по методу IQR составил 21 423 доллара в год, что является крайне большой разницей (при условии, что среднее значение целевой переменной составляет 16410, а медиана – 6184), свидетельствующей о достаточном

сильном неравенстве благополучия между странами, что в принципе и так ожидаемо. Относительный показатель квартильной вариации для нашей целевой переменной составил 1.73, что в несколько раз превышает граничное значение, при котором мы могли бы назвать наши данные однородными, что также свидетельствует о весомой разнице между ВВП на душу населения между странами мира.

Проверив наличие выбросов и аномалий из данных тремя методами (3σ, 1.5IQR & 3IQR), мы ожидаемо получили несколько разные результаты о наблюдениях, которые следует признавать выбросами (первый метод признал выбросами две наиболее успешных страны мира, второй – 11 подобных стран, заключительный – лишь Люксмебург). Но, исходя из соображений логики и визуального анализа распределения нашей переменной, выбрали наиболее жёсткий метод 1.5IQR, удаляющий из выборки наибольшее количество наблюдений.

Таким образом, мы исключили из рассмотрения такие страны, как Люксембург, Швейцария, Сингапур, США, Великобритания, Австралия, Ирландия, Норвегия, Дания, Исландия и Катар. Сравним ящичковые диаграммы для целевой переменной до и после удаления выбросов:

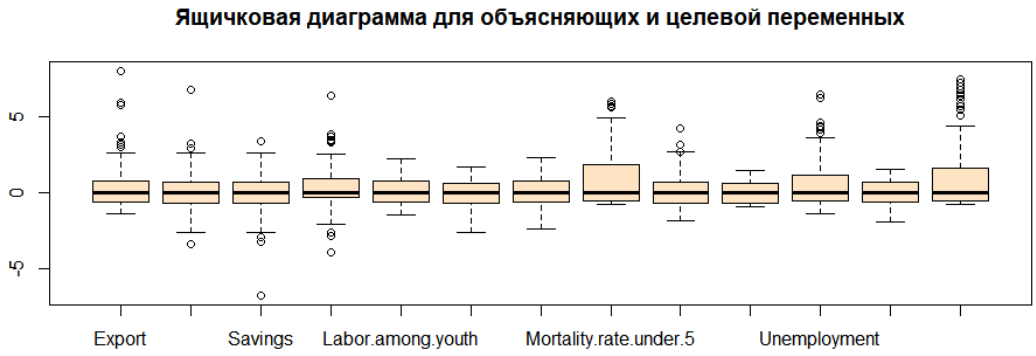




Видим, что после удаления выбросов у нас образовались новые выбросы (относительно новых квартилей данных), которые, возможно, также следовало бы последовательно убирать вплоть до момента, когда у нас больше не будет выбросов значений относительно предполагаемого максимума. Однако попытка сделать даже три подобные итерации привела бы к сокращению имеющихся для рассмотрения наблюдений до несущственной в смысле проведения регрессионного анализа, сохранив при этом в себе некоторые выбросы.

Стандартизуем очищенные от выбросов в целевой переменной данные и проверим, какие выбросы существуют в других переменных. Заранее оговорим, что мы не будем

убирать наблюдения со значительными выбросами среди объясняющих переменных, поскольку это хоть и влияет на предсказательную силу этих переменных, но чересчур сильно сожмёт нашу выборку, что также негативно повлияет на качество модели. Названия всех переменных на рисунке ниже не указаны ввиду отсутствия функции поворота подписей данных к диаграмме в R на 90 градусов.

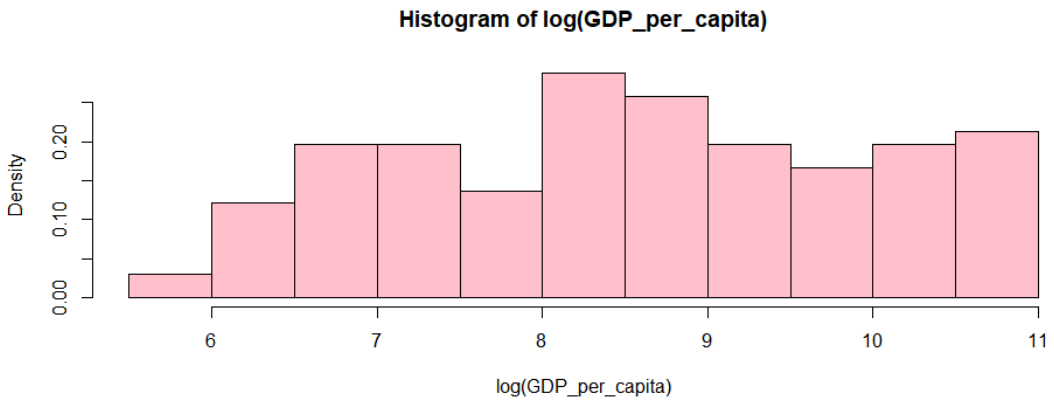


Воспользовавшись тестом Граббса для проверки H_0, предполагающей, что максимальное или минимальное наблюдение в выборке является типичным наблюдением, мы получили p-value=0.2 и 1 соответственно, то есть на любом адекватном уровне значимости гипотеза о наличии выбросов в данных отвергается с вероятностью ошибки первого рода α.

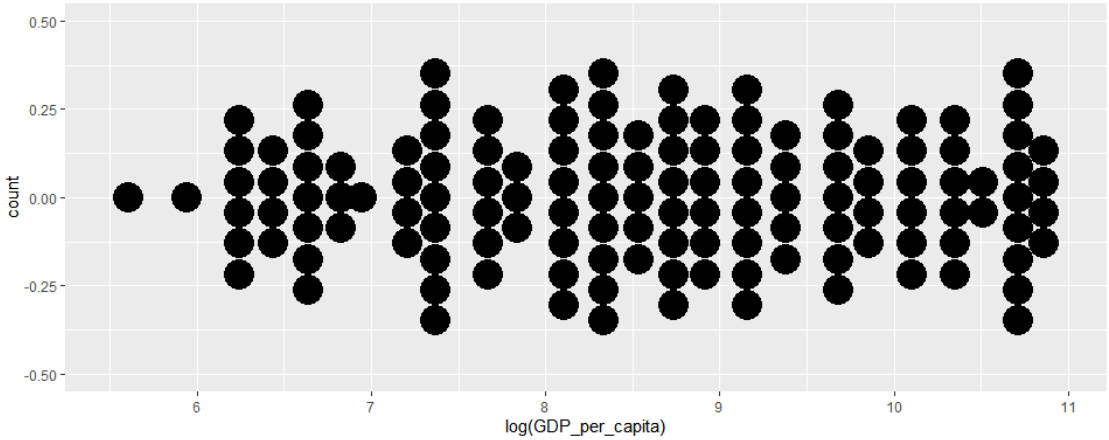
В результате проведения теста Рознера (H_0: k наибольших наблюдений взяты из той же генеральной совокупности, что и $(n-k)$ первых наблюдений) мы также удостоверились в отсутствии выбросов для очищенных данных. Таким образом мы можем сделать вывод, что, применив правило 1.5IQR, мы избавились от выбросов и можем идти дальше. Это очень важно для возможности дальнейшего анализа.

