全球人口期望寿命

期末报告

刘睿鹏 李媛媛 唐慕尧

1 数据背景

世界、人口、健康。Covid-19 的爆发让健康问题再一次展现在人们的面前，经济快速发展的同时全球变暖，能源枯竭，病毒肆虐等等问题同样也在影响着人类的身体健康。Our World in Data（https://ourworldindata.org）是一个收集全世界各地区人口、经济、城市发展、健康等等数据的网站，借助可视化与数学模型开展对于人类期望寿命的改变的全球的不均衡性原因的分析。在 Our World in Data 之上有着详尽的各类数据的可视化报告，却缺少各个因素之间相互影响的相关研究，因此我选取每一个板块中最具有代表性的数据，开展下列关于人口期望寿命的相关研究。

1.1 数据收集与清洗

1.1.1 数据选择

人类期望寿命的影响因素颇多，其涉猎众多领域，健康，食物供应，收入增长和分配，暴力，战争，文化，能源使用，教育和环境变化的趋势等等。经过对于数据的完整度的考量，我们在每个模块中进行遴选，选择在 1990-2020 年之间全世界数据最完整，代表性较强数据加入到整个模型中进行后续的分析，相关变量如下：

- 出生率、死亡率、人口迁移率、自然增长率
- 总人口、人口密度、城市人口率、城市宜居度（离散变量）
- 赡养比，人口老龄化程度（离散变量）
- 空气污染死亡率，每人能源消耗，交通覆盖率
- 税收比重，平均受教育年限, 国家所处地域地形（离散变量）

1.1.2 聚合数据与具体数据

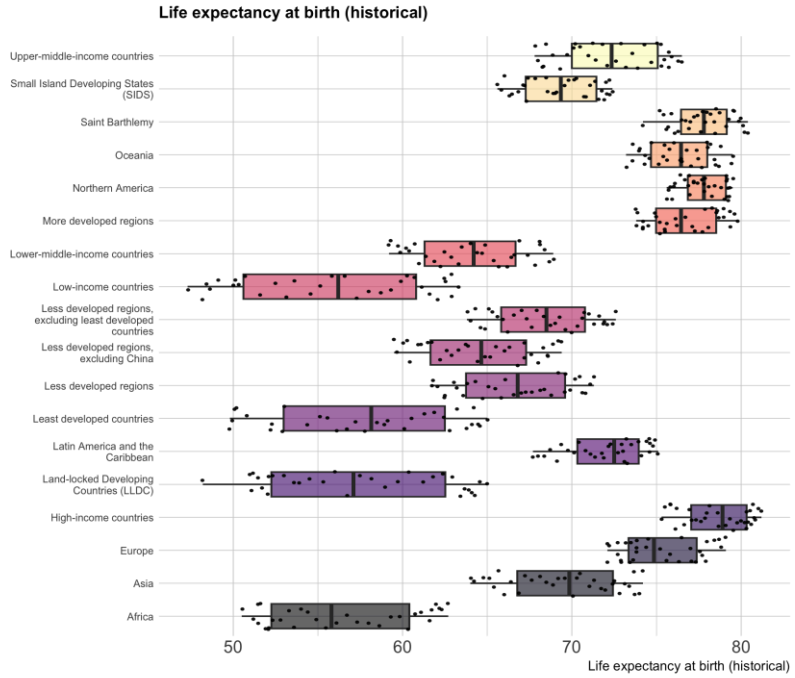
我们通过左连接的方式，与所有国家名称列表匹配国家（Entity），国家代码（Code），年份（Year）三个变量，生成最终的数据集。我们就此得到了 1990-2020 年间所有注册国家与地区的人口，经济，环境的面板数据，并基于此展开后续的分析研究。

在上述的数据表中，除去常规的国家数据，同时还包括若干地区数据。地区数据包括五大洲的聚合数据，以及发达国家、欠发达国家、高收入国家、低收入国家等等不同分类维度的聚合数据。这些聚合数据是对相关地区与国家的数据的一个综合的展示，在数据集中他们并不存在国家代码（Code）参数，因此在后续的研究中，我们将从两方面开展研究与讨论：单个国家层面和同类型国家层面。

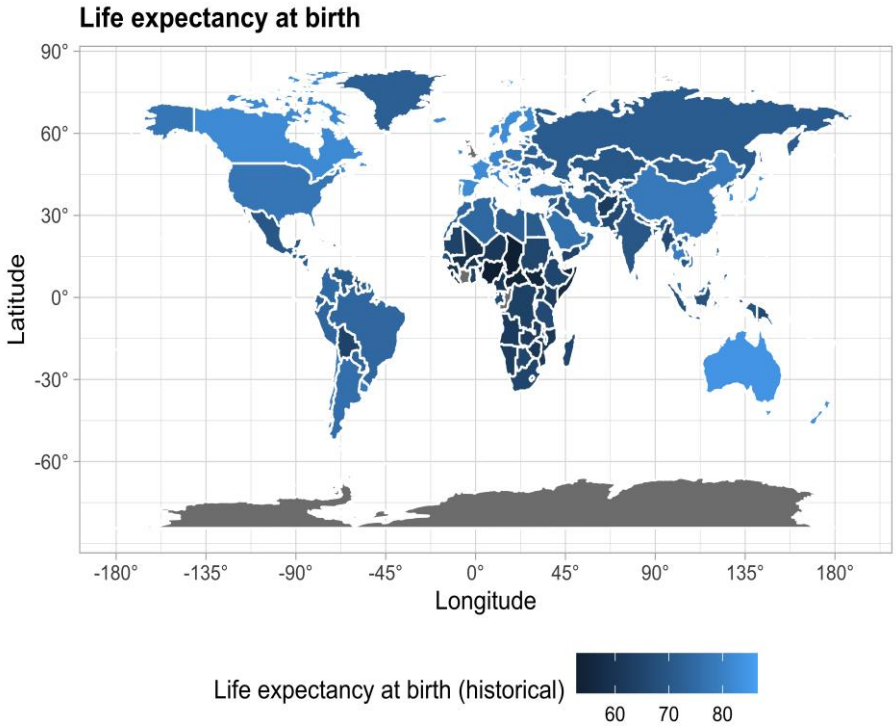
2 探索性数据分析

2.1 数据可视化

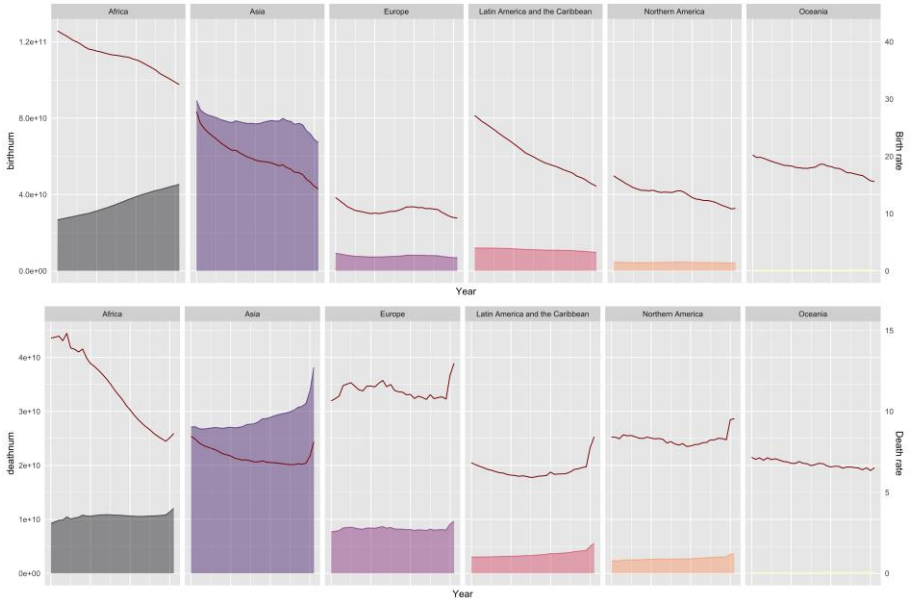
我们使用箱线图与散点图相结合，展现核心变量预期寿命 Life Expectancy 的分布情况，依照聚合数据进行分组展示。根据箱线图，可以清晰地发现北美，大洋洲以及高收入国家的预期寿命最高，其中位数位于 75-80 岁之间，同时这些地区的散点图相对密集，意味着在 1990-2020 年之间变化相对较小。而非洲、低收入国家和欠发达国家的预期寿命的中位数则远远低于其余地区，仅处于 55-60 岁之间，但是其箱线图的宽度较宽，散点图分布较为稀疏，说明在过去的 30 年之中，这些欠发达地区的预期寿命的变化非常大，预期寿命增长迅速。



