Национальный исследовательский университет «Высшая школа экономики»

Факультет: Московский институт электроники и математики им.Тихонова

Образовательная программа: Прикладная математика

Отчет по модульной работе №1 по майнору «Прикладной статистический анализ»

Работу выполнила

студентка 3 курса

Беломытцева Алена

Владимировна

Москва, 2023г.

Введение

Для анализа в этой модульной работе был выбран датасет от Всемирной организации здравоохранения (WHO) об ожидаемой продолжительности жизни в различных странах. В нем рассматриваются такие переменные как:

- Country - страна
- Life expectancy - Ожидаемая продолжительность жизни в годах
- Adult Mortality - Показатели смертности взрослого населения обоих полов (вероятность смерти в возрасте от 15 до 60 лет на 1000 человек населения)
- infant deaths - Число младенческих смертей на 1000 человек населения
- Alcohol - Потребление алкоголя, зарегистрированное на душу населения (15+) (в литрах чистого алкоголя)
- percentage expenditure - Расходы на здравоохранение в процентах от валового внутреннего продукта на душу населения (%)
- Hepatitis B - Охват иммунизацией против гепатита B (HepB) среди детей в возрасте 1 года (%)
- Measles - число зарегистрированных случаев на 1000 человек населения
- BMI - Средний индекс массы тела всего населения
- under-five deaths - Число смертей в возрасте до пяти лет на 1000 человек населения
- Polio - Охват иммунизацией от полиомиелита (Pol3) среди детей в возрасте 1 года (%)
- Total expenditure - Общие государственные расходы на здравоохранение в процентах от общих государственных расходов (%)
- Diphtheria - Охват иммунизацией против дифтерии, столбнячного анатоксина и коклюша (DTP3) среди детей в возрасте до 1 года (%)
- HIV/AIDS - Смертность на 1 000 живорождений ВИЧ/СПИД (0-4 года)
- GDP - Валовой внутренний продукт на душу населения (в долларах США)
- Population - Население страны
- thinness 1-19 years - Распространенность худобы среди детей и подростков в возрасте от 10 до 19 лет (%)
- thinness 5-9 years - Распространенность худобы среди детей в возрасте от 5 до 9 лет (%)
- Income composition of resources - Индекс развития человеческого потенциала с точки зрения структуры доходов и ресурсов (индекс в диапазоне от 0 до 1)
- Schooling - Количество лет обучения в школе (лет)

Также в датасете данные разбиты на несколько лет. Для того чтоб выборка стала пространственной, для анализа было решено взять данные на 2015 год по всем странам.
Задачей этой модульной работы является изучение влияния выбранных показателей на продолжительность жизни с помощью построения модели линейной регрессии.
Модель линейной регрессии помогает определить, как изменение значений независимых переменных влияет на изменение зависимой переменной. Эта информация позволяет прогнозировать значения зависимой переменной на основе известных значений независимых переменных.
Гипотезой исследования является то, что на продолжительность жизни в стране положительно влияют расходы государства на здравоохранение, охват иммунизации от гепатита, полиомиелита и дифтерии, а также индекс развития человеческого потенциала и средняя продолжительность обучения. Отрицательно влияют все показатели смертности: от СПИДА и вич, взрослого человека и детей, распространенность худобы, популяция. Незначительно влияют на продолжительность жизни средний индекс массы тела.

Предварительная обработка данных

Уже было сказано, что данные содержат в себе информацию сразу по 15 годам, из которых для анализа были выбраны данные за 2015 год. После удаления других годов, были удалены переменные, в которых отсутствовала большая часть данных - столбцы Alcohol и Total expenditure. Далее были удалены страны, в которых в некоторых ячейках данные отсутствовали.

```
df_n.info()
```

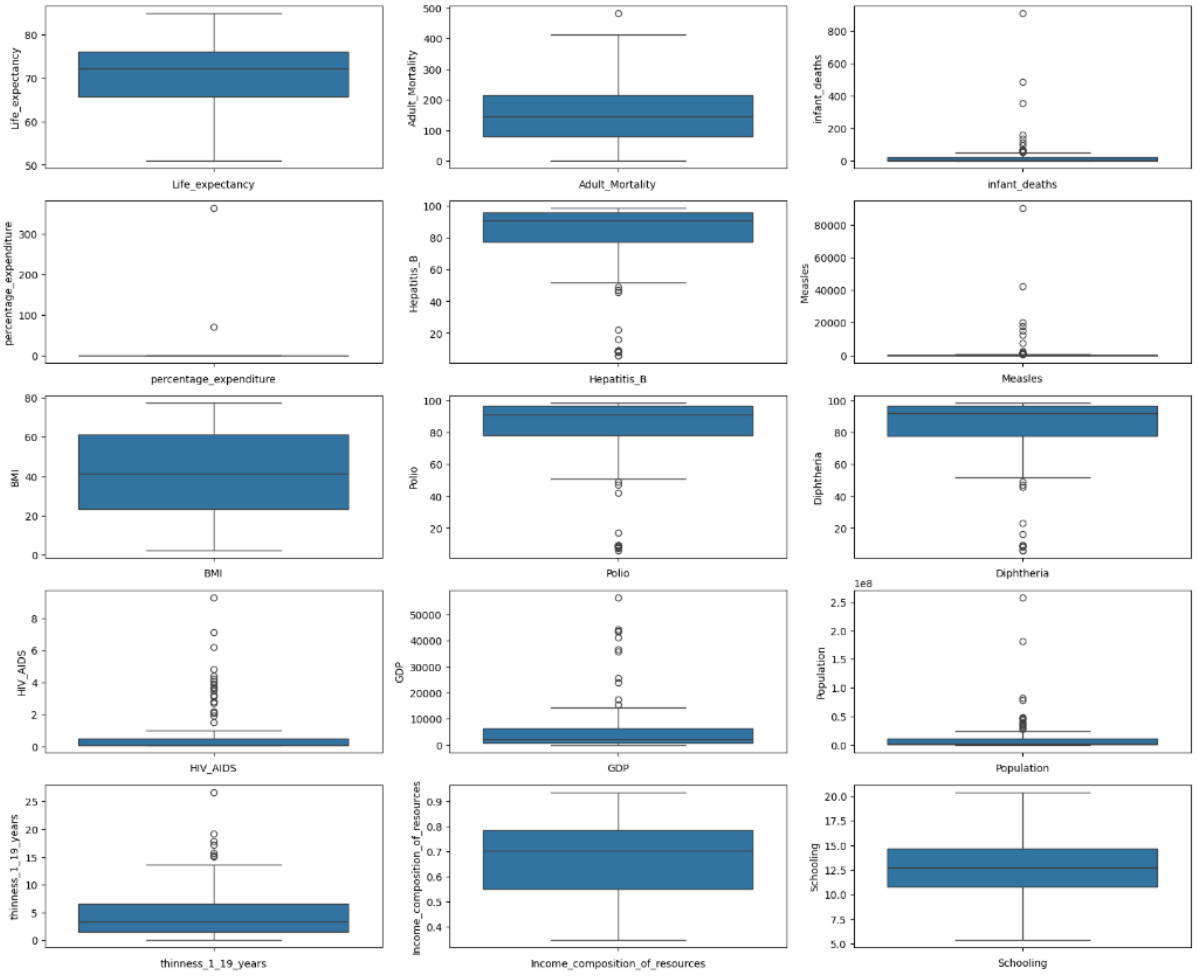
```
<class 'pandas.core.frame.DataFrame'>
Index: 183 entries, 0 to 2922
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Country                          183 non-null    object 
 1   Year                             183 non-null    int64  
 2   Status                           183 non-null    object 
 3   Life_expectancy                  183 non-null    float64
 4   Adult_Mortality                  183 non-null    float64
 5   infant_deaths                    183 non-null    int64  
 6   Alcohol                          6 non-null      float64
 7   percentage_expenditure           183 non-null    float64
 8   Hepatitis_B                      174 non-null    float64
 9   Measles                          183 non-null    int64  
 10  BMI                              181 non-null    float64
 11  under_five_deaths                183 non-null    int64  
 12  Polio                            183 non-null    float64
 13  Total_expenditure                2 non-null      float64
 14  Diphtheria                       183 non-null    float64
 15  HIV_AIDS                         183 non-null    float64
 16  GDP                              154 non-null    float64
 17  Population                       142 non-null    float64
 18  thinness_1_19_years              181 non-null    float64
 19  thinness_5_9_years               181 non-null    float64
 20  Income_composition_of_resources  173 non-null    float64
 21  Schooling                        173 non-null    float64
dtypes: float64(16), int64(4), object(2)
memory usage: 32.9+ KB
```