Проверка соответствия эмпирического распределения нормальному распределению

После логарифмирования целевой переменной при группировке наблюдений по методу Стёрджеса мы получили нечто хотя бы отдалённо напоминающее график нормального распределения для нашей целевой переменной.



Построив точечную диаграмму, мы вновь можем убедиться в том, что логарифмированные данные имеют структуру, похожую на нормальное распределение, несколько стремящееся к равномерному.



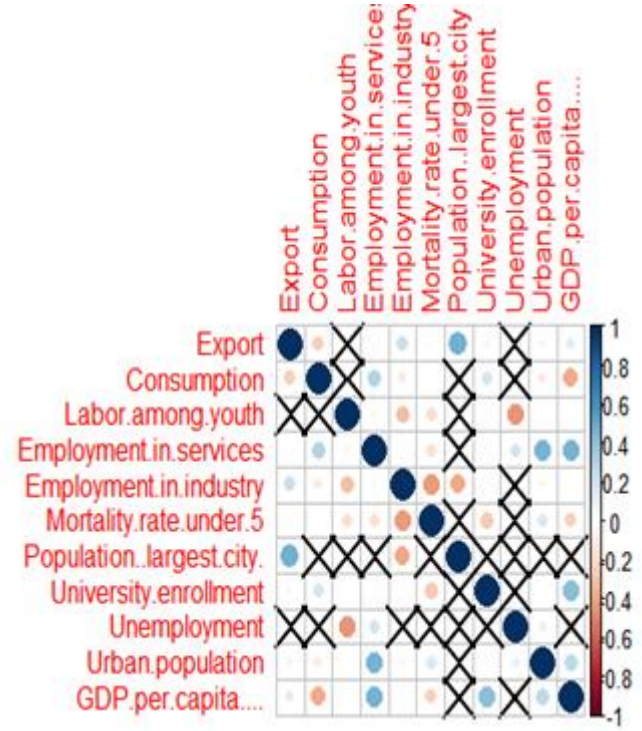
Для проверки соответствия эмпирических распределений некоторым теоретическим используются критерии согласия. Для того, чтобы было возможно применить все последующие многомерные статистические методы, нам необходимо проверить гипотезу о нормальности распределения целевой переменной:

Применив сначала наиболее простой критерий согласия - критерий Пирсона, мы получили p-value=0.85. Традиционный метод дал крайне высокое значение p-value, не совсем соответствующее увиденному на графиках даже после линейных преобразований. Поэтому мы применили наиболее мощные критерии - Жарке-Бера и Колмогорова-Смирнова, не отвергнув которые можно будет смело использовать гипотезу о нормальности распределения при дальнейшем анализе. Первый тест показал p-value=0.05, второй – 0.17.

Наиболее мощные тесты дают несколько разные результаты при проверке целевой переменной на нормальность, но, в целом, на уровне значимости 0.05, который принято считать наибольшим адекватным значением ошибки первого рода, гипотезы не отвергаются. Поэтому мы можем принять решение о нормальности полученных через логарифмирование ВВП на душу населения данных.

Корелляционный анализ

Визуализируем результаты расчёта матрицы частных коэффициентов корреляций для целевой и тех объясняющих переменных, которые не оказались функционально зависимыми друг от друга:



Результаты расчетов показывают, что ни один из признаков не имеет значение коэффициента, указывающего на умеренную или сильную связь, цвет и радиус кружочков отражает направление и силу взаимозависимости признаков, крестик представляет те парные корреляции, которые не были признаны значимыми на уровне 0.05.

Частные коэффициенты корреляции всех переменных модели в абсолютном значении оказались меньше парных коэффициентов. Это говорит о том, что взаимосвязь признаков с ВВП на душу населения обусловлена воздействием на них фиксируемых факторов. Исключением является только безработица: частный коэффициент корреляции больше парного для данного показателя, то есть остальные признаки ослабляют связь между рассматриваемыми двумя. Для заданной надежности доверительные оценки частных коэффициентов не указывают на наличие сильной степени сопряженности ВВП на душу населения и какого-либо показателя при постоянных значениях других параметров.

Важно заметить, что доверительные интервалы коэффициентов корреляции между такими переменными, как экспорт, занятость среди молодежи, занятость в промышленности и численность городского населения, и целевой переменной включают 0, то есть частные коэффициенты корреляции перечисленных показателей статистически незначимы. Интервальные и точечные оценки частных коэффициентов показывают, что наибольшая степень связи имеет место между ВВП и занятостью в сфере услуг, зачислениями в вузы и потреблением соответственно.

Множественный коэффициент корреляции для нашей целевой переменной, отражающий тесноту связи между результативной переменной и всеми остальными факторами в наборе данных, составляет 0,9406. Это достаточно хороший результат, показывающий тот факт, что мы можем практически полностью объяснить разницу между странами в ВВП на душу населения за счёт рассматриваемых объясняющих переменных. Максимально достижимое значение предсказательной силы будущих регрессионных моделей не будет превышать этого значения.

Это очень хороший вывод, показывающий наличие действительной пользы для понимания факторов, определяющих значение нашей целевой переменной на основе взятых факторов. Также это означает, что мы выбрали верные предпосылки для данной работы и выделили верные влияющие переменные.

Регресионный анализ. Линейная регресионная модель

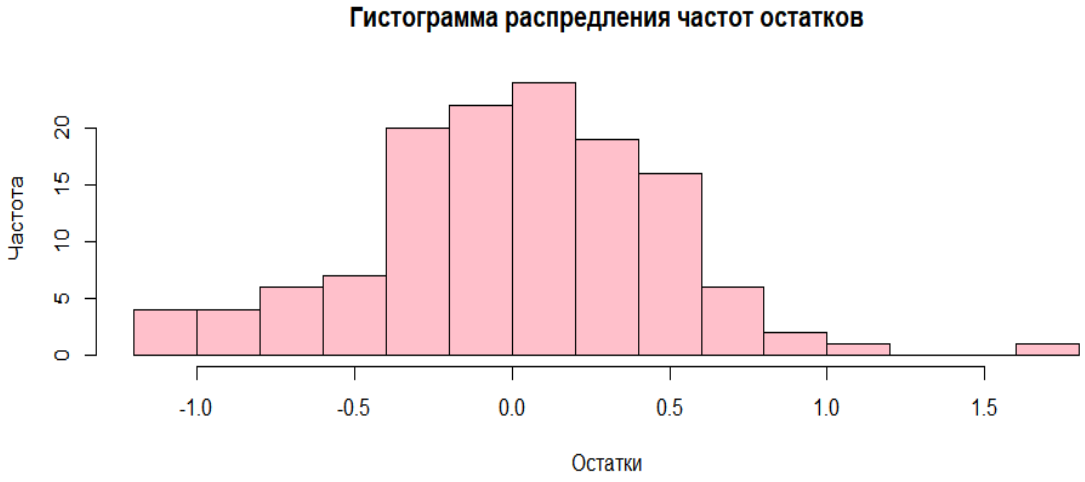
Дабы не загромождать отчёт множеством графиков и моделей, оставим приведённые в оригинале компьютерной работы десятки регрессионных моделей и оставим в ней лишь оптимальные.