图中的全球期望寿命地图，是利用 maps 包中的 worldmap 数据集，该数据集包含了全世界国家的经纬度信息，利用这些经纬度的坐标信息，与 2020 年的全球国家的期望寿命进行左连接，进行可视化地图的绘制，大致可以看到全世界期望寿命的分布情况。其中较为深色的是期望寿命相对较短的国家，我们可以发现几种在非洲以及一部分的拉丁美洲，而较为浅色的地区说明该地区的期望寿命较长，集中分布在北美、欧洲、大洋洲和东亚地区。

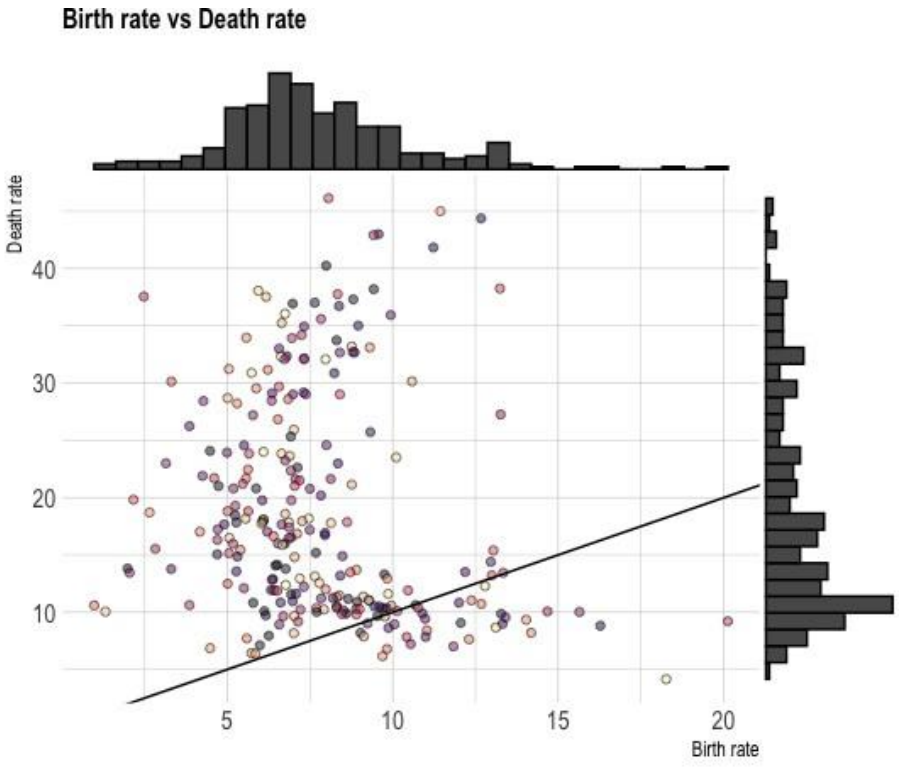


出生率，死亡率，迁移率，自然增长率是衡量一个地区人口增长的最主要的因素，而人口的期望寿命也与这些指标息息相关。我们首先对不同类型的国家进行分析。我们考虑 2020 年数据分布。如图所示，图中的横坐标表示该地区出生率与死亡率的比值，因此我们可以看到靠近图的右侧的是非洲、欠发达地区及低收入地区，这些地区有着较大出生率与死亡率的比值，较高的自然增长率，同时迁移率也大于发达国家的迁移率。(迁移率反映的是跨国移民对于本国人口的影响大小)



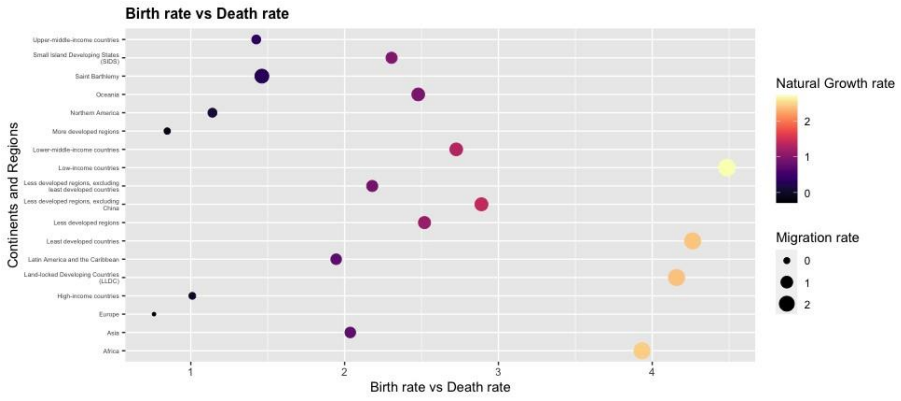
我们接下来考察所有国家的出生率与死亡率的分布情况, 如图所示，大量的国家的散点位于辅助线之上，这意味着在全世界二百多国家中，大部分国家的出生率大于死亡率，少部分国家已经进入老龄化社会。全世界各个国家出生率大约集中在 10%左右，而死亡率则大致位于 5%-10%之间。

在分析完 2020 年全世界出生/死亡率等数据后，我们对 1990-2020 年间的各大洲的出生人口、死亡人口、出生率和死亡率随着年份的变化进行可视化。图中的面积图为出生/死亡人口数量，折线图为变化率的大小。我们可以看到在过去的 30 年间，全世界所有的大洲的出生率都在不断降低，非洲的死亡率也大幅度降低，与之前的期望寿命的快速增长相吻合。尽管亚洲的死亡率在不断降低，但总死亡人数却在不断上升，以中国为代表的亚洲国家正在快速进入老龄化社会阶段。其余大洲的各项数据相对稳定，也与之前散点图中的期望较小的方差相吻合。