Также были удалена переменная thinnes 5 - 9, так как есть очень похожая на нее переменная, но для 10-19 лет. Скорее всего они бы сильно коррелировали и не несли бы новой информации в модель. По тем же соображениям была удалена переменная смертности детей до 5 лет, так как есть данные о младенческой смертности, одна включает в себя другую.

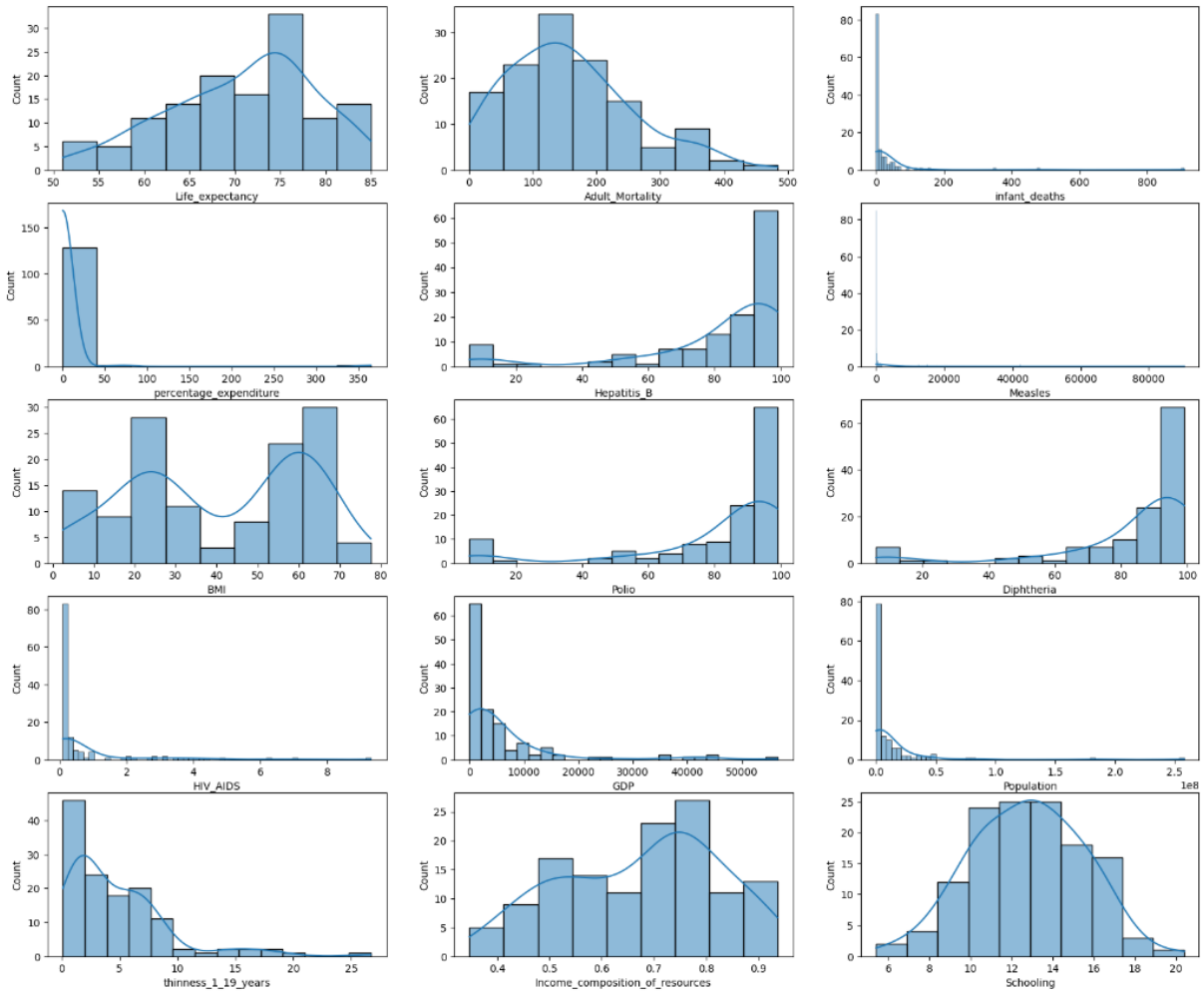
После удаления всех данных, которые не способствовали дальнейшему анализу получилось, что количество исследуемых переменных вместе с результирующей сократилось до 15, объем выборки равен 130.

```
Index: 130 entries, 0 to 2922
Data columns (total 15 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Life_expectancy                  130 non-null    float64
 1   Adult_Mortality                  130 non-null    float64
 2   infant_deaths                    130 non-null    int64
 3   percentage_expenditure           130 non-null    float64
 4   Hepatitis_B                      130 non-null    float64
 5   Measles                          130 non-null    int64
 6   BMI                              130 non-null    float64
 7   Polio                            130 non-null    float64
 8   Diphtheria                       130 non-null    float64
 9   HIV_AIDS                         130 non-null    float64
 10  GDP                              130 non-null    float64
 11  Population                       130 non-null    float64
 12  thinness_1_19_years              130 non-null    float64
 13  Income_composition_of_resources  130 non-null    float64
 14  Schooling                        130 non-null    float64
dtypes: float64(13), int64(2)
```

Рассмотрим каждую переменную подробнее.
Для анализа выбросов были построены графики boxplot для каждой переменной.



Можно заметить, что в каждой переменной кроме Life expectancy, BMI, Income composition of resources, schooling имеют выбросы. Но удалять выбросы нецелесообразно, потому что тогда не останется объектов, на которых проводился бы дальнейший анализ. Так что несмотря на то, что выбросы могут негативно влиять на будущую регрессионную модель, значительно увеличивая квадраты ошибок и некоторые другие проблемы. Далее посмотрим распределение величин - по хорошему они должны иметь нормальное распределение.