Построим линейную модель множественной регрессии со всеми имеющимися в нашем распоряжении объясняющими переменными:

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.2973066  0.5368353  13.593  < 2e-16 ***
Export                      0.0029867  0.0020357   1.467  0.14491
Consumption                -0.0158201  0.0035603  -4.443 1.97e-05 ***
Labor.among.youth          -0.0005431  0.0041082  -0.132  0.89505
Employment.in.services      0.0322569  0.0056426   5.717 7.97e-08 ***
Employment.in.industry     -0.0013263  0.0083159  -0.159  0.87355
Mortality.rate.under.5     -0.0068409  0.0025187  -2.716  0.00758 **
Population..largest.city. -0.0046526  0.0030256  -1.538  0.12672
University.enrollment       0.0122084  0.0024434   4.996 1.99e-06 ***
Unemployment               -0.0218995  0.0101770  -2.152  0.03340 *
Urban.population            0.0115591  0.0039205   2.948  0.00383 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4955 on 121 degrees of freedom
Multiple R-squared:  0.8848, Adjusted R-squared:  0.8753
F-statistic: 92.94 on 10 and 121 DF,  p-value: < 2.2e-16
```

Значимы только Consumption (потребление),Employment.in.services (доля занятых в сфере услуг), Mortality.rate.under.5 (уровень смертности), University.enrollment (количество поступивших в университет), unemployment (уровень безработицы) и Urban.population (доля городского населения). Данная модель имеет крайне большое значение скорректированного R^2, что говорит о её высокой предсказательной силе.

Одним из важнейших этапов валидации результатов является анализ остатков, проверим гипотезу H_0 о нормальном распределении остатков:



Визуально кажется, что остатки действительно распределены нормально и имеется лишь один значительный выброс при прогнозировании логарифма ВВП на душу населения одной из наиболее развитых стран среди оставшихся в выборке. В дальнейшем все модели будут базироваться на предположении о нормальности остатков.

Результаты тестов Пирсона и Жарке-Бера показали достаточно большое значение p-value (0.77 и 0.38 соответственно), что говорит о том, что нулевая гипотеза о принадлежности остатков нормальному закону распределения действительно не отвергается.

Примем получившуюся модель за лучшую и запишем уравнение регрессии формально:
$Y = 7{,}297 + 0{,}003x_1 - 0{,}016x_2 - 0{,}0005x_3 + 0{,}032x_4 - 0{,}001x_5 - 0{,}007x_6 - 0{,}005x_7 + 0{,}012x_8 - 0{,}021x_9 + 0{,}012x_{10}$,

где $x_1 - Export, x_2 - Consumption, x_3 - Labor.among.youth, x_4 - Employment.in.services, x_5 - Employment.in.industry, x_6 - Mortality.rate.under.5, x_7 - Population..largest.city., x_8 - University.enrollment, x_9 - Unemployment, x_{10} - Urban.population$

Найдем для данной модели коэффициенты эластичности и проинтерпретируем их.

```
[x1] 0.014
[x2] -0.148
[x3] -0.003
[x4] 0.205
[x5] -0.003
[x6] -0.024
[x7] -0.018
[x8] 0.06
[x9] -0.018
[x10] 0.082
```

Коэффициент эластичности показывает, на сколько процентов в среднем изменяется результативный признак y при изменении факторного признака x на 1%. Таким образом, лишь 4 переменные имеют положительное влияние на результативный признак.

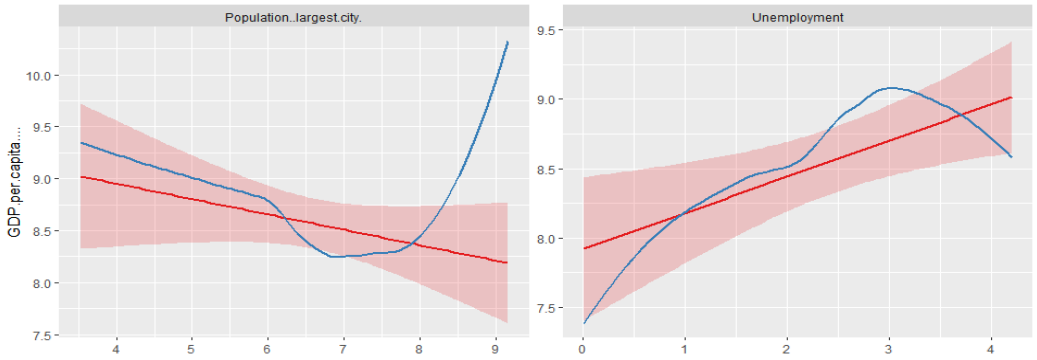
Примечательно, что при увеличении безработицы на 1% модель предсказывает увеличение и показателя ВВП, хотя логически должно быть наоборот, однако данная ошибка будет исправлена нами в следующей части данной работы путем изменения характера связи данной переменной на нелинейную.

Оставшиеся 6 коэффициентов эластичности являются отрицательными. Это значит, что при увеличении факторного признака, результативный уменьшится. Из отрицательных коэффициентов эластичности наиболее значим тот, что относится к потреблению, это может быть объяснено тем, что при росте потребления сокращается доля сбережений и инвестиций домохозяйств, что косвенно может снизить совокупный выпуск государства.

Но стоит отметить, что абсолютное значение всех коэффициентов крайне мало, нет ни одного, превышающего значения в 0,205; из чего можно сделать вывод о несущественности влияния факторов на результативный.

Нелинейная (степенная) регрессионная модель

Проанализировав все одномерные регрессионные модели, мы попробовали с помощью Desmos подогнать примерные выходы модели в зависимости от значений показателя безработицы и населения в наибольшем городе, однако новая регрессия не дала существенного улучшения результатов (исходные графики регрессий для этих признаков приведены ниже):



Попробовав прологарифмировать все имеющиеся объясняющие переменные, чтобы привести все данные к одному виду, мы получили результаты, в соответствии с которыми потребление и занятость в сфере услуг имели значительно отклоняющиеся от других переменных коэффициенты. Отказавшись от логарифмирования этих двух переменных, мы получили следующие результаты:

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.563838   1.080683   7.924 1.27e-12 ***
Export                     0.010245   0.085379   0.120  0.90469
Consumption               -0.017706   0.003180  -5.567 1.58e-07 ***
Labor.among.youth         -0.125814   0.149461  -0.842  0.40157
Employment.in.services     0.028232   0.005061   5.578 1.51e-07 ***
Employment.in.industry    -0.229790   0.104222  -2.205  0.02936 *
Mortality.rate.under.5    -0.545869   0.069062  -7.904 1.41e-12 ***
Population..largest.city. -0.062423   0.074590  -0.837  0.40431
University.enrollment      0.057315   0.075153   0.763  0.44716
Unemployment              -0.175463   0.064997  -2.700  0.00794 **
Urban.population           0.697118   0.167759   4.155 6.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4418 on 121 degrees of freedom
Multiple R-squared:  0.9085, Adjusted R-squared:  0.9009
F-statistic: 120.1 on 10 and 121 DF,  p-value: < 2.2e-16
```