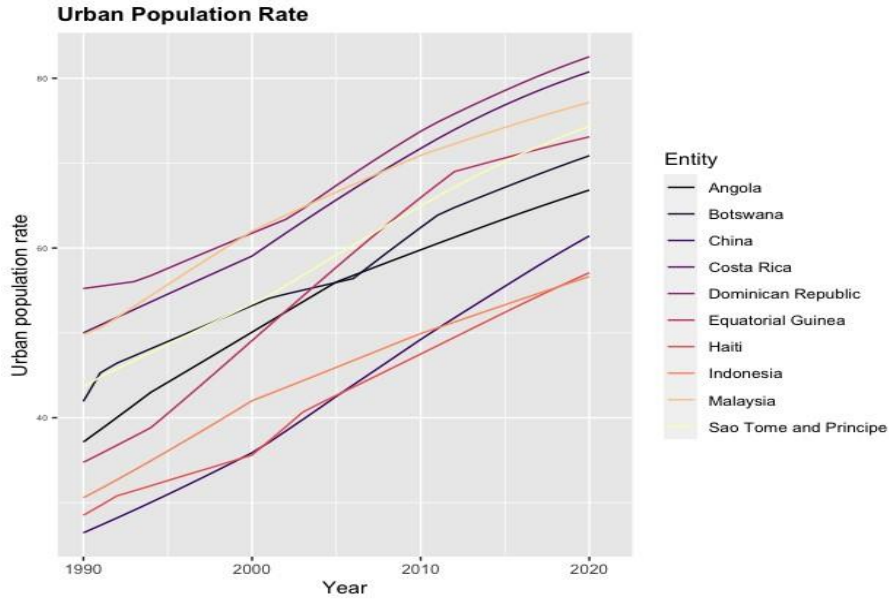


当我们对各地区、不同发展程度的国家进行分析，我们可以看到越靠近图的右侧，非洲、发达地区、低收入地区，出生率与死亡率的比值越大，其自然增长率也越高，同时移民影响

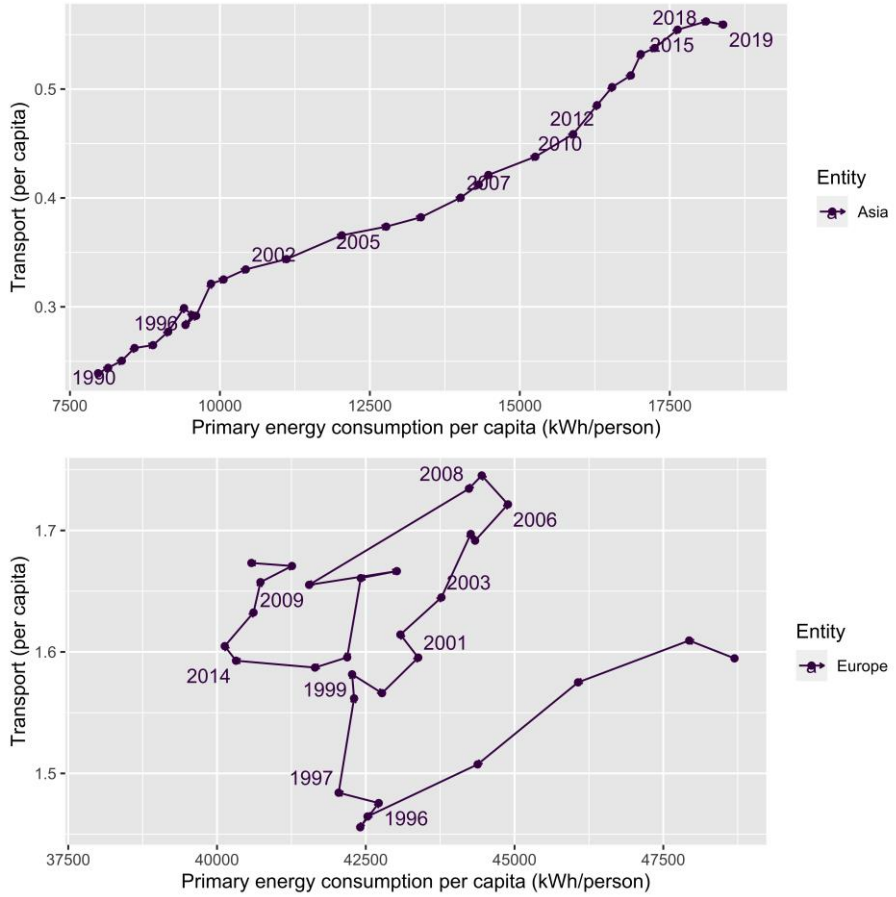
率（How large of an impact does migration have on population changes across the world?）也大于发达国家的影响率。



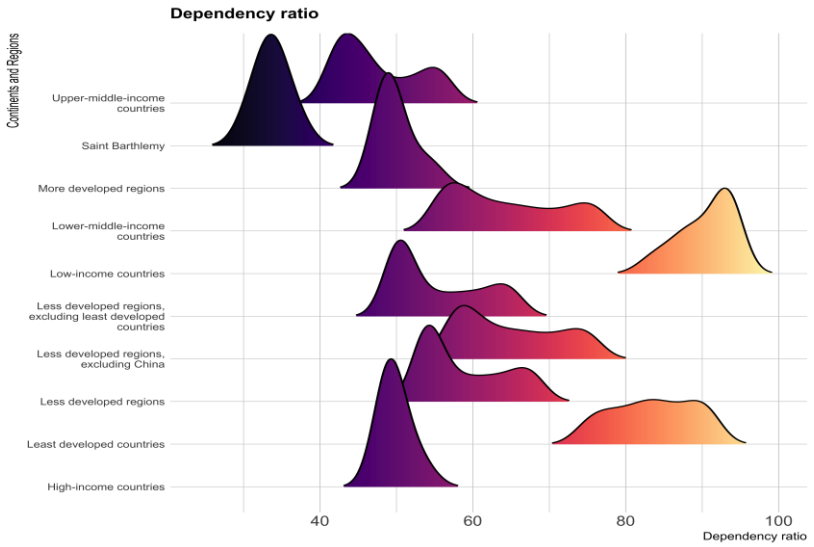
对于城市化率，我们利用 tidyverse 对数据集进行处理，计算 2020 年与 1990 年之间所有城市的城市化率的增长百分比，并按照其大小进行降序排序，绘制增长率最高的十座城市。



对于不同地区的交通的发展和能源消耗，我们选取亚洲和欧洲进行分析。可以看到在过去的 30 年间，亚洲国家的能源消耗稳步增加，交通发展迅猛，整个曲线呈现一个单调递增的趋势。亚洲国家以发展中国家为主，地区发展迅猛，城市化进程加速，能源消耗和交通的覆盖正在不断增长。而对于欧洲这个以发达国家居多的地区而言，主要能源的消耗正在逐渐地减少，同时交通的发展相对比较平稳，没有较大的变化。



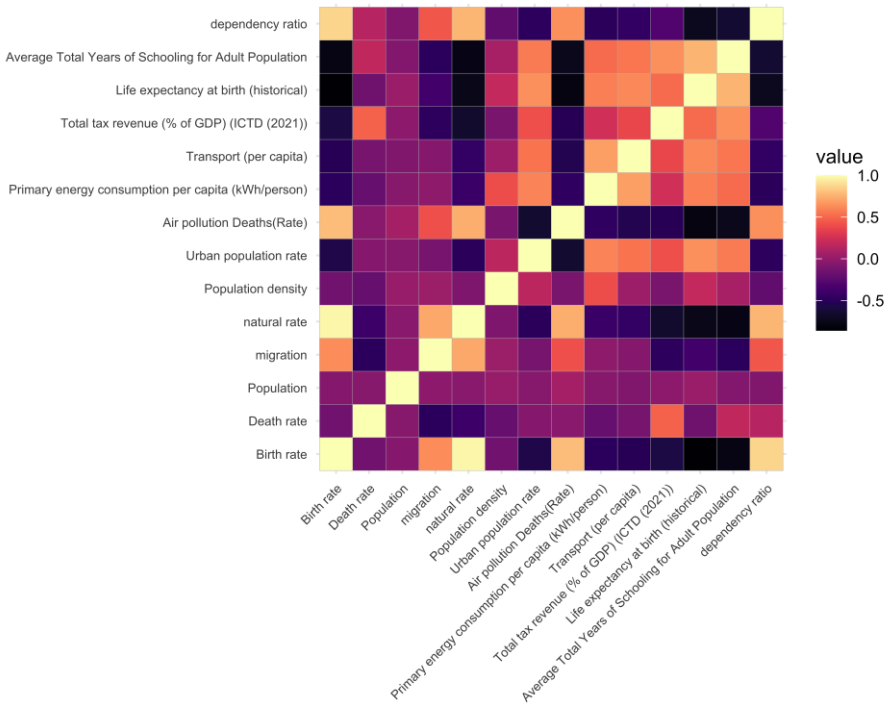
Dependency Ratio 赡养比率的连续数据，反应地区年龄架构的指标，抚养比越大，表明劳动力人均承担的抚养人数就越多，即意味着劳动力的抚养负担就越严重。我们可以看到低收入国家和欠发达地区的在近 30 年间的赡养比率远远大于发达地区与高收入地区，这说明有很多发展中国家同样也有很大的老龄化问题。