На графиках видно, что нормальному распределению приблизительно соответствует только продолжительность обучения, но проверим это еще с помощью тестов на нормальность.
H0: распределение нормально
H1: распределение ненормально

```
Life_expectancy         NormaltestResult(statistic=5.845897801057178, pvalue=0.05377487649508612)
Adult_Mortality         NormaltestResult(statistic=11.05804982647664, pvalue=0.003969858158581466)
infant_deaths         NormaltestResult(statistic=215.84901646127685, pvalue=1.3458033756873541e-47)
percentage_expenditure         NormaltestResult(statistic=274.48471201369046, pvalue=2.4911627585819876e-60)
Hepatitis_B         NormaltestResult(statistic=56.77347990119794, pvalue=4.696728762088265e-13)
Measles         NormaltestResult(statistic=231.42307403782104, pvalue=5.586219289788382e-51)
BMI         NormaltestResult(statistic=122.08022344706481, pvalue=3.094683958359167e-27)
Polio         NormaltestResult(statistic=58.031581265429764, pvalue=2.503815018749549e-13)
Diphtheria         NormaltestResult(statistic=68.48376244155489, pvalue=1.3456742305867998e-15)
HIV_AIDS         NormaltestResult(statistic=104.83420574535026, pvalue=1.7200515583963406e-23)
GDP         NormaltestResult(statistic=103.75395259121105, pvalue=2.951993908611266e-23)
Population         NormaltestResult(statistic=189.71302898165175, pvalue=6.372930525704794e-42)
thinness_1_19_years         NormaltestResult(statistic=68.817939617226, pvalue=1.138608139610463e-15)
Income_composition_of_resources         NormaltestResult(statistic=16.35936887124493, pvalue=0.0002802903758323781)
Schooling         NormaltestResult(statistic=0.1816686649257625, pvalue=0.9131689808246882)
```

Итак, по этому тесту гипотеза на нормальное распределение на уровне 0,05 не отвергается в случаях в life expectancy, schooling.
Тест Шапиро-Уилка:
H0: распределение нормально
H1: распределение ненормально

```
Life_expectancy         ShapiroResult(statistic=0.9672056436538696, pvalue=0.0030873878858983517)
Adult_Mortality         ShapiroResult(statistic=0.9556861519813538, pvalue=0.0003180943022016436)
infant_deaths         ShapiroResult(statistic=0.277055025100708, pvalue=1.540323366429561e-22)
percentage_expenditure         ShapiroResult(statistic=0.07973605394363403, pvalue=6.508366243438531e-25)
Hepatitis_B         ShapiroResult(statistic=0.6969553232192993, pvalue=4.9115806839933156e-15)
Measles         ShapiroResult(statistic=0.1908230185508728, pvalue=1.2440061951740339e-23)
BMI         ShapiroResult(statistic=0.9107385277748108, pvalue=2.998459649461438e-07)
Polio         ShapiroResult(statistic=0.6790836453437805, pvalue=1.7668189307768118e-15)
Diphtheria         ShapiroResult(statistic=0.670691967010498, pvalue=1.108765069164695e-15)
HIV_AIDS         ShapiroResult(statistic=0.5228351354598999, pvalue=9.507087667019763e-19)
GDP         ShapiroResult(statistic=0.5784101486206055, pvalue=1.084433038594468e-17)
Population         ShapiroResult(statistic=0.3858938217163086, pvalue=5.259743843295283e-21)
thinness_1_19_years         ShapiroResult(statistic=0.8044607639312744, pvalue=7.104763825871441e-12)
Income_composition_of_resources         ShapiroResult(statistic=0.9647559523582458, pvalue=0.0018631733255460858)
Schooling         ShapiroResult(statistic=0.9967318177223206, pvalue=0.9928616285324097)
```

По этому тесту гипотеза на нормальное распределение на уровне 0,05 не отвергается только для переменной schooling.
Так как в двух случаях предположение о нормальности не имеет оснований для отвержения говорим о нормальности распределения переменной schooling.

Как переменные могут влиять на результирующую переменную life expectancy.

Adult Mortality - показатели смертности людей среднего возраста - если много людей умирают не в старом возрасте, тем теоретически меньше средняя продолжительность жизни.

Infant deaths - число младенческих смертей на 1000 человек населения - большая младенческая смертность может говорить о недостатке медицины или о плохих условиях жизни населения, что может сказываться на среднюю продолжительность жизни. Да и напрямую тоже влияет.

Percentage expenditure - расходы на здравоохранение в процентах от валового внутреннего продукта на душу населения. Чем лучше развита

медицина в стране, тем своевременнее и доступнее помощь подобного рода для любого человека в стране. А это значит что продолжительность жизни населения увеличится.

Hepatitis B, Polio, Diphtheria - охват иммунизацией против гепатита B (HepB), полиомиелита, дифтерии, столбнячного анатоксина и коклюша среди детей в возрасте 1 года (%) - способствует иммунитету против болезни, которая тяжело протекает, от которой большая смертность. Чем больше людей защищены от нее на начальном этапе жизни, тем меньше заражаемость и способность противостоять опасной болезни, а значит и продолжительность жизни должна увеличиваться.

Measles - число зарегистрированных случаев на 1000 человек населения. Корь опасное заболевание с возможным летальным исходе, или же с серьезными осложнениями. А значит на здоровье и продолжительность жизни влияет достаточно сильно.

BMI - средний индекс массы тела всего населения - массовое ожирение или анорексия может говорить о проблемах образа жизни населения и соответственно о проблемах со здоровьем, а следственно и сокращении продолжительности жизни.

HIV/AIDS - смертность на 1 000 живорождений ВИЧ/СПИД говорит о распространенности опасного заболевания, от которого легко получить летальный исход, если не развита медицина или не иметь доступа к соответствующим медикаментам.

GDP - валовой внутренний продукт на душу населения (в долларах США) - показатель уровня экономической активности и качества жизни населения, большой показатель говорит о том, что жители имеют в среднем нормальный уровень качества жизни и соответственно доступ к необходимой медпомощи, а значит и должны быть здоровее.

Population - население страны наиболее вероятно не будет влиять на продолжительность жизни, но, например, в бедных странах, где живут много бедных людей могут быть распространены инфекции и другие заболевания.

thinness 1-19 years - распространенность худобы среди детей и подростков в возрасте от 10 до 19 лет (%) худоба - свидетельствует о недостатке правильного питания или о заболеваниях, а значит может сокращать продолжительность жизни человека.

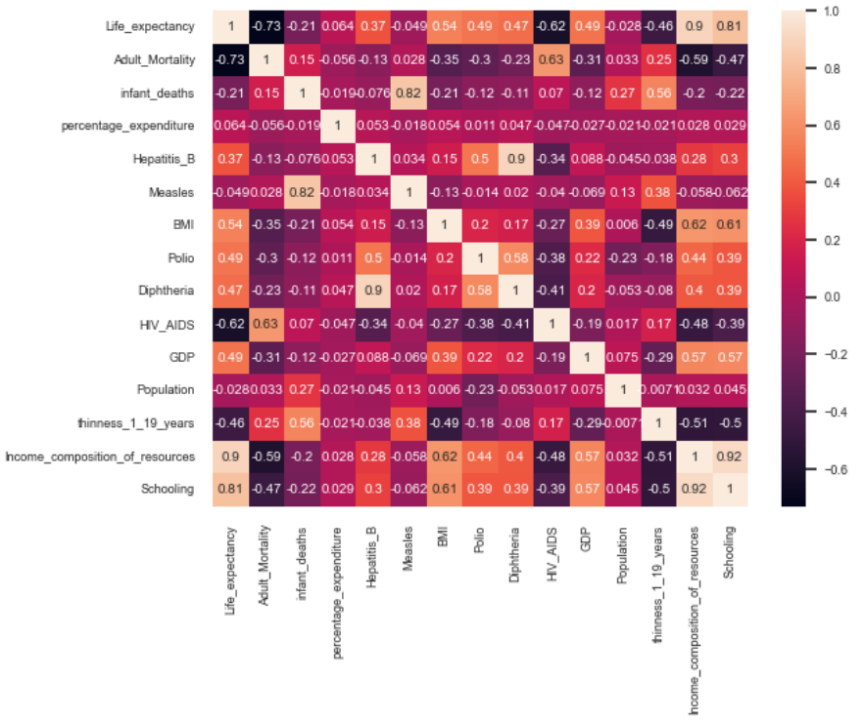
Income composition of resources - Индекс развития человеческого потенциала с точки зрения структуры доходов и ресурсов (индекс в диапазоне от 0 до 1) чем лучше развита страна и возможности для человека, тем он здоровее и тем дольше продолжительность его жизни.