Сравнивая с лучшей линейной моделью, мы видим, что стандартная ошибка остатков лучшей нелинейной модели, в которой прологарифмированы все переменные, кроме доли населения, занятой в сфере услуг и потребления, уменьшилась на 0.051 и скорректированный R^2 увеличился на 0.026, что является достаточно прорывным результатом, показывающим эффективность проведённых преобразований. Эту модель было принято считать лучшей в классе нелинейных, она принимает следующий вид:

$$\ln Y = 13.473 + 0{,}04\ln x_1 - 1{,}398x_2 - 0{,}129\ln x_3 + 0{,}028x_4 - 0{,}23\ln x_5 - 0{,}562\ln x_6 - 0.071\ln x_7 + 0{,}049\ln x_8 - 0{,}176\ln x_9 + 0.671\ln x_{10}$$

где $x_1 - Export, x_2 - Consumption, x_3 - Labor.among.youth, x_4 - Employment.in.services, x_5 - Employment.in.industry, x_6 - Mortality.rate.under.5, x_7 - Population..largest.city, x_8 - University.enrollment, x_9 - Unemployment, x_{10} - Urban.population$

Промежуточные выводы по итогам первой части работы:

Увеличение благосостояния общества, выражаемое в виде ВВП на душу населения, становится возможным, если мы обеспечим устойчивое развитие инфраструктуры городов, стимулируя граждан заселяться не только в наиболее крупном городе. Также крайне важно улучшать качество медицины, поскольку это является ключом к сохранению человеческого капитала государства. Сокращение потребления населения, вызванное увеличением сбережений, является также позитивным индикатором для развития экономики страны ввиду того, что возможность сберегать в принципе отражает то, что не все доходы населения идут на сиюсекундное потребление, поэтому развитие стимулов к увеличению сбережений является весьма важной задачей государства для увеличения благосостояния её горожан. Уменьшение доли занятых в производстве и соответственное увеличение доли занятых в сфере услуг также является очень важной задачей, которую можно достичь через увеличение субсидий,

направленных на автоматизацию производства и открытие малых и средних бизнесов, не связанных с ручным трудом. Также очевидно, что создание рабочих мест и сокращение безработицы является позитивным фактором для благосостояния горожан.

Особенно примечательно, что столь важные взаимосвязи стали прослеживаться после логарифмирования почти всех переменных, что можно интерпретировать математически. Убывающая отдача от масштаба объясняющих переменных, описываемая функцией логарифма, свидетельствует о критической важности начальных стадий развития указанных факторов. Например, если в государстве почти нет городов и большинство людей проживают в мелких поселениях без доступа к широкой информации, развитие инфраструктуры до уровня городов существенно раскроет интеллектуальный потенциал близлежащего населения и поспособствует развитию страны в целом. Однако если государство уже достаточно развито, вряд ли будет какой-либо толк пытаться увеличить долю городского населения с 96% до 97%.

Вывод касательно потребления и сбережений кажется несколько неверным ввиду того, что вероятнее всего, на деле причино-следственная связь работает наоборот: если граждане живут достаточно благополучно, то они могут позволить себе не только выживать, проживая “от зарплаты до зарплаты”, но и откладывать на будущие крупные приобретения. Исходя из приведённых соображений, представим ещё одну модель, в которой не рассматривается потребление:

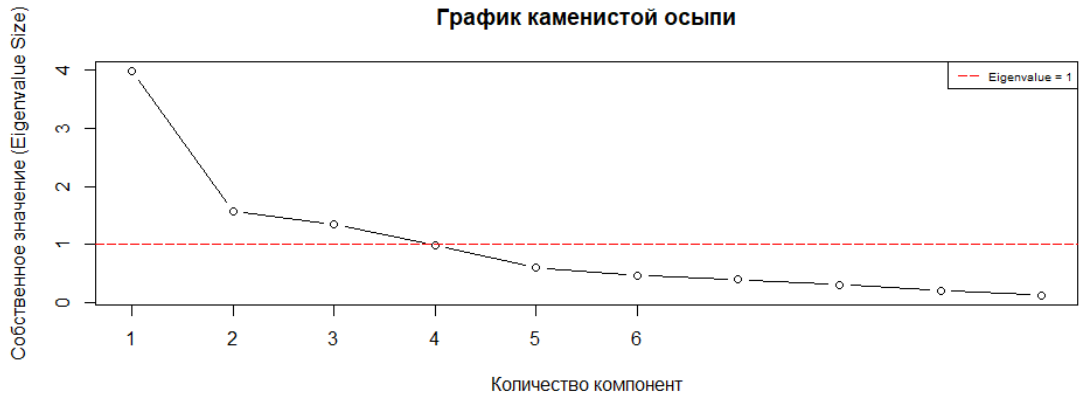
```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.494818   1.132673   5.734 7.25e-08 ***
Export                     0.166045   0.090035   1.844 0.067575 .
Labor.among.youth         -0.147376   0.166770  -0.884 0.378593
Employment.in.services     0.026678   0.005641   4.730 6.10e-06 ***
Employment.in.industry    -0.137041   0.114835  -1.193 0.235037
Mortality.rate.under.5    -0.565976   0.076980  -7.352 2.46e-11 ***
Population..largest.city. -0.130769   0.082121  -1.592 0.113886
University.enrollment      0.007133   0.083279   0.086 0.931883
Unemployment              -0.249651   0.071007  -3.516 0.000616 ***
Urban.population           0.839977   0.185047   4.539 1.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4931 on 122 degrees of freedom
Multiple R-squared:  0.885,  Adjusted R-squared:  0.8765
F-statistic: 104.3 on 9 and 122 DF,  p-value: < 2.2e-16
```

Выделение главных компонент

Представим результаты вычисления главных компонент на отмасштабированных объясняющих переменных:

Номер компоненты	Собственное значение	Доля объясненной дисперсии	Доля накопленного уровня объясненной дисперсии
1	3,98	39,79	39,79
2	1,58	15,75	55,55
3	1,34	13,43	68,97
4	0,98	9,82	78,79
5	0,60	6,01	84,80
6	0,47	4,67	89,47
7	0,40	4,00	93,47
8	0,31	3,12	96,59
9	0,21	2,14	98,73
10	0,13	1,27	100,00