3 回归分析

3.1 相关系数

绘制变量之间的相关系数图来初步判断参数之间的相关性。可以看到色块越浅，越具有正的相关性，色块越深，越具有负的相关性，因此我们可以断定变量之间存在相互关联，可以进行后续的分析。



3.2 参数估计

我们考察全模型的 R^2,根据下方表格所得全模型的 R^2 为 0.9244，调整后的 R^2 为 0.9182，R^2 较大，说明模型拟合效果很好。

3.3 假设检验

我们首先进行进行全模型的 F 检验和对于每一个变量的显著性 t 检验。如表格所示，F 统计量为 147.9，p <0.05，则拒绝原假设，全模型存在显著性。对于每一个变量我们进行 t 检验，可以看到，存在有多个变量都与人口的期望寿命正相关/负相关。

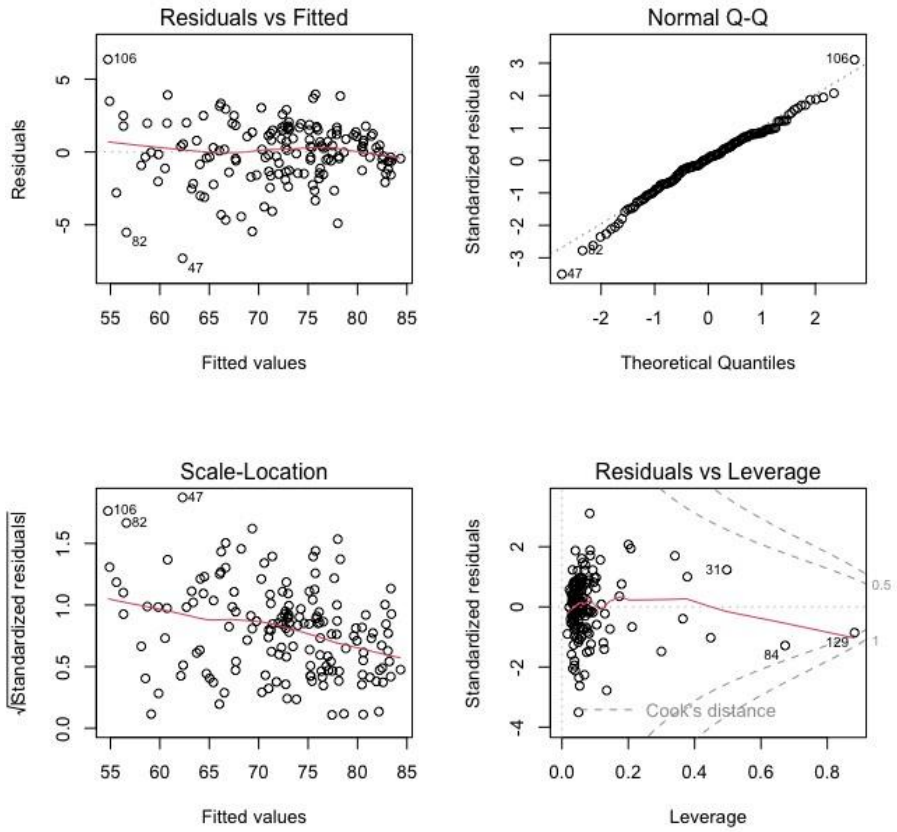
```
Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        7.970e+01  1.592e+00  50.052  < 2e-16 ***
`Birth rate`                                      -9.138e-01  6.772e-02 -13.493  < 2e-16 ***
`Death rate`                                      -1.220e+00  9.180e-02 -13.287  < 2e-16 ***
Population                                        -3.395e-10  1.142e-09  -0.297    0.767
migration                                          2.702e-01  2.095e-01   1.290    0.199
`Population density`                               2.146e-04  3.020e-04   0.711    0.478
`Urban population rate`                            1.801e-02  1.147e-02   1.570    0.119
`Air pollution Deaths(Rate)`                      -1.222e-03  4.171e-03  -0.293    0.770
`Primary energy consumption per capita (kWh/person)` 3.962e-06  9.410e-06   0.421    0.674
`Transport (per capita)`                           1.116e-01  2.343e-01   0.476    0.635
`Total tax revenue (% of GDP) (ICTD (2021))`       5.430e-02  2.634e-02   2.062    0.041 *
`Average Total Years of Schooling for Adult Population` 4.526e-01  1.053e-01   4.297 3.16e-05 ***
`dependency ratio`                                 2.429e-01  2.737e-02   8.873 2.44e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.134 on 145 degrees of freedom
Multiple R-squared:  0.9244,    Adjusted R-squared:  0.9182
F-statistic: 147.9 on 12 and 145 DF,  p-value: < 2.2e-16
```

3.4 模型诊断

接下来对于模型完成必要的模型诊断，检验模型是否符合高斯马尔科夫条件，并且关心数据中有没有异常点或强影响点。可以看到 QQ 图基本上成为一条直线，正态性假设基本成立，残差图的分布没有趋势性。同时我们根据 Cook 距离来检验是否有异常点或影响点，在图中我们可以看到异常值点较少，删除 129,84 的异常值点。接着我们借助 VIF 函数来考察变量的多重共线性问题。所有的变量中只有出生率这个变量的 VIF 值大于 10，出现了多重共线性的问题，其余变量的 VIF 值都远远小于 10，并且非常非常接近 1，当选择直接删去出生率这个变量后，多重共线性的问题消失。

随后我们将同时使用岭回归来尝试解决多重共线性问题。

```
                                `Birth rate`                                `Death rate`
                                       14.85                                        2.18
                                  Population                                   migration
                                        1.08                                        2.81
                         `Population density`                        `Urban population rate`
                                        1.41                                        2.30
                     `Air pollution Deaths(Rate)`  `Primary energy consumption per capita (kWh/person)`
                                        3.86                                        3.13
                          `Transport (per capita)`       `Total tax revenue (% of GDP) (ICTD (2021))`
                                        2.61                                        2.69
`Average Total Years of Schooling for Adult Population`                    `dependency ratio`
                                        3.56                                        7.35
```



3.5 变量选择

我们基于 AIC 准则和 BIC 准则来进行变量的选择。如图所示，与期望寿命强相关的变量较多，包括死亡率，人口千余率，空气污染死亡率，交通覆盖率，税收，平均受教育年限及赡养比。在此标准之下的 AIC 的值为 828.9266，当我们采用的 BIC 准则时候，BIC 的值为 868.7404。对比 AIC 准则和 BIC 准则，BIC 准则对于变量的选择确实更加苛刻，二者都认为死亡率、迁移率、空气污染死亡率、税收和平均受教育年限这五个变量对于预期寿命来说至关重要，而保守的 AIC 准则之下，认为交通覆盖率和赡养比同样对人们的期望寿命有所影响。

```
Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    75.9705128  2.2919053  33.147  < 2e-16 ***
`Death rate`                                   -0.9396852  0.1333771  -7.045 6.50e-11 ***
migration                                      -0.8375498  0.2856736  -2.932 0.003905 **
`Population density`                            0.0006371  0.0004027   1.582 0.115771
`Urban population rate`                         0.0291115  0.0163546   1.780 0.077124 .
`Air pollution Deaths(Rate)`                   -0.0253412  0.0054983  -4.609 8.67e-06 ***
`Transport (per capita)`                        0.6571573  0.2948291   2.229 0.027325 *
`Total tax revenue (% of GDP) (ICTD (2021))`    0.1566395  0.0376798   4.157 5.43e-05 ***
`Average Total Years of Schooling for Adult Population`  0.5234464  0.1549027   3.379 0.000929 ***
`dependency ratio`                             -0.0557938  0.0246123  -2.267 0.024845 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AIC 准则