Schooling - Количество лет обучения в школе (лет) - интересно посмотреть реальное влияние на результирующую переменную - учеба может вызывать хронический стресс, а люди подверженные постоянному переживанию живут меньше согласно исследованиям. С другой стороны после обучения люди становятся более ответственными и способными позаботиться о себе как следует.

Корреляционный анализ

Перед тем как строить линейную регрессию необходимо проверить корреляцию, чтоб понять как связаны объясняющие переменные с результирующей переменной. Если взаимосвязь между регрессорами и результатом будет мала, то большая вероятность что в модели линейной регрессии она будет незначима. Также для хорошей модели линейной регрессии необходима маленькая корреляция между объясняющими переменными, иначе будет мультиколлинеарность приводит к неустойчивым коэффициентам.

Итак, для выбранных переменных построим матрицу парных корреляций Пирсона.



Как можно заметить, что результирующая переменная Life expectancy имеет коэффициент корреляции больше 0,45 с переменными

- Adult_Mortality
- Hepatitis_B
- BMI
- Polio
- Diphtheria
- HIV_AIDS
- GDP
- thinness_1_19_years
- Income_composition_of_resources
- Schooling

что говорит о средней и сильной взаимосвязи. Другие переменные не будут оказывать на результирующую переменную сильного влияния. Также необходимо отметить сильную корреляцию между некоторыми переменными, например, Schooling и Income_composition_of_resources или Diphtheria и Hepatitis_B. Между этими переменными сильная взаимосвязь, что может свидетельствовать о сильной мультиколлинеарности в данных.

Мультиколлинеарность

Было выдвинуто предположение о сильной мультиколлинеарности в данных. Это можно проверить также с помощью коэффициента инфляции дисперсии VIF.

	Variable	VIF
0	Adult_Mortality	6.523864
1	infant_deaths	5.474308
2	percentage_expenditure	1.023457
3	Hepatitis_B	64.954329
4	Measles	3.789818
5	BMI	8.633566
6	Polio	20.176092
7	Diphtheria	89.134117
8	HIV_AIDS	2.405855
9	GDP	1.951469
10	Population	1.484530
11	thinness_1_19_years	3.863694
12	Income_composition_of_resources	176.967996
13	Schooling	172.134631

Предположения о сильной мультиколлинеарности данных подтвердились. Для отсутствия мультиколлинеарности коэффициент VIF должен быть меньше 4, или в крайних случаях допускается повышение его до 8. В случае исследуемых данных, он повышается до 172, что говорит о мультиколлинеарности.

Линейная модель регрессии

Построим линейную модель регрессии со всеми выбранными для анализа переменными и постепенно будем удалять незначимые переменные и пытаться избавиться от мультиколлинеарности.

OLS Regression Results

Dep. Variable:	Life_expectancy	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	71.71
Date:	Sun, 19 Nov 2023	Prob (F-statistic):	5.75e-50
Time:	00:38:34	Log-Likelihood:	-306.36
No. Observations:	130	AIC:	642.7
Df Residuals:	115	BIC:	685.7
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	49.0369	2.136	22.954	0.000	44.805	53.268
Adult_Mortality	-0.0209	0.004	-5.792	0.000	-0.028	-0.014
percentage_expenditure	0.0052	0.007	0.701	0.485	-0.009	0.020
infant_deaths	0.0004	0.006	0.072	0.943	-0.011	0.011
Hepatitis_B	0.0436	0.023	1.918	0.058	-0.001	0.089
Measles	4.78e-06	5e-05	0.096	0.924	-9.44e-05	0.000
BMI	-0.0081	0.015	-0.529	0.598	-0.038	0.022
Polio	0.0094	0.013	0.740	0.461	-0.016	0.035
Diphtheria	-0.0111	0.026	-0.424	0.672	-0.063	0.041
HIV_AIDS	-0.5218	0.220	-2.377	0.019	-0.957	-0.087
GDP	6.363e-06	2.97e-05	0.214	0.831	-5.26e-05	6.53e-05
Population	-7.617e-09	9.04e-09	-0.843	0.401	-2.55e-08	1.03e-08
thinness_1_19_years	-0.0990	0.082	-1.205	0.230	-0.262	0.064
Income_composition_of_resources	34.8560	4.959	7.029	0.000	25.034	44.678
Schooling	-0.0513	0.240	-0.213	0.831	-0.527	0.425

Omnibus:	4.109	Durbin-Watson:	2.136
Prob(Omnibus):	0.128	Jarque-Bera (JB):	4.383
Skew:	-0.189	Prob(JB):	0.112
Kurtosis:	3.816	Cond. No.	6.82e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.82e+08. This might indicate that there are
strong multicollinearity or other numerical problems.

У модели хороший коэффициент детерминации, равный 0,897, но при этом можно заметить, что большинство коэффициентов незначимы или имеют очень маленький коэффициент. Сумма ошибок равна 851.6019772036441. Избавимся регрессоров, несущих малую информативность для результирующей переменной, а затем последовательно уберем все незначимые переменные.

OLS Regression Results

Dep. Variable:	Life_expectancy	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	333.4
Date:	Sun, 19 Nov 2023	Prob (F-statistic):	1.00e-59
Time:	01:09:14	Log-Likelihood:	-311.87
No. Observations:	130	AIC:	631.7
Df Residuals:	126	BIC:	643.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	47.1497	1.711	27.555	0.000	43.763	50.536
Adult_Mortality	-0.0255	0.003	-8.604	0.000	-0.031	-0.020
Hepatitis_B	0.0453	0.010	4.556	0.000	0.026	0.065
Income_composition_of_resources	35.5445	2.012	17.667	0.000	31.563	39.526

Omnibus:	7.039	Durbin-Watson:	2.089
Prob(Omnibus):	0.030	Jarque-Bera (JB):	7.007
Skew:	-0.434	Prob(JB):	0.0301
Kurtosis:	3.736	Cond. No.	2.13e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.13e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

В получившейся модели так и не удалось избавиться от мультиколлинеарности, что не очень хорошо. Коэффициент Durbin-Watson = 2.089 - для отсутствия автокорреляции должен быть от 2 до 4, значит значение лежит в пределах нормы.

Вид получившейся линейной регрессии:

$$\hat{y} = 47.1497 - 0.0255 * Adult\_Mortality + 0.0453 * Hepatitis\_B + 35.5445 * Income\_composition\_of\_resources$$

Коэффициент детерминации равен 0.888 что говорит о том, что модель объясняет 88,8% исходных данных - это хороший показатель. При этом по сравнению с изначальной моделью сумма ошибок оказалась равной 923.04. Не особо сильное повышение, так как было удалено больше половины изначальных переменных.