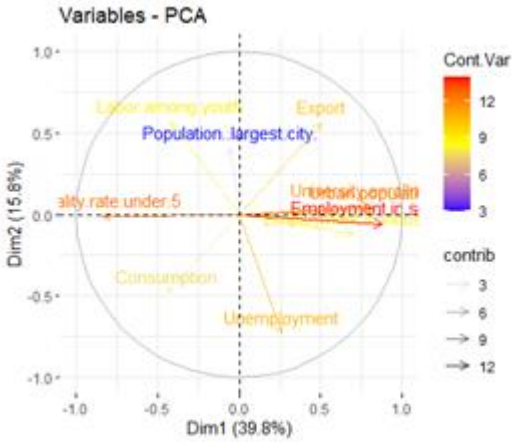
Очевидно, что в соответствии критерием Кайзера нам надо выбрать 3 главных компоненты, поскольку у первых трёх собственное значение больше 1. Однако следует обратить внимание на то, что у последней из главных компонент собственное значение лишь на несколько сотых меньше единицы. Поэтому необходимо принять решение о включении этой компоненты на основании дополнительного метода, воспользуемся критерием каменистой осыпи (Кэттеля) и взглянем на график отсортированных по убыванию собственных значений:



Видим, что по критерию Кэттеля нам следует оставить всего 2 главных компоненты, поскольку дальше разница между смежными собственными значениями крайне невелика. Поскольку в данном случае предполагаемая разница между анализом на двух и трёх главных компонентах будет крайне существенна - воспользуемся третьим методом для принятия окончательного решения, посмотрим на график накопленной объяснённой дисперсии. Накопленная дисперсия превышает критическое значение 70% при четрёх главных компонентах, потому следует принять следующее решение: ищем среднее арифметическое от оптимального количествах главных компонент по версии разных методов и получаем (2+3+4)/2=3. Именно столько мы и будем использовать для нашего анализа, тогда как остальные компоненты следует принять за “факториальную осыпь”.

Визуализируем зависимость трёх главных компонент от исходных объясняющих переменных:

	PC1	PC2	PC3
Export	0.5	0.56	0.32
Consumption	-0.44	-0.48	0.2
Labor.among.youth	-0.42	0.57	-0.3
Employment.in.services	0.88	-0.06	0.15
Employment.in.industry	0.7	-0.12	-0.41
Mortality.rate.under.5	-0.83	-0.01	0.23
Population..largest.city.	-0.05	0.4	0.82
University.enrollment	0.8	0.05	-0.09
Unemployment	0.26	-0.73	0.39
Urban.population	0.83	0.04	0.19



Мы видим, что первая главная компонента главным образом зависит от занятости в сфере услуг (положительно), смертности детей младше 5 лет (отрицательно), доли городского населения и доли поступающих в высшие учебные заведения среди соответствующей возрастной группы. Показательно, что именно эти параметры лучше всего обуславливают разницу между странами, очевидно, эта компонента войдёт в уравнение регрессии с положительным коэффициентом.

Вторая компонента меньше скореллирована с исходными признаками, но главным образом обуславливается отрицательной зависимостью с безработицей и положительной - с трудоустройством среди молодёжи и уровнем экспорта (вероятно, будет иметь положительный коэффициент в регрессии).

Последняя рассматриваемая нами компонента главным образом зависит от доли населения, сосредоточенной в наибольшем по размеру городе, лишь незначительно завися от остальных переменных (по остальным зависимостях можно предположить, что она также войдёт в итоговое уравнение с положительным коэффициентом). Важно отметить, что все три главных компоненты не описывают значительной зависимости с потреблением, хотя и все каким-то образом зависят от него.

Построение уравнения регрессии с использованием выделенных главных компонент

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.364e-16  3.562e-02   0.000    1.000
components$Dim.1  4.501e-01  1.786e-02  25.207  < 2e-16 ***
components$Dim.2  1.154e-01  2.838e-02   4.066  8.3e-05 ***
components$Dim.3 -4.544e-02  3.074e-02  -1.478    0.142
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4093 on 128 degrees of freedom
Multiple R-squared:  0.8363, Adjusted R-squared:  0.8325
F-statistic:   218 on 3 and 128 DF,  p-value: < 2.2e-16
```

Нелинейная модель на главных компонентах оказалась качественно хуже своего “собрата” по всем рассматриваемым метрикам качества (Residual St.Error & Adjusted R^2).

Попробуем использовать продвинутые критерии сравнения и воспользуемся информационными критериями для сравнения представленных в работе регрессионных моделей:

```
AIC(lm) 201.76

AIC(nlm1) 171.43

AIC(nlm2) 199.54

AIC(lm_pca) 144.68

BIC(lm) 236.36

BIC(nlm1) 206.03

BIC(nlm2) 231.25

BIC(lm_pca) 159.10
```

Мы видим, что и в соответствии с информационным критерием Акайке, и Байесовским критерием Шварца наилучшей моделью будет линейная модель на главных компонентах. По критериям скорректированного R^2 нелинейная модель с сохранением переменной потребления остаётся лучшей, имея наилучшую предсказательную силу.

Таким образом, сравнив качество модели через продвинутые информационные критерии, модель, полученная по результатам метода главных компонент, будет на порядок лучше, поскольку она выделяется очень важным свойством отсутствия коллинеарности и имеет высокую предсказательную силу, используя крайне небольшое количество объясняющих переменных. При этом модель обладает достаточно высоким уровнем логарифмической функции правдоподобия.

Кластерный анализ

Применим правило выбора числа кластеров, основанное на WSS (elbow method). Оптимальным будет то число кластеров, после которого убывание WSS начинает замедляться.

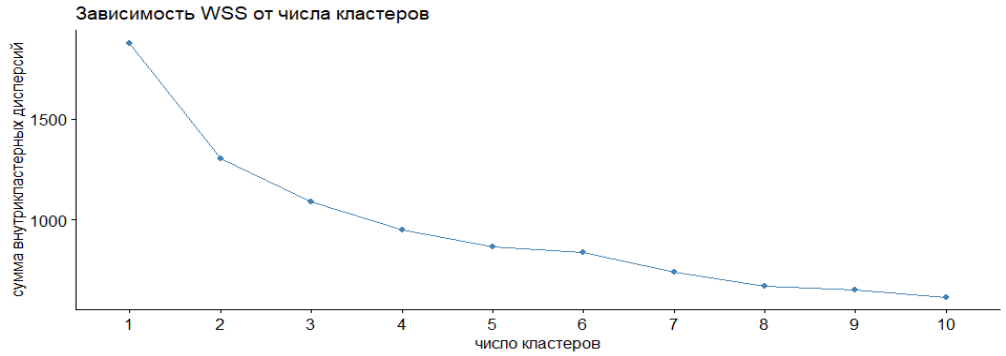
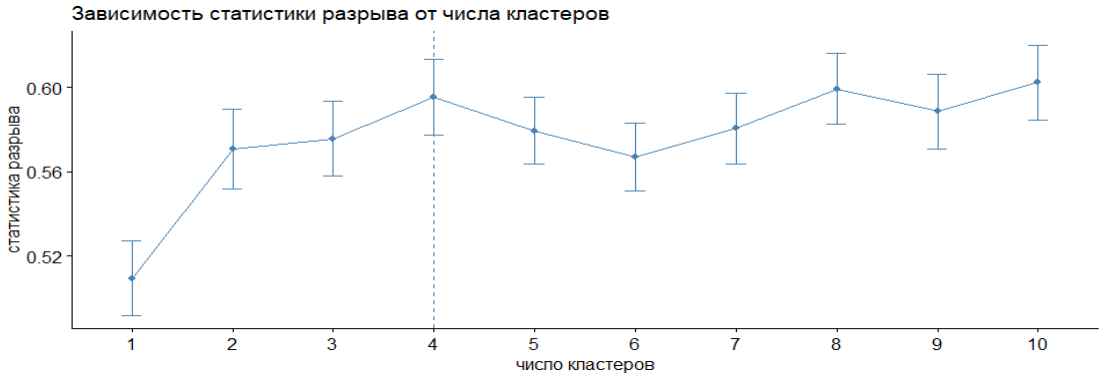


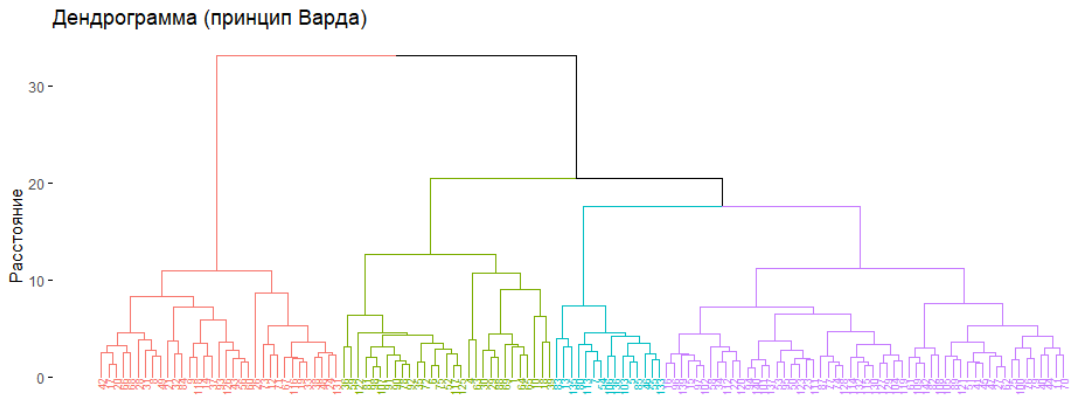
График показывает, что оптимальным будет примерно 4-5 кластеров, поскольку именно там скорость изменения суммы внутрикластерных дисперсий значительно уменьшается.

Также рассчитаем статистику разрыва:



В соответствии с методом локтя и статистики разрыва, 4 кластера будут являться оптимальным количеством.

Дендрограмма иллюстрирует поэтапный процесс кластеризации, отображая по горизонтальной оси объекты, в нашем случае, страны, а по вертикальной – расстояния.



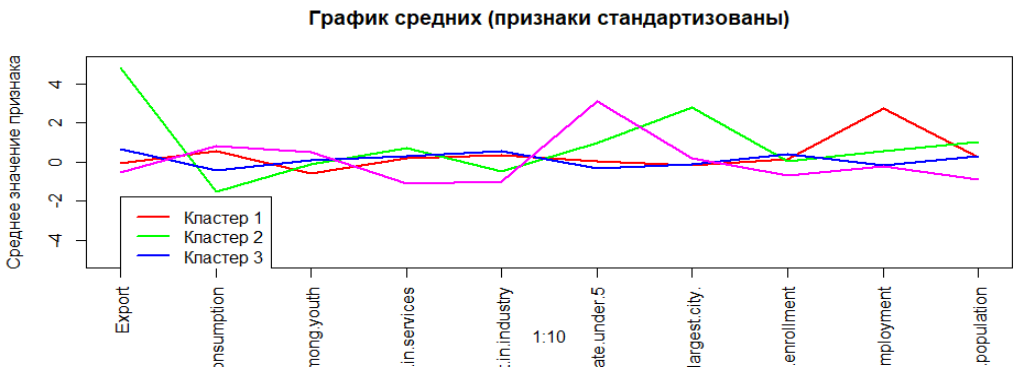
Как видно из дендрограммы, построенной через метод Варда, наиболее однородным является третий кластер, так как объединения стран в данном кластере происходило на наименьших расстояниях. Третий кластер содержит только небольшие государства, каждое из которых за 2018 год имеет положительные темпы роста ВВП. В первый кластер входят в основном африканские страны, исключениями являются Афганистан и Непал. Второй кластер объединяет страны, расположенные в Восточной Европе, и в Юго – Восточной Азии. В последний класс вошли страны, в которых невысокий уровень безработицы и, как правило, высокие расходы на потребление.

Мы не будем представлять иные дендрограммы, поскольку они либо дали достаточно похожий по своей сути результат, либо не могут быть легко проинтерпретированы, либо не дали в принципе адекватных результатов кластеризации.

Использование метода к-средних для классификации объектов:

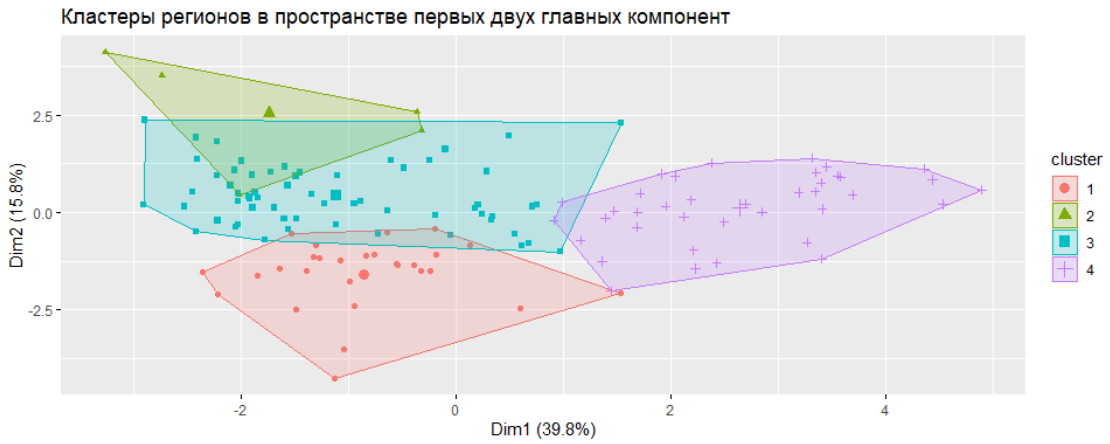
В результате применения метода получаем 4 кластера разной размерности, наименьший включает в себя 5 стран, в то время как крупнейший - 61 страну.

Визуализируем результаты кластеризации через график средних. По графику средних даем интерпретацию полученным кластерам.



Таким образом, самый малочисленный кластер (кластер 2) из 5 стран характеризуется наибольшей долей экспорта в отличие от других групп, что неудивительно, учитывая размер регионов, входящих в кластер. Также значительно возвышается над другими кластерами значение численности населения крупнейшего города, так, например, население Гонконга превышает 7,5 миллионов (из которых ввиду особенностей

площади страны в крупнейшем городе живёт более 99% жителей) и высока детская смертность. Минимален в данном кластере показатель потребления. Кроме того, примечательно распределение рабочей силы, так, в странах второго кластера самая высокая доля занятых в сфере услуг, в то время как доля занятых в индустриальной сфере находится на втором месте с конца, это может быть обосновано максимальной долей городского населения, так как в городе преимущественно преобладает занятость в сфере услуг.



Из графика видно разбиение регионов по кластерам в пространстве двух первых главных компонент. Учитывая наше знание о знаке перед коэффициентами этих двух компонент, мы можем утверждать, что линии уровня благосостояния государств в данном пространстве будут идти вправо вверх (хотя угол наклона идёт ближе к виртуальной оси OX и соответствует примерно 20 градусам), поскольку обе компоненты значимы и положительно влияют на ВВП на душу населения.

Поэтому можем заключить, что первый кластер соответствует странам, наименее преуспевающим в своём экономическом развитии; второй - одним из наиболее успешных стран мира среди тех, ВВП на душу населения которых не было принято считать выбросами; третий - среднестатистический кластер, внутри которого содержатся развивающиеся страны мира; четвёртый - наиболее успешные страны мира, показатель ВВП на душу населения которых существенно отличается от остальных кластеров.

Построение регрессионных моделей в кластерах (типологическая регрессия)

При построении типологических регрессий мы старались отобрать для каждого из кластеров только значимые переменные, при этом по возможности максимально усилив её предсказательную силу.

Регрессия для первого кластера

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.08490    0.05792   1.466 0.155700
Consumption             -0.14820    0.04404  -3.365 0.002569 **
Employment.in.services   0.41933    0.09186   4.565 0.000126 ***
Mortality.rate.under.5  -0.30827    0.06560  -4.699 8.92e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.252 on 24 degrees of freedom
Multiple R-squared:  0.7222, Adjusted R-squared:  0.6874
F-statistic: 20.79 on 3 and 24 DF,  p-value: 7.355e-07
```