```
Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    7.573e+01  2.342e+00  32.333  < 2e-16 ***
`Death rate`                                  -9.304e-01  1.336e-01  -6.964 1.04e-10 ***
Population                                     1.935e-09  1.691e-09   1.144  0.25433
migration                                     -8.538e-01  2.877e-01  -2.968  0.00351 **
`Population density`                           4.788e-04  4.510e-04   1.062  0.29015
`Urban population rate`                        2.481e-02  1.715e-02   1.447  0.15004
`Air pollution Deaths(Rate)`                  -2.653e-02  5.576e-03  -4.757 4.67e-06 ***
`Primary energy consumption per capita (kWh/person)`  1.134e-05  1.406e-05   0.807  0.42123
`Transport (per capita)`                       5.317e-01  3.476e-01   1.530  0.12824
`Total tax revenue (% of GDP) (ICTD (2021))`   1.542e-01  3.783e-02   4.075 7.51e-05 ***
`Average Total Years of Schooling for Adult Population`  5.213e-01  1.575e-01   3.311  0.00117 **
`dependency ratio`                            -4.833e-02  2.520e-02  -1.918  0.05709 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

BIC 准则

3.6 模型的预测

我们分别使用生成的全模型，AIC 准则下的选择模型，BIC 准则下的选择模型对 2021 年的期望寿命进行预测，并于 2021 年的真实数据进行比对，比较真实数据与预测数据，获得残差。如下表格所示，如果考虑全模型，该预测误差为 1.5604，在此基础上经过 AIC 变量选择后的模型精度为 2.312，与全模型相比反而较差，经过 BIC 模型筛选后的与 AIC 模型的精度相近。

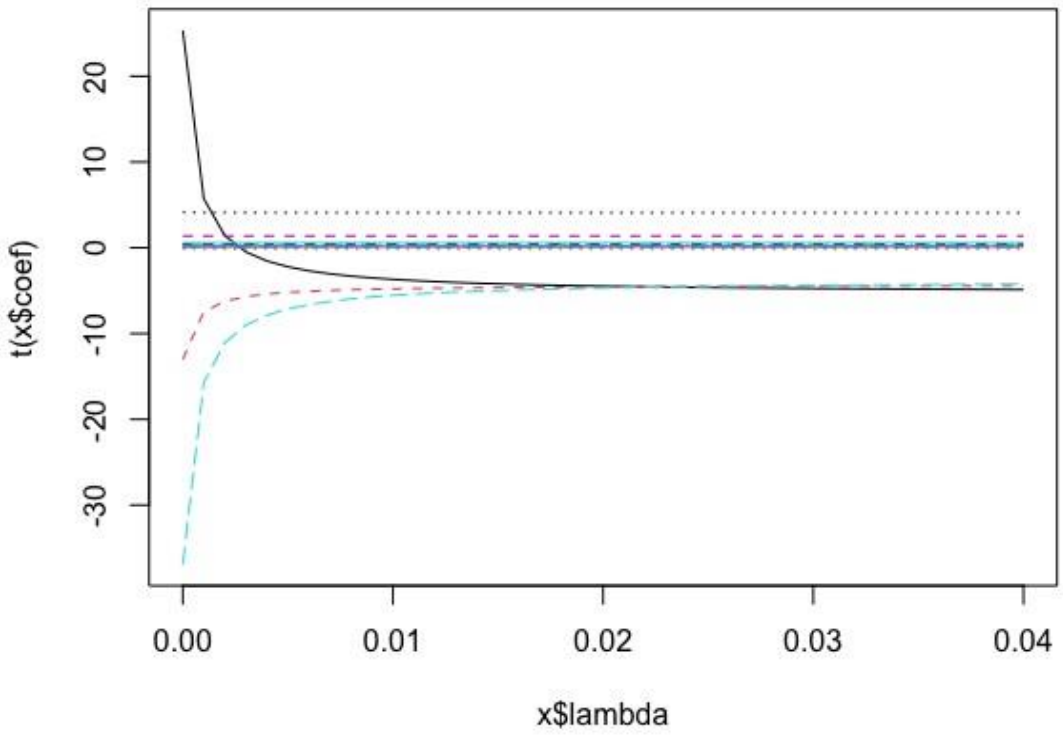
全模型	AIC 模型	BIC 模型
1.560464	2.312448	2.306004

3.7 岭回归

在前文的模型诊断中，借助 VIF 函数，我们考察了每一个变量的多重共线性，而后发现出生率 Birth rate 具有较大的多重共线性，当时我们选择了删除改变量进行后续的回归分析，在此我们借助岭回归来对解决多重共线性的问题。

我们计算 Hoerl-Kennard 公式下的 k 的值，并且使用岭迹法对不同的 k 下的各个回归系数的岭估计的值进行绘制，观察岭迹图，得到以下结果。

HKB	岭迹法
0.02296082	0.02



岭迹法在 □ =

0.02 的条件下的岭回归，各变量的系数如下所示：

	`Birth rate`
7.969077e+01	-4.633566e-01
`Death rate`	Population
-1.668904e+00	-3.395368e-10
migration	`natural rate`
2.683131e-01	-4.493928e+00
`Population density`	`Urban population rate`
2.118205e-04	1.801984e-02
`Air pollution Deaths(Rate)`	`Primary energy consumption per capita (kWh/person)`
-1.255350e-03	4.039038e-06
`Transport (per capita)`	`Total tax revenue (% of GDP) (ICTD (2021))`
1.129931e-01	5.434971e-02
`Average Total Years of Schooling for Adult Population`	`dependency ratio`
4.524103e-01	2.426353e-01

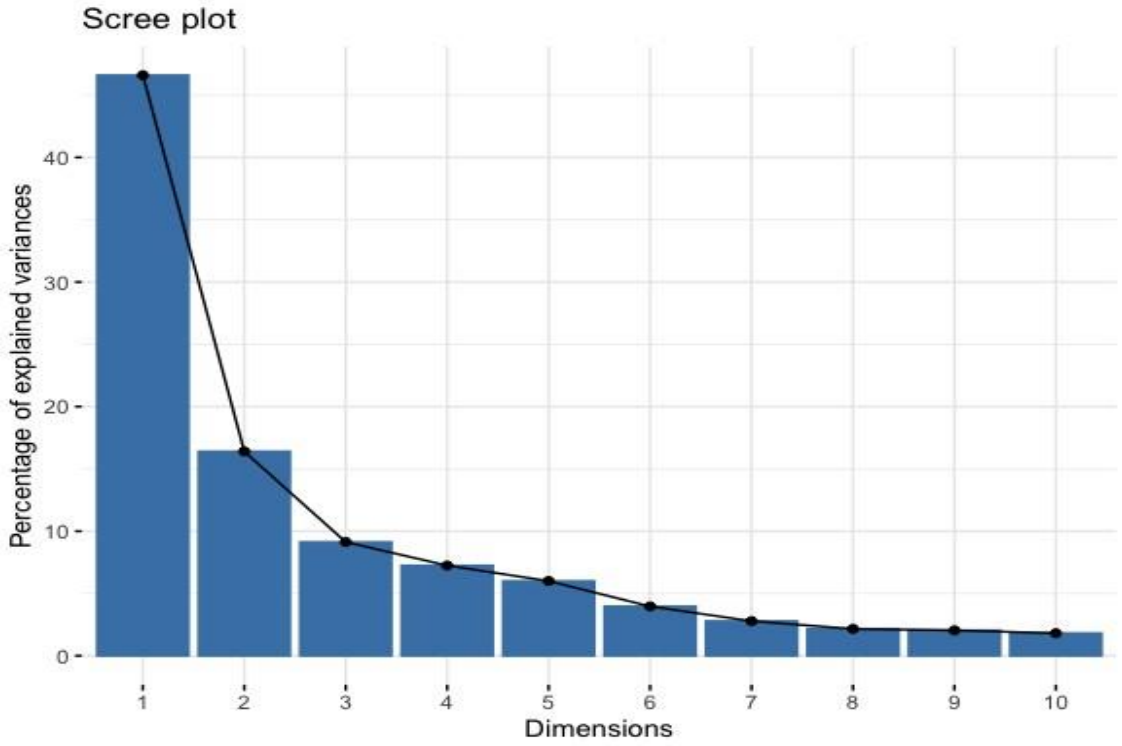
3.8 主成分估计

对于模型中存在的复共线性，我们同时也尝试了一下主成分估计。我们首先计算设计矩阵 □□□ 的特征向量的线性组合所组成的主成分，排在第一排的新变量对应于 □□□ 的最大特征值，即为第一主成分，排在第二列的即为第二主成分，计算所得到的全部的主成分如下：