Далее построим модель, с прологарифмированной результирующей переменной:

OLS Regression Results

Dep. Variable:	np.log(Life_expectancy)	R-squared:	0.880
Model:	OLS	Adj. R-squared:	0.878
Method:	Least Squares	F-statistic:	309.5
Date:	Sun, 19 Nov 2023	Prob (F-statistic):	6.34e-58
Time:	01:21:19	Log-Likelihood:	232.61
No. Observations:	130	AIC:	-457.2
Df Residuals:	126	BIC:	-445.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.9198	0.026	150.993	0.000	3.868	3.971
Adult_Mortality	-0.0004	4.5e-05	-8.849	0.000	-0.000	-0.000
Hepatitis_B	0.0007	0.000	4.569	0.000	0.000	0.001
Income_composition_of_resources	0.5041	0.031	16.514	0.000	0.444	0.564

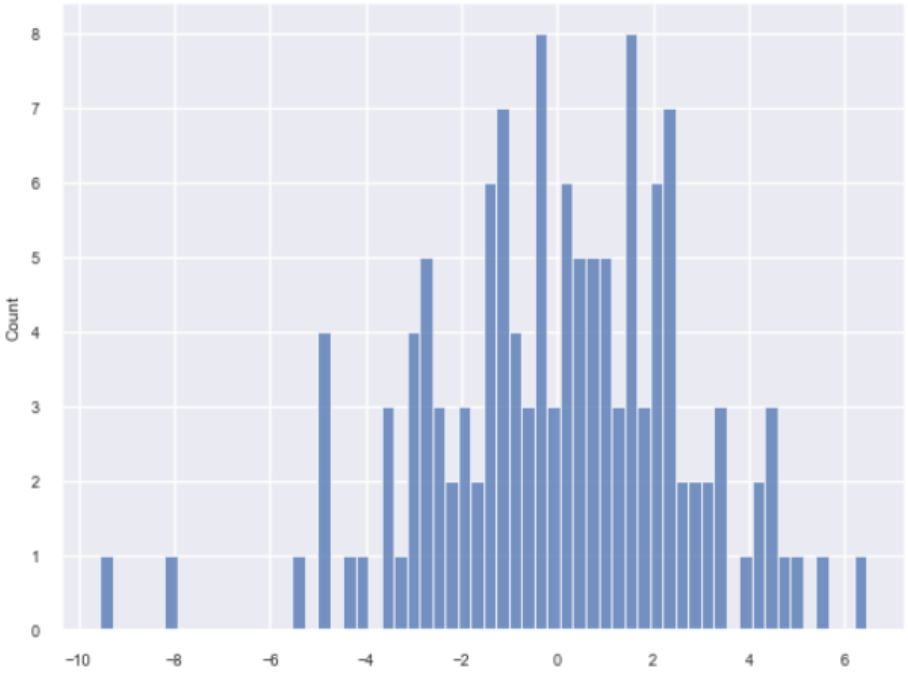
Omnibus:	14.401	Durbin-Watson:	2.071
Prob(Omnibus):	0.001	Jarque-Bera (JB):	19.182
Skew:	-0.618	Prob(JB):	6.84e-05
Kurtosis:	4.418	Cond. No.	2.13e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.13e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

R^2 снизился и две переменные стали незначимыми, логарифмировать смысла не имеет.

Анализ ошибок построенной модели

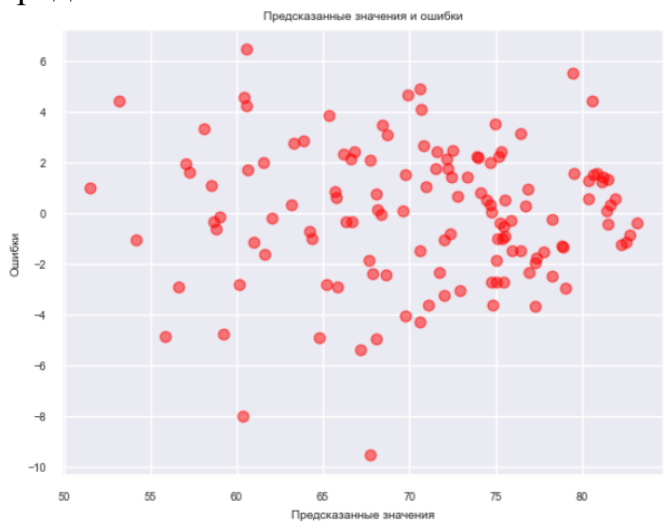
Анализ будет проводиться для модели без логарифмирования. Для начала посмотрим на распределение ошибок - оно должно быть нормальным.



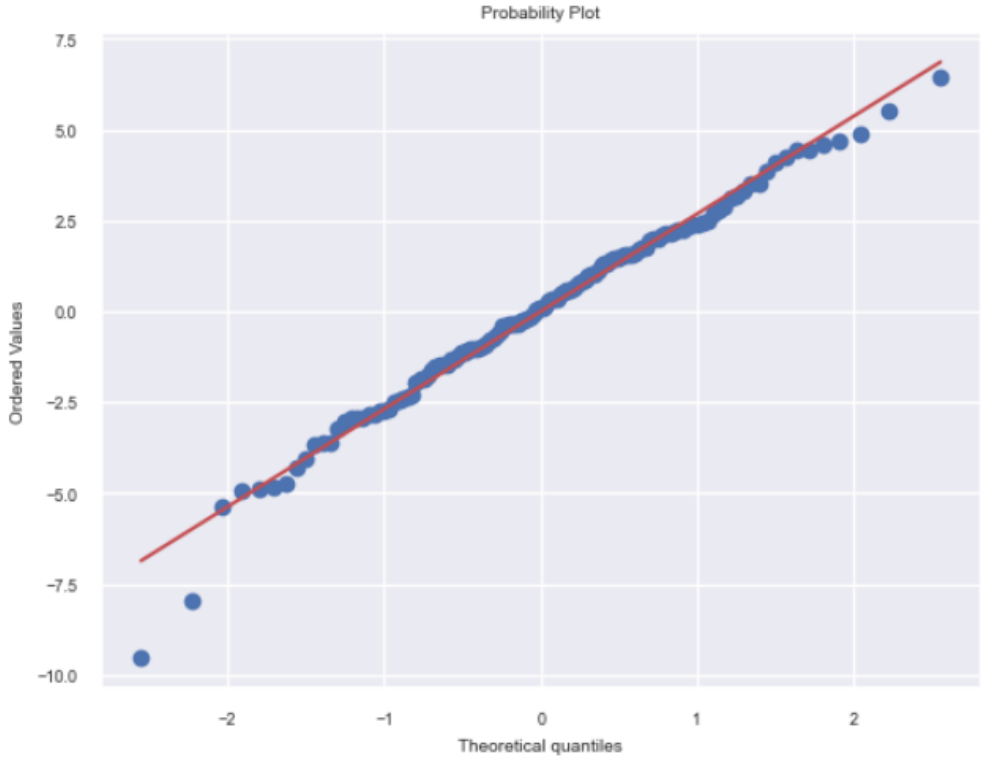
На нормальное распределение это не особо похоже, но возможно, если бы выборка была больше, что-то из этого и получилось бы.
Далее рассмотрим график предсказанных и реальных значений.

Точки образуют широкую прямую, а значит что в целом модель нормально предсказывает продолжительность жизни.