Внутри кластера стран третьего мира (в соответствии с интерпретацией кластеров в пространстве двух первых компонент) мы можем определить благосостояние проживающих там жителей на основании переменных потребления (чем меньше потребляют и, соответственно, больше сберегают, тем лучше качество жизни там), занятости в сфере услуг (страны движутся в верном направлении, развивая сферу услуг) и смертности детей младше 5 лет (развитие медицины позволяет значимо отличиться и двигаться в сторону выхода из данного кластера).

$y = -0.35 - 0.24x_1 + 0.23x_2 - 0.44x_3$

$x_1 - Consumption, x_2 - employment.in.services, x_3 - Mortality.rate.under.5$

Но при условиях полученных коэффициентов, остаётся очевидным, что если просто создавать рабочие места по профессиям типа продавцов, парикмахеров и т.д. делая большие ставки по сберегательным депозитам в странах по типу Зимбабве - вряд ли это хоть как-то положительно повлияет на уровень жизни в стране, поскольку тут скорее обратная зависимость: если в стране высокий уровень жизни, то у её граждан есть возможность сходить в магазин, постричься вне дома и оставить какие-то сбережения для будущих лет.

Регрессия для второго кластера

Включим в модель только ту переменную, которые в наименьшей степени характеризуют второй кластер (пользуясь выводом, полученным по результатам регрессии для первого кластера), это доля поступающих в университеты. Берём только одну объясняющую переменную, поскольку большее количество нарушало бы все предпосылки для построения линейной регрессии.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.42849    0.07390   5.799  0.01021 *
University.enrollment  0.97779    0.09713  10.067  0.00209 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1646 on 3 degrees of freedom
Multiple R-squared:  0.9712, Adjusted R-squared:  0.9617
F-statistic: 101.3 on 1 and 3 DF,  p-value: 0.002087
```

$y = 0.42849 + 0.97779x, x - University.enrollment$

Таким образом, мы можем сделать вывод, что увеличение доли поступающих в университете в кластере одних из наиболее развитых стран мира существенно влияет на уровень ВВП на душу населения, что говорит о необходимости углубленно заниматься развитием института образования в развитых странах.

Регрессия для третьего кластера

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.08490    0.05792   1.466 0.155700
Consumption            -0.14820    0.04404  -3.365 0.002569 **
Employment.in.services  0.41933    0.09186   4.565 0.000126 ***
Mortality.rate.under.5 -0.30827    0.06560  -4.699 8.92e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.252 on 24 degrees of freedom
Multiple R-squared:  0.7222, Adjusted R-squared:  0.6874
F-statistic: 20.79 on 3 and 24 DF,  p-value: 7.355e-07
```

Показательно, что в первом и третьем кластере значимыми оказались одни и те же переменные. Это можно легко объяснить тем фактом, что занятость в сфере услуг, возможность сберегать доходы и уровень развития медицины наибольшим образом коррелируют с нашими переменными, также имея при этом наиболее показательные коэффициенты эластичности.

Регрессия для четвертого кластера

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.06189    0.19175   0.323  0.74896
Consumption           -0.13297    0.03400  -3.911  0.00045 ***
Mortality.rate.under.5 -0.05760    0.03270  -1.762  0.08770 .
University.enrollment  0.47743    0.23650   2.019  0.05196 .
Unemployment          -0.07037    0.05322  -1.322  0.19549
Urban.population       0.47595    0.08739   5.446 5.43e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2699 on 32 degrees of freedom
Multiple R-squared:  0.6381, Adjusted R-squared:  0.5816
F-statistic: 11.28 on 5 and 32 DF,  p-value: 2.499e-06
```

Лучшая модель для 4-го кластера включает в себя 5 переменных-предикторов и имеет уравнение вида: $y = 0.06189 + 0.47743x_1 + 0.47595x_2 - 0.13297x_3 - 0.07037x_4 - 0.0576x_5$

где $x_1 - University.enrollment, x_2 - Urban.population, x_3 - Consumption, x_4 - Unemployment, x_5 - Mortality.rate.under.5$

Примечательно, что, в отличие от других кластеров, в четвертом уровень детской смертности играет слабую роль в образовании целевой переменной. Также необходимо отметить, что для данной модели, в отличии от модели на 10 переменных, выполняются необходимые требования для проведения регрессионного анализа по соотношению количества наблюдений и объясняющих переменных.

Дискриминантный анализ

Полученная по итогам компьютерной работы дискриминантная функция:

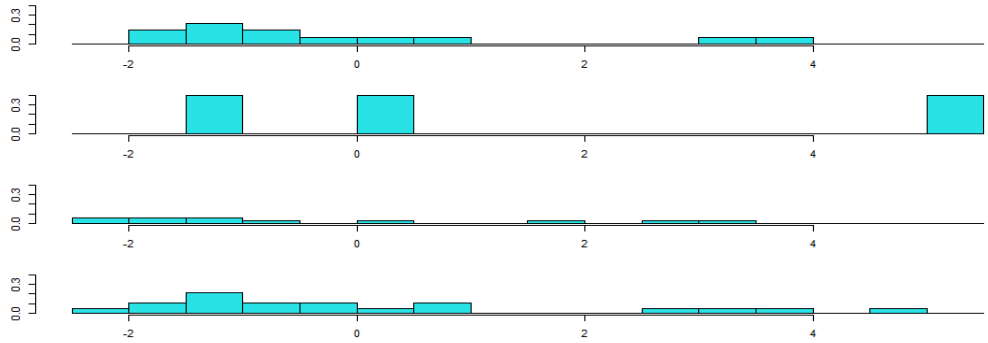
```
Coefficients of linear discriminants:
                                  LD1         LD2         LD3
Export                     -0.13735177 -0.74841365  0.59798556
Consumption                -0.08372618  0.27395398 -0.05329539
Labor.among.youth           0.19393574  0.13514647  0.21123303
Employment.in.services      0.23935736 -0.68084016 -0.40136452
Employment.in.industry     -0.23571091 -0.01947286 -0.35178428
Mortality.rate.under.5      0.87164154  0.08466651  0.11119043
Population..largest.city.   0.08631207 -0.11380682  0.13088948
University.enrollment       0.16163650  0.20660661 -0.82054110
Unemployment               -0.47489263  0.62995443  0.53176204
Urban.population           -0.61432294  0.37161362  0.54248244
```