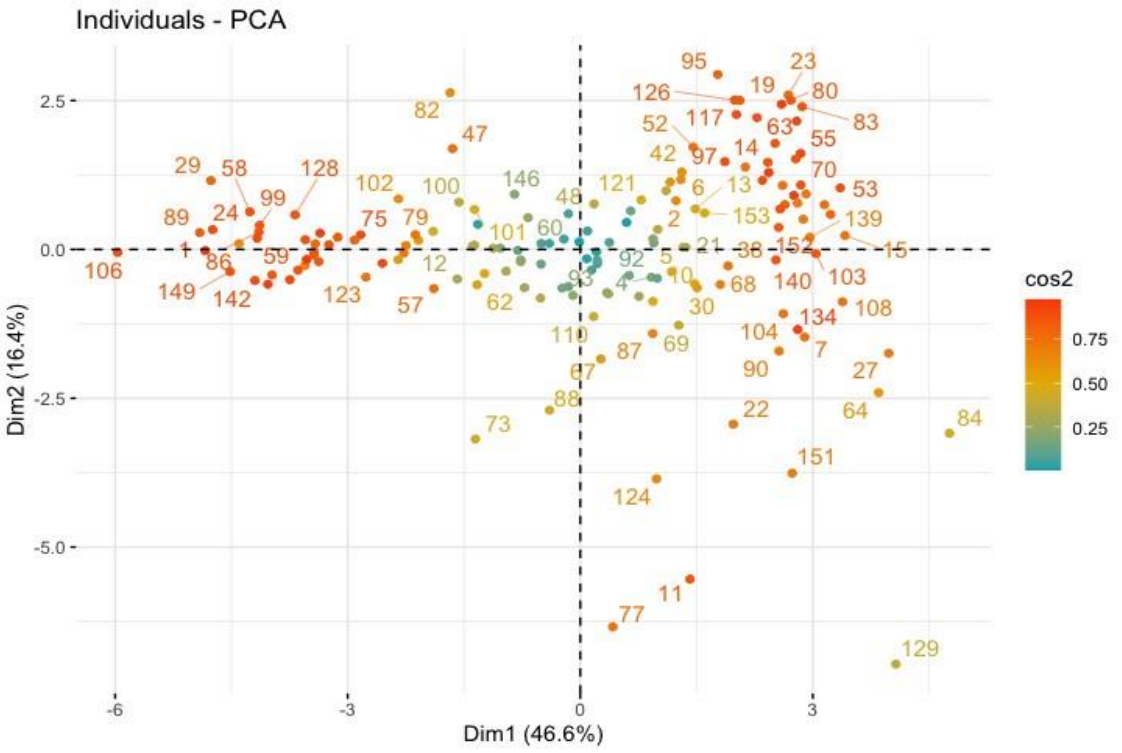
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
`Birth rate`	-0.385326032	-0.04790336	0.18237940	-0.002995885	0.16975391	-0.03701792	0.03050084	-0.006759931	-0.214754435	-0.10255017
`Death rate`	0.075815492	0.55034498	0.15722382	0.185932114	0.43105165	0.11679745	-0.39415359	0.146307239	0.293860967	-0.26165489
Population	-0.006669095	0.02408088	-0.64855594	-0.604410801	0.43845362	-0.03233245	0.01286944	0.083169755	-0.069823847	-0.08543578
migration	-0.237104676	-0.41338636	0.23270765	-0.183400138	0.16812075	-0.13585343	0.17125192	0.020451139	0.731730364	-0.24520644
`natural rate`	-0.376728431	-0.18866193	0.12810973	-0.051759600	0.04427814	-0.06469686	0.13139248	-0.044525822	-0.276595475	-0.02637944
`Population density`	0.066151746	-0.31645407	-0.36905398	0.685851141	0.38769638	-0.06530718	0.21257190	0.181703233	0.004611426	0.08021759
`Urban population rate`	0.282684750	-0.22967407	0.19992604	-0.107861196	0.18455676	-0.62876066	-0.45608316	-0.214113717	-0.120589534	0.16528024
`Air pollution Deaths(Rate)`	-0.346288723	0.05351762	-0.06307861	0.035330909	0.15357863	0.38294049	-0.11364184	-0.659608311	0.038204376	0.14275538
`Primary energy consumption per capita (kWh/person)`	0.253750408	-0.402147S5	0.09163320	0.068137265	0.29926269	0.33746820	-0.23399035	-0.298675552	-0.133960868	-0.11382297
`Transport (per capita)`	0.264104527	-0.27432468	0.30323693	-0.257835271	0.09271661	0.52974589	-0.01910691	0.420621639	-0.088238600	0.19609667
`Total tax revenue (% of GDP) (ICTD (2021))`	0.275598296	0.28192273	0.24652981	-0.095501374	0.34069127	-0.08484108	0.58320560	-0.243488400	0.095959583	0.45850781
`Average Total Years of Schooling for Adult Population`	0.355772707	0.03214499	0.09942613	-0.014665920	-0.01200439	-0.01369392	0.36388714	-0.239440415	-0.233062112	-0.73041265
`dependency ratio`	-0.326675273	0.12048351	0.31570035	0.007859885	0.37494815	-0.11467939	0.05642268	0.269330009	-0.381360831	-0.06021377

	PC11	PC12	PC13
`Birth rate`	0.0342803886	0.5311632996	-6.669705e-01
`Death rate`	0.0592980131	0.2485301964	1.891329e-01
Population	0.0001053093	0.0298991504	6.221824e-05
migration	0.0017243053	-0.1143673399	1.858664e-04
`natural rate`	0.0157846205	0.4256408708	7.206790e-01
`Population density`	0.2223239827	0.0150855201	3.149378e-04
`Urban population rate`	0.3010271211	-0.0005271037	-1.501365e-05
`Air pollution Deaths(Rate)`	0.4214445120	-0.2330477211	-5.295064e-05
`Primary energy consumption per capita (kWh/person)`	-0.6182528482	0.0175884462	-3.360453e-04
`Transport (per capita)`	0.4258553740	0.0495541665	-1.291915e-04
`Total tax revenue (% of GDP) (ICTD (2021))`	-0.1304072994	0.1246596989	1.091648e-04
`Average Total Years of Schooling for Adult Population`	0.2865017593	-0.0480708291	-3.229631e-05
`dependency ratio`	-0.1213136378	-0.6211275127	-8.275461e-04