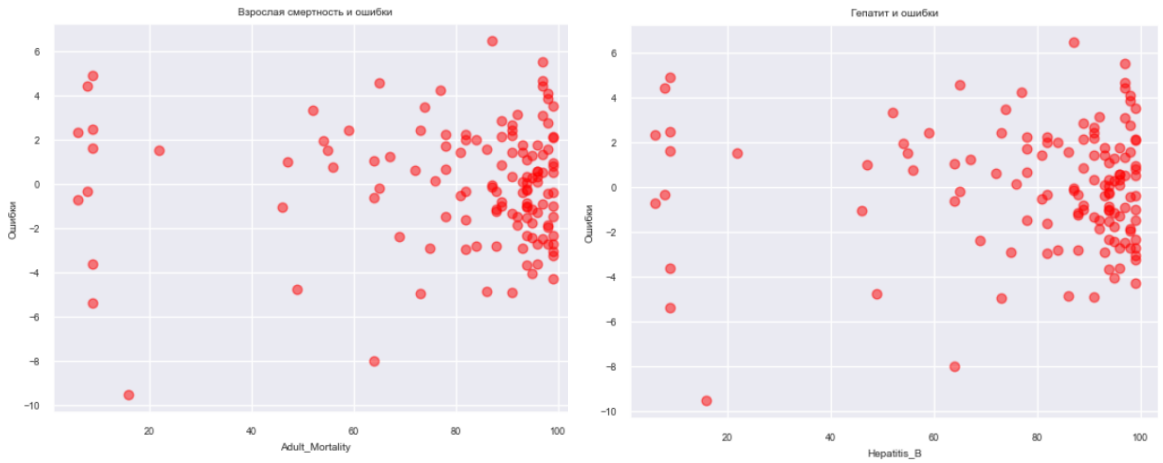
Далее рассмотрим облако рассеяния ошибок от предсказания продолжительности жизни.

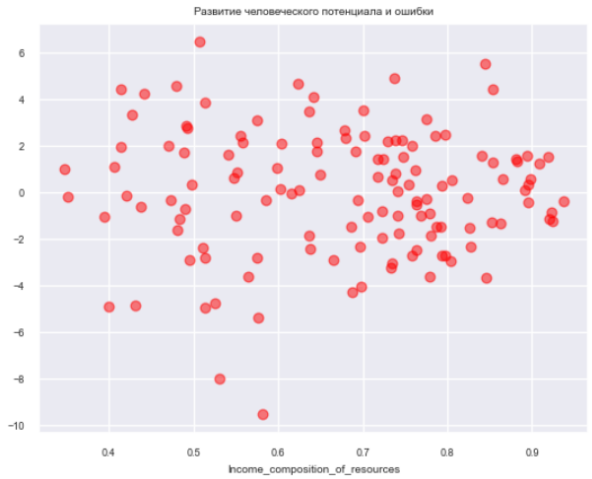


На графике нет подобия конуса, а значит предполагаем, что гетероскедастичность отсутствует.
Далее посмотрим, соответствует ли распределение ошибок нормальному:



На графике красной линией обозначено то, как выглядело бы нормальное распределение. Можно видеть, как синие кружочки - ошибки незначительно колеблются вокруг этой линии, за исключением некоторых выбросов на концах. Можно сделать вывод о нормальности распределения ошибок.
Затем, посмотрим какие переменные влияют на гетероскедастичность ошибок.





На графиках видно, что нет гетероскедастичности, так как ошибки образуют облако.

Тесты на гетероскедастичность.

Тест Бройша-Пагана:

H0: ошибки гомоскедастичны

H1: присутствует гетероскедастичность

Тест Бройша-Пагана на гетероскедастичность показал следующие результаты:

```
Статистика теста Лагранжа: 13.031033482345661
p-value: 0.004569972126598085
F-Statistic: 4.679047977874645
F-Test p-value: 0.003917492051701245
```

Так как p-value меньше уровня значимости 0,05, имеем основания отвергнуть нулевую гипотезу о гомоскедастичности данных. Значит гетероскедастичность все-таки присутствует, несмотря на красивые графики.

В подтверждение проведем тест Уайта на гетероскедастичность:

H0: ошибки гомоскедастичны

H1: присутствует гетероскедастичность

Тест Уайта на гетероскедастичность показал следующие результаты:

```
Статистика теста: 24.132188715365125
p-value: 0.004097143324297448
F-статистика: 3.0392856176064864
F-Test p-value: 0.0026150115440515863
```

P-value в этом тесте также меньше уровня значимости, а значит нулевая гипотеза об отсутствии гетероскедастичности отвергается.

Гетероскедастичность присутствует по результатам двух тестов. Необходимо сделать поправку Уайта.

OLS Regression Results

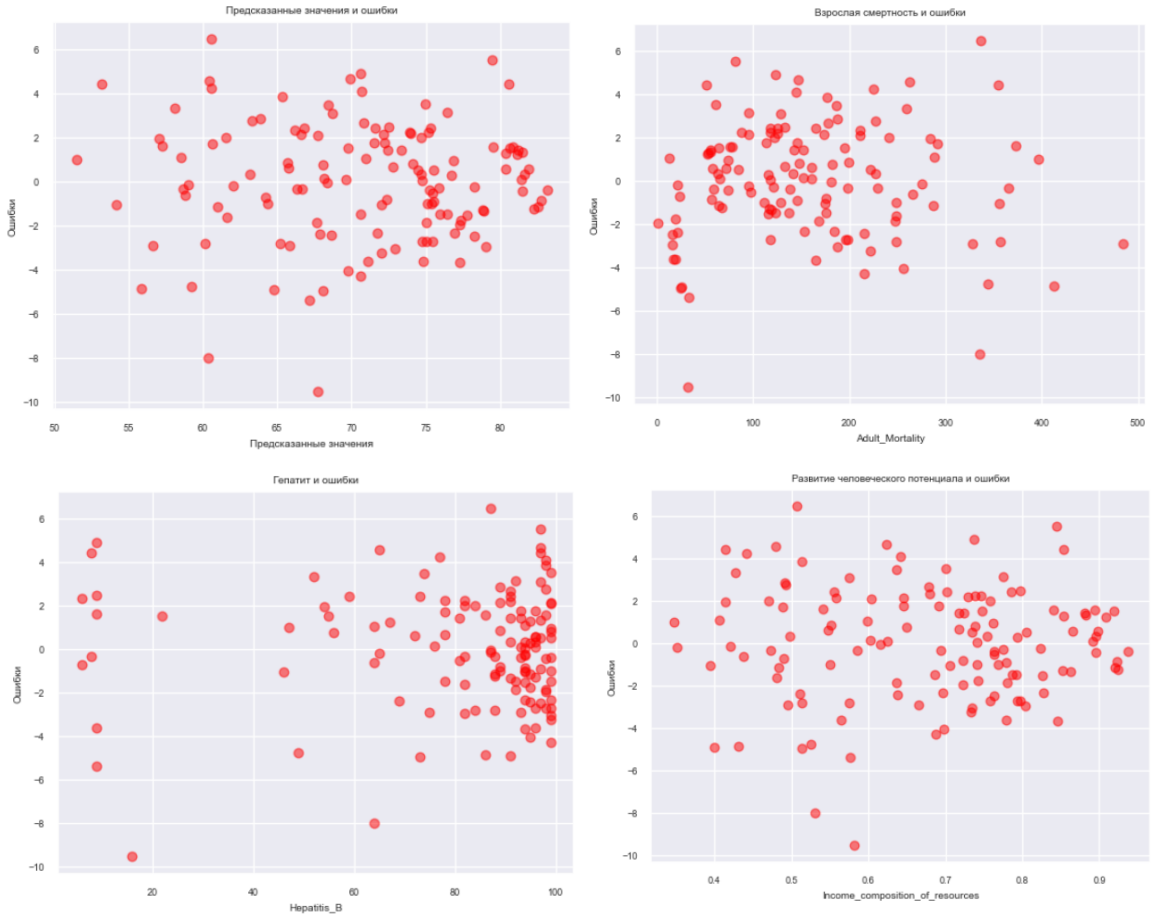
Dep. Variable:	Life_expectancy	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	343.2
Date:	Sun, 19 Nov 2023	Prob (F-statistic):	1.97e-60
Time:	02:19:07	Log-Likelihood:	-311.87
No. Observations:	130	AIC:	631.7
Df Residuals:	126	BIC:	643.2
Df Model:	3		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	47.1497	2.342	20.129	0.000	42.559	51.741
Adult_Mortality	-0.0255	0.004	-6.303	0.000	-0.033	-0.018
Hepatitis_B	0.0453	0.013	3.363	0.001	0.019	0.072
Income_composition_of_resources	35.5445	2.134	16.655	0.000	31.362	39.727

Omnibus:	7.039	Durbin-Watson:	2.089
Prob(Omnibus):	0.030	Jarque-Bera (JB):	7.007
Skew:	-0.434	Prob(JB):	0.0301
Kurtosis:	3.736	Cond. No.	2.13e+03

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 2.13e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Вывод построения линейной регрессии после применения поправки Уайта.
Для анализа ошибок после поправки Уайта построим те же графики и сравним их.