```
Proportion of trace:
   LD1    LD2    LD3
0.5863 0.2737 0.1400
```

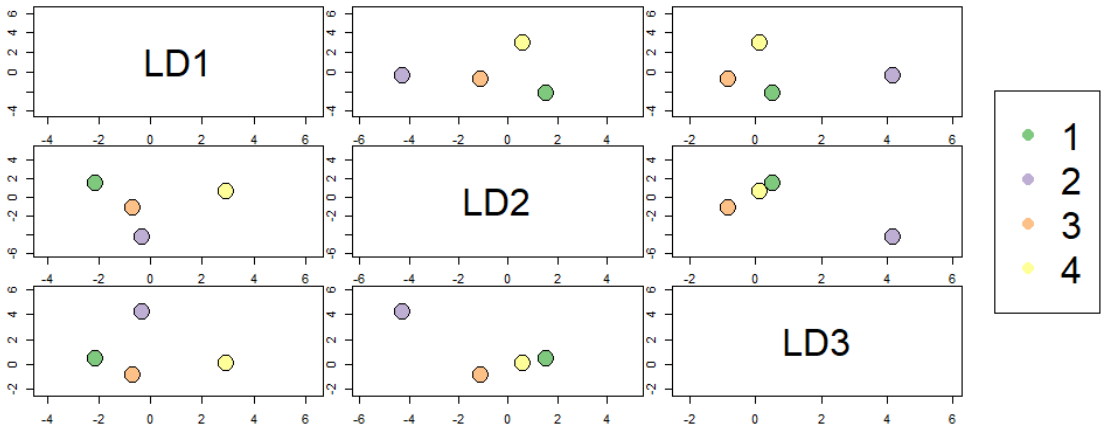
LDA в языке программирования R вычисляет принадлежность к классу по методу «all vs all» из машинного обучения. LDA модель высчитывает среднюю и вариацию для каждого класса и находит ковариацию для дискриминации каждого класса. Для предсказания

модель высчитывает для каждого класса вероятность неправильной классификации, используя теорему Байеса.

Гистограмма значений дискриминантной функции для каждого кластера из тренировочной выборки:



Для визуализации центроидов групп, был создан следующий график:



Таким образом, мы можем видеть местоположение центроида каждого кластера в пространстве трех дискриминантных функций. В целом, на большинстве графиков можно увидеть довольно репрезентативную визуализацию, что свидетельствует о высоком уровне классификации. Исключение составляет четвертый график: разбиение на 1 и 4 кластер не столь очевидны ввиду пересечения этих двух центроидов.

Попробуем предсказать значения выходов для тестовой выборки стран:

```
Classification table:
    obs
pred  1  2  3  4
   1  4  0  2  0
   2  0  1  0  0
   3  0  1 24  1
   4  0  0  0 11
Misclassification errors:
    1     2     3     4
 0.00 50.00  7.69  8.33
```

Из-за очень малого количества наблюдений во втором кластере, мы можем попытаться причислить лишь 2 наблюдения к тестовой выборке, из-за чего ошибка классификации моментально взлетает до критических значений при хотя бы одном неверно определённом наблюдении. В остальном, были получены достаточно хорошие значения, качественно относящие рассматриваемые наблюдения к соответствующим кластерам. Наибольшая доля ошибок концентрируется вокруг кластера 3, два наблюдения из которого были причислены к кластеру отстающих стран, при этом два наблюдения из других кластеров были неверно отнесены к кластеру 3.

Для нашей функции была получена Лямбда Уилкса, равная 0.012 при низком p-value, что говорит о высоком качестве представленной модели. По итогам проведённого анализа мы можем сделать вывод о том, что дискриминантный анализ позволяет достаточно точно прогнозировать значения внутри классов. Таким образом, при возможном добавлении в анализ новых стран, мы можем с высокой уверенностью утверждать о том, что они относились бы к тому или иному кластеру, характеризующемуся определёнными свойствами (например, с высоким уровнем безработицы или с низкой вовлеченностью в профессиональное образование), благодаря чему можно было бы значительно усовершенствовать экономическую политику развития страны, заранее зная их проблемные места.

Итоги проведенного исследования

В результате выполнения компьютерной работы наша команда не только смогла освоить навыки практического применения изученных по ходу курса многомерных статистических методов, но и лучше понять столь важный макроэкономический показатель, как разница между уровнем ВВП на душу населения в различных странах мира.

Ключевыми выводами, которые можно сделать по итогам всей работы, следует считать тот факт, что около 11 стран мира явно выделяются в лучшую сторону по уровню благосостояния граждан, тогда как из оставшихся 131 стран несколько десятков выделяются в два кластера наиболее преуспевающих среди оставшихся в рассмотрении государств. В свою очередь, остальные страны мира, преимущественно расположенные в Африке, Южной Америке и Азии, представляют собой весьма однородную по своей сути группу, среди которой сложно найти такую объясняющую переменную, которая обуславливала бы внутрикластерные различия.

Также весьма существенным выводом следует считать возможность предсказать, к какому именно кластеру стран относится государство, зная хотя бы часть всей генеральной совокупности стран. Столь важный метод исследования паттернов стран стал возможен благодаря применению дискриминантного анализа.

Наиболее значимым результатом нашей работы следует считать возможность предсказания благосостояния жителей конкретной страны за счёт применения методов регрессионного анализа, где в качестве объясняющих переменных рассматривается уровень развития институтов образования, медицины, труда и урбанизма.