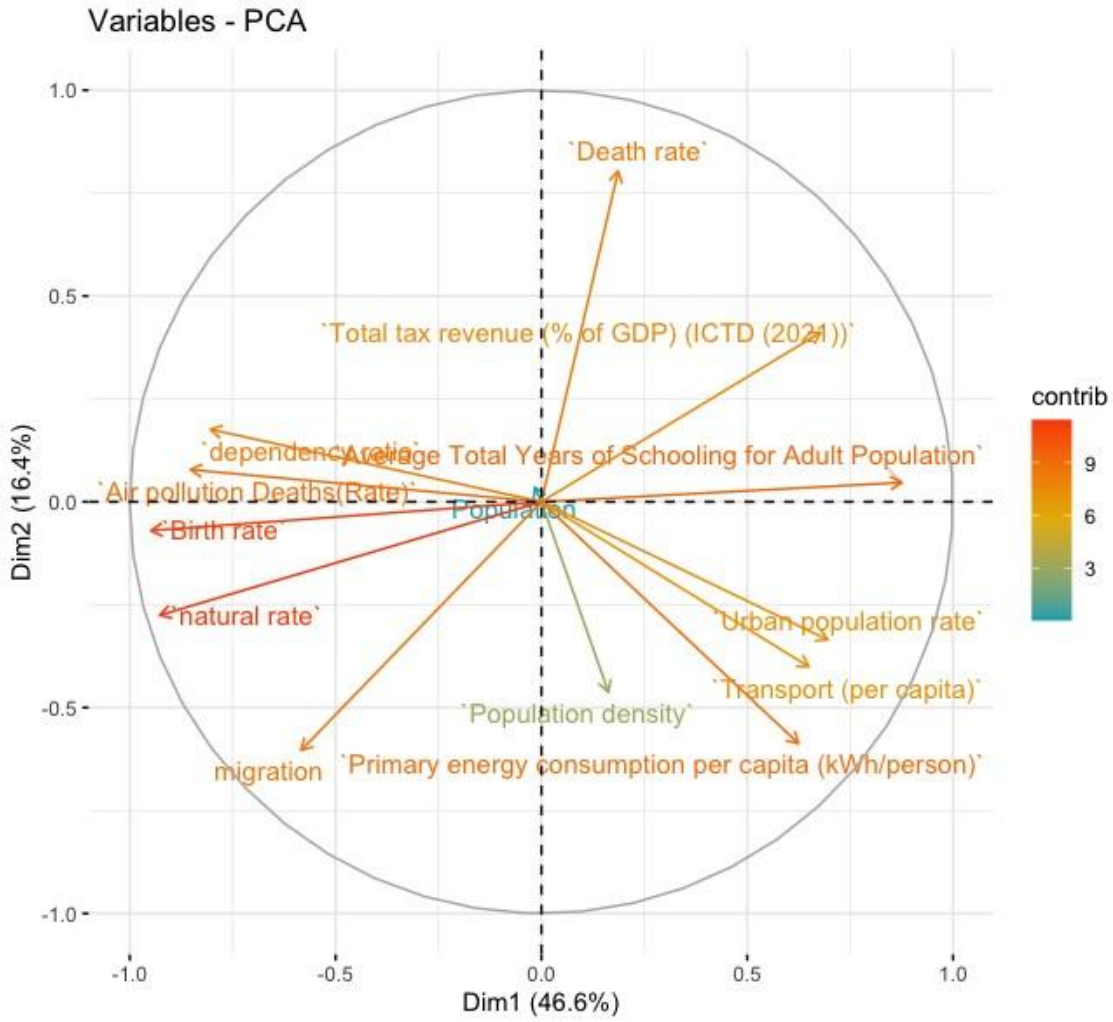
下列三张图中，第一张展示的是各个样本在第一主成分和第二主成分这两个维度中的大小，第二章展示的是不同变量在第一主成分和第二主成分这两个维度中贡献程度和方向，第三张图片展示的是不同的样本和变量在第一主成分和第二主成分这两个维度中的方向以及相对位置。若与给定的变量处于同一侧，则对该变量具有较高的值。



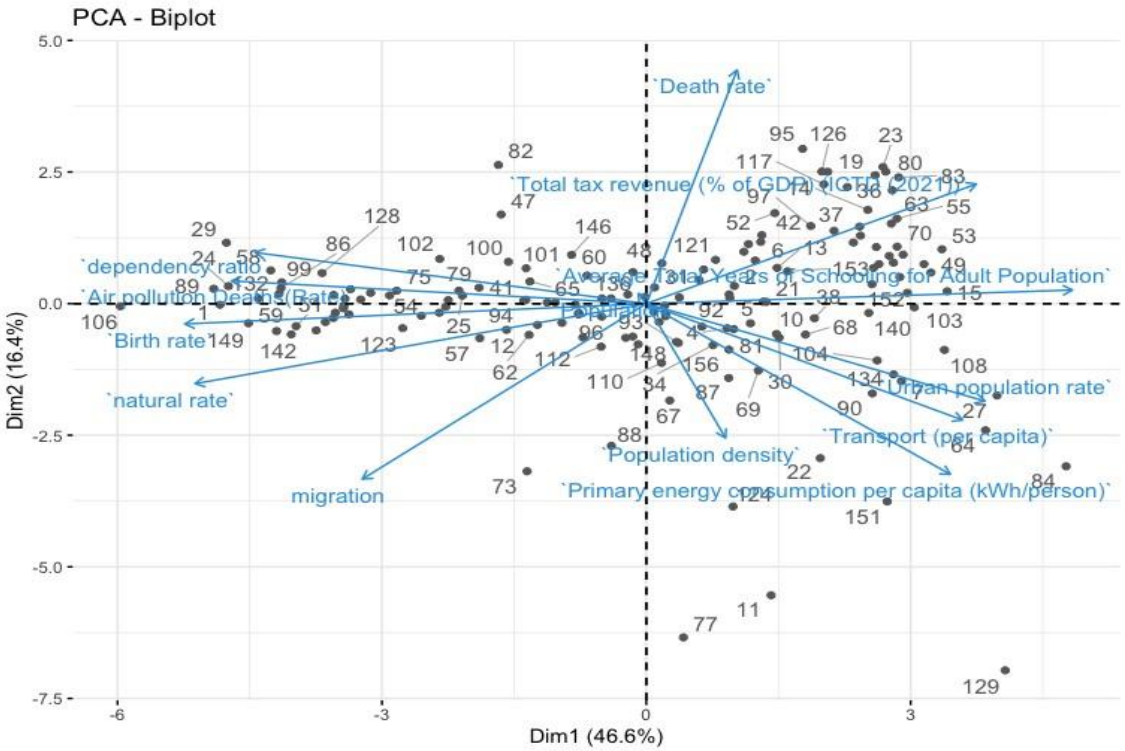
通过该图，可以观察到第一主成分贡献接近一半，而从第六主成分开始，主成分贡献程度的减退速度明显变慢，考虑选取前 5-6 个主成分来解释模型。



通过该图，可以发现大部分样本点位于第四象限，少部分集中在 x 周负半轴附近，即大量样本点与第一主成分有正向关系且与第二主成分有负向关系，而和第一主成分负相关的样本与第二主成分相关性低。