Видно, что ошибки сместились и не образуют два разрозненных облака для взрослой смертности, но график для гепатита и ошибок выглядит примерно также как и для модели без поправки.

Модель линейной регрессией с регуляризацией и кросс валидацией

Для этого сначала была произведена нормировка данных и построена модель регрессии с преобразованными данными. Затем была построена модель с помощью регуляризации Ридж и кросс валидацией.
Полученные коэффициенты:

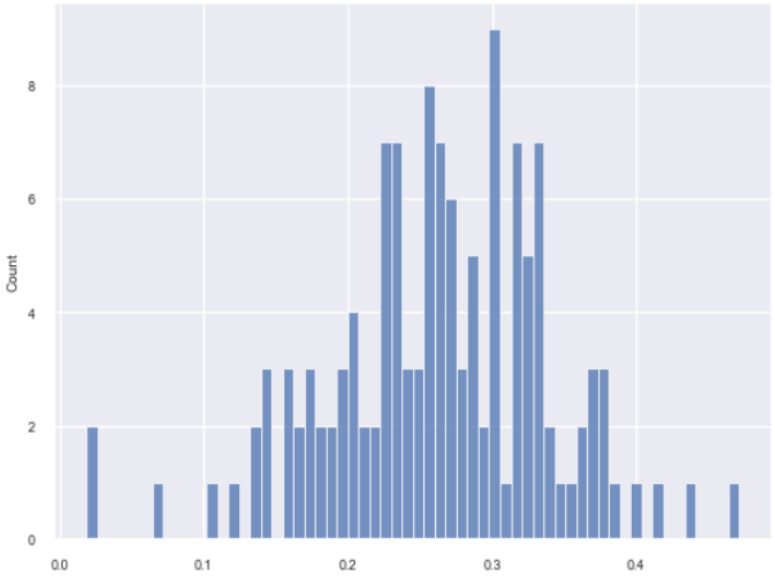
	0	factors
0	-0.295031	Adult_Mortality
1	0.003622	infant_deaths
2	0.038054	percentage_expenditure
3	0.073438	Hepatitis_B
4	0.014658	Measles
5	0.006319	BMI
6	0.042902	Polio
7	0.018127	Diphtheria
8	-0.146671	HIV_AIDS
9	0.022158	GDP
10	-0.036631	Population
11	-0.076796	thinness_1_19_years
12	0.450874	Income_composition_of_resources
13	0.115224	Schooling

А коэффициент альфа равен 0.5. При этом коэффициент детерминации равен 0.8929147164732575.

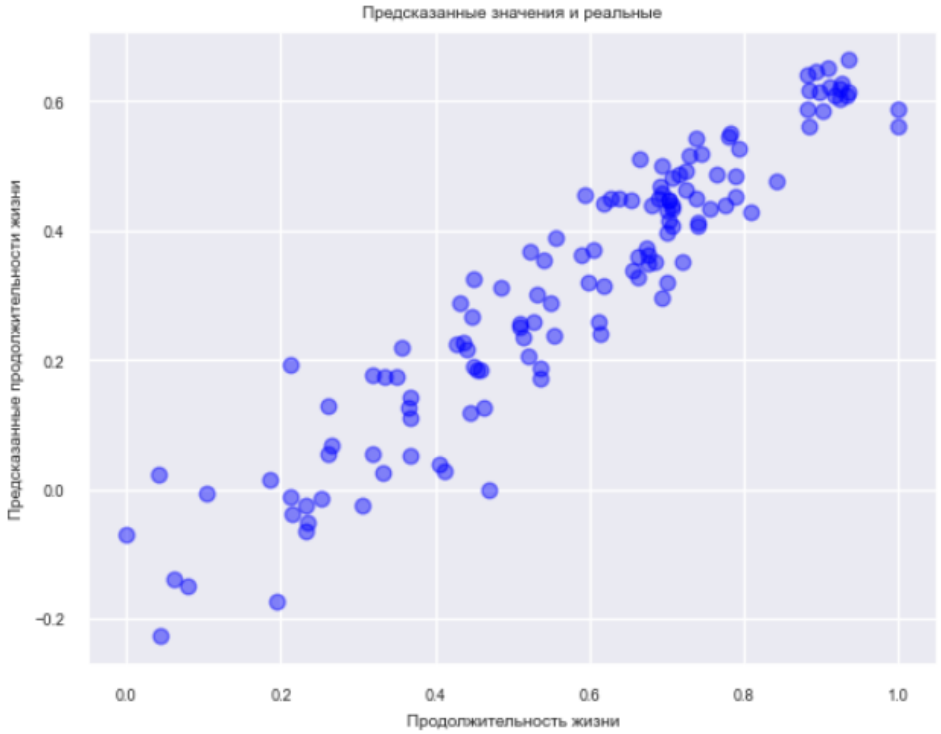
Далее проведена регуляризация Лассо с кросс валидацией и были получены следующие коэффициенты:

	0	factors
0	-0.282242	Adult_Mortality
1	-0.000000	infant_deaths
2	0.077409	Hepatitis_B
3	0.000000	BMI
4	0.028010	Polio
5	0.000000	Diphtheria
6	-0.089683	HIV_AIDS
7	0.000000	GDP
8	-0.000000	Population
9	-0.000000	thinness_1_19_years
10	0.598531	Income_composition_of_resources
11	0.000000	Schooling

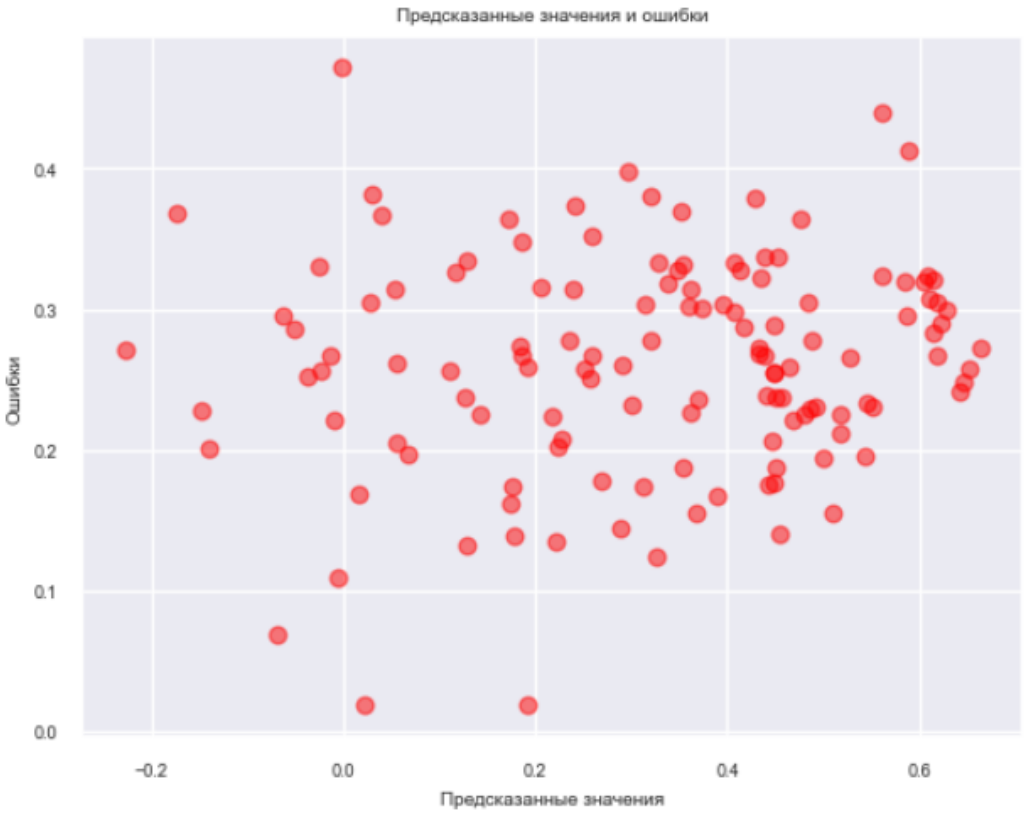
Для такой линейной модели коэффициент детерминации равен 0.8914420030245664. Это меньше, чем для регуляризации Ридж, но стоит обратить внимание на то, что в данных присутствует мультиколлинеарность и разреживание переменных, которое происходит с помощью Лассо, в какой-то степени помогает избавиться от этой помехи. К тому же, эта модель оставила больше переменных, в сравнении с обычной моделью линейной регрессии, что позволит сделать больше выводов о влиянии исходных переменных на результирующую. Проведем анализ ошибок модели с регуляризацией Лассо.



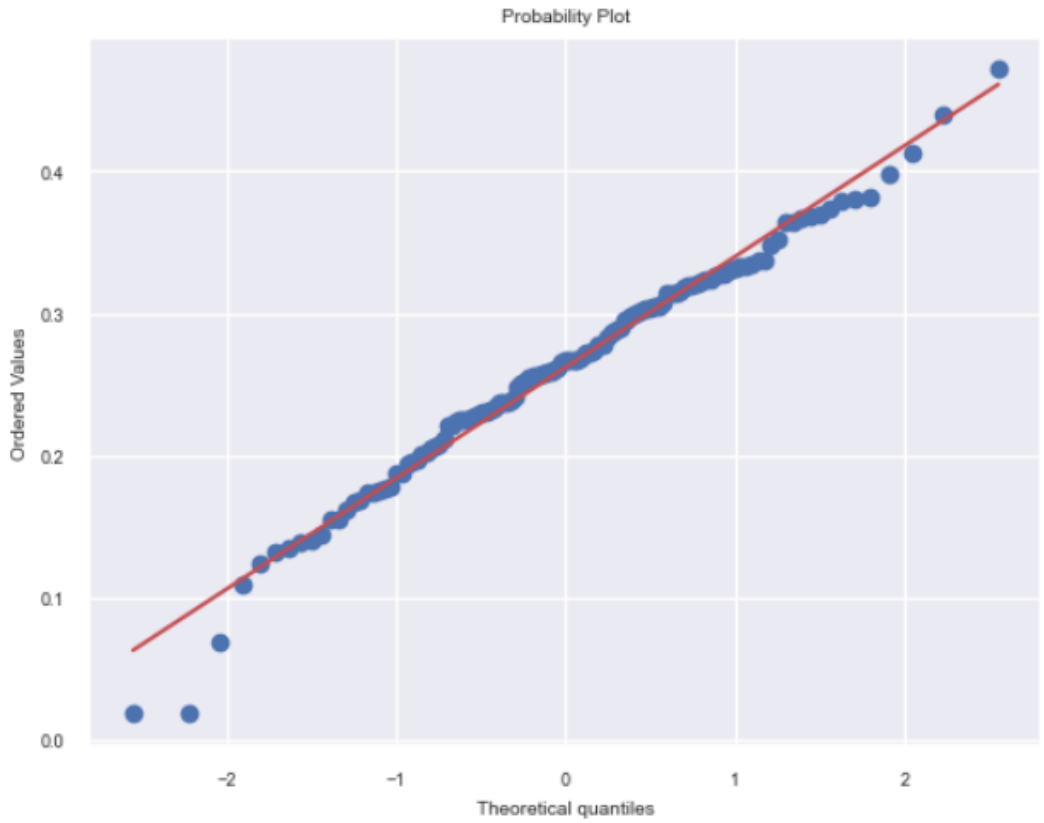
Распределение ошибок похоже на распределение, построенное для модели линейной регрессии, построенной вручную.



Точки образуют широкую прямую, а значит что в целом модель нормально предсказывает продолжительность жизни.



Облако рассеяния ошибок не образует конусовидную форму, а значит, что гетероскедастичности не должно быть(но лучше бы провести тесты).



А также можно сказать, что ошибки также имеют нормальное распределение с некоторыми погрешностями да концах.

Интерпретация и выводы

Лучшей моделью линейной регрессией вышла модель с регуляризацией Лассо, так как получилась зависимость от пяти переменных с коэффициентами, которые не стремятся к нулю кроме одной. Она имеет достаточно большой коэффициент детерминации , а значит имеет способность предсказывать результат. И этой модели хочется доверять, так как она по возможности убрала мультиколлинеарность данных.

Итак, уравнение линейной регрессии выглядит:

$$\widehat{y} = -0.282242 * Adult\_Mortality + 0.077409 * Hepatitis\_B + 0.028010 * Polio - 0.089683 * HIV\_AIDS + 0.598531 * Income\_composition\_of\_resources$$

Отрицательный коэффициент перед взрослой смертностью - чем больше человек умирает в возрасте до 60 лет, тем меньше продолжительность жизни в годах.

Чем больше иммунизаций от Гепатита-Б и полиомиелита было произведено, тем больше продолжительность жизни. Прививание способствует продолжительности жизни.

Смертность от СПИД и вич отрицательно влияет на продолжительность жизни человека.

Индекс развития потенциала человека в большей мере положительно влияет на результирующую переменную, что говорит о том, что чем лучше развита страна и чем лучше уровень жизни в ней, тем человек дольше живет. Логично, так как при хорошем уровне жизни доступна своевременная медицинская помощь и другие преимущества.

Таким образом была построена модель линейной регрессии, которая могла бы предсказывать продолжительность жизни в конкретной стране на 2015 год.
Предположения о влиянии переменных, высказанные в начале подтвердились:

- смертность людей среднего возраста отрицательно влияет на продолжительность жизни
- общая иммунизация способствует увеличению продолжительности жизни.
- смертность от спида негативно сказывается на средней продолжительности жизни в стране
- Индекс развития человеческого потенциала положительно влияет на среднюю продолжительность жизни в стране.

Другие переменные не вошли в модель, но хотелось бы сказать что индекс развития человеческого потенциала теоретически включает себя данные и о других переменных, так, например, в странах с большим индексом среднее количество лет обучения будет больше, чем в странах с маленьким. Так же при подсчете этого индекса используются данные о валовой внутренний продукт на душу населения.

Интересно, что число заболеваемостью корью не влияет на продолжительность жизни.
Население страны, как и предполагалось, не оказывает влияния на среднюю продолжительность жизни в стране.

Число младенческих жизней и распространенность худобы также не влияет на продолжительность жизни.