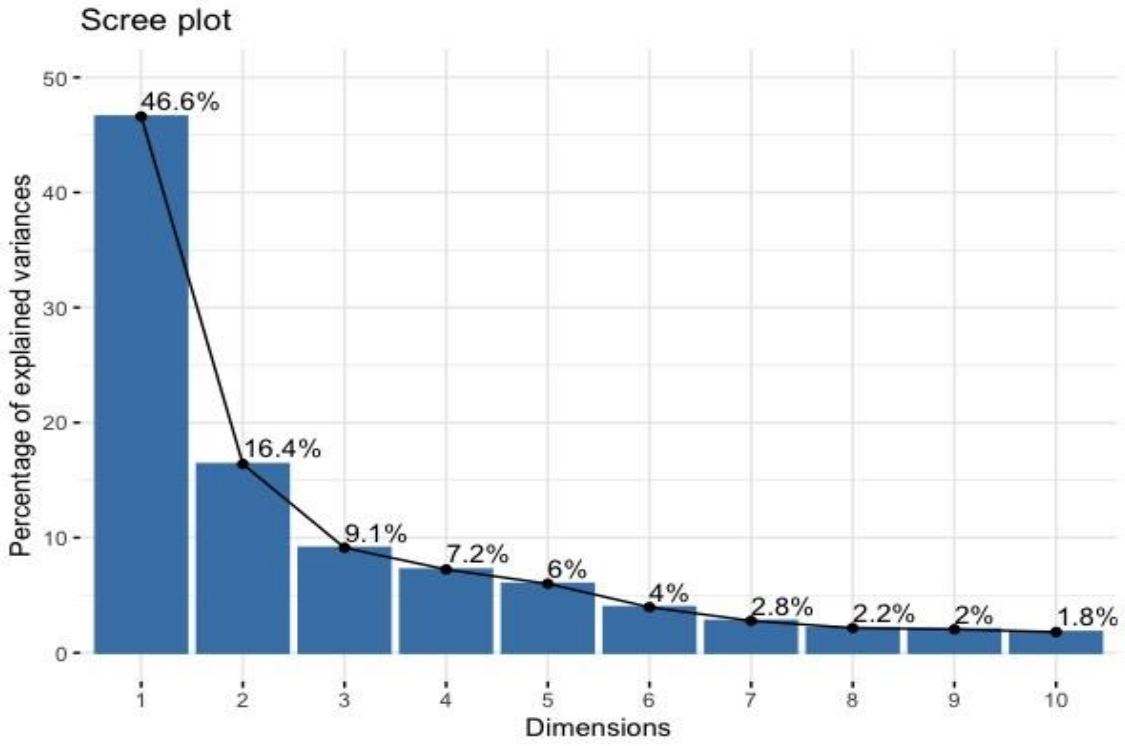


通过本图，可以观察到：人口密度、城市人口密度、交通、能源密度位于第一象限，和第一、二主成分都是正向关系；自然增长率、高抚养率和移民率位于第二象限，和第一主成分有负向关系，和第二主成分有正向关系；出生率位于第三象限，和第一主成分有负向关系，和第二主成分有负向关系；死亡率、平均教育年限、中等抚养率和税收收入位于第四象限，和第一主成分有正向关系，和第二主成分有负向关系。



图三将样本点和变量结合起来，便于更清楚的观察样本点、变量和第一、二主成分之间的关系。

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	6.057724e+00	4.659788e+01	46.59788
Dim.2	2.131857e+00	1.639890e+01	62.99678
Dim.3	1.186458e+00	9.126598e+00	72.12337
Dim.4	9.414676e-01	7.242059e+00	79.36543
Dim.5	7.797884e-01	5.998372e+00	85.36380
Dim.6	5.151890e-01	3.962992e+00	89.32680
Dim.7	3.615049e-01	2.780807e+00	92.10760
Dim.8	2.803617e-01	2.156629e+00	94.26423
Dim.9	2.651837e-01	2.039874e+00	96.30411
Dim.10	2.351330e-01	1.808716e+00	98.11282
Dim.11	1.823698e-01	1.402845e+00	99.51567
Dim.12	6.295980e-02	4.843061e-01	99.99997
Dim.13	3.564387e-06	2.741836e-05	100.00000

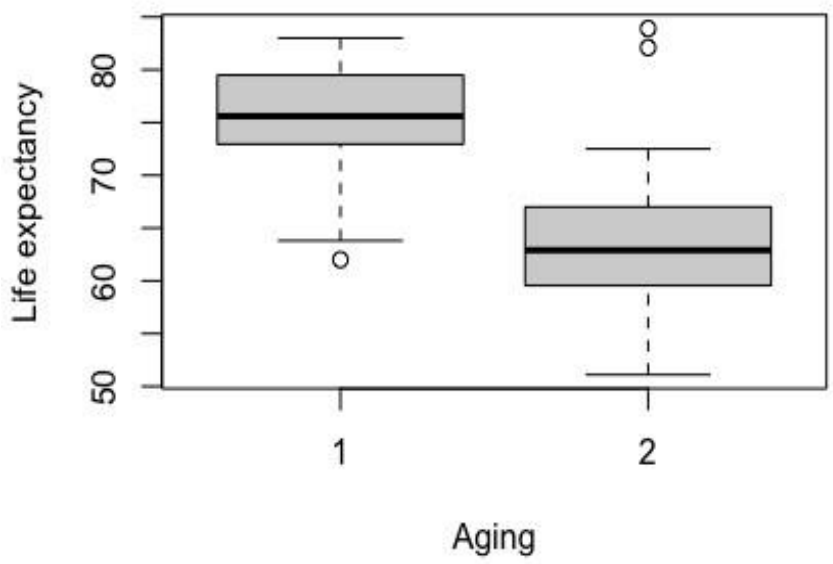


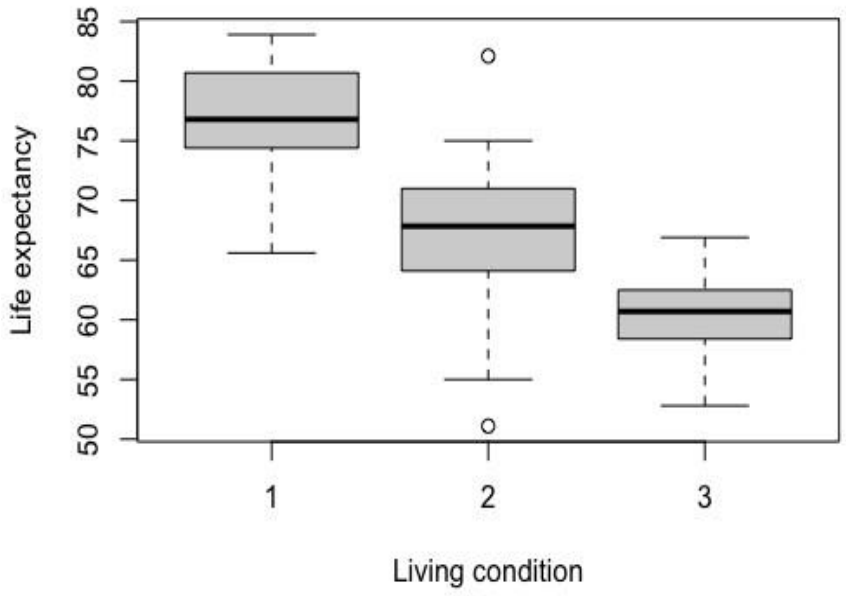
计算每个主成分对原始变量的解释程度，按照降序排列，并计算累计贡献率。通过以上两个图表，可以很清晰地观察到，前 6 个主成分的贡献和超过了 80%，且从第七主成分起，方差的解释率解释率在 5% 以下而且递减速度放缓，说明该点是一个拐点，建议提取前 6 个主成分。

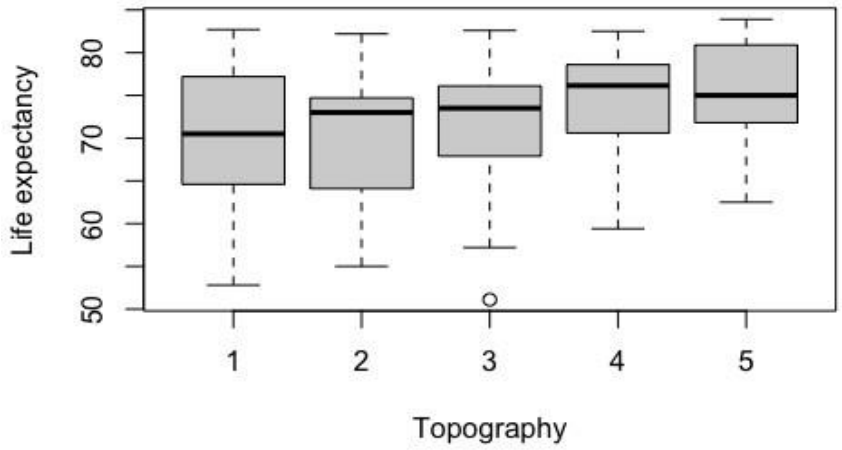
4 协方差分析

在准备变量时，我们同时还准备了若干离散变量，在上文中我们并没有将他们纳入模型。利用协方差分析，我们综合考虑离散变量和连续变量，对模型进行分析。

4.1 数据可视化


Life expectancy
50
60
70
80
1
2
Aging


Life expectancy
50
55
60
65
70
75
80
85
1
2
3
Living condition



我们首先绘制箱线图, 根据箱线图，我们大致可以判断这三者与期望寿命之间的关联。其中是否是老龄化国家与期望寿命有着显著的关联，城市宜居程度由高到低所对应的期望寿命亦是逐渐变短，而最后一个国家所处地形，其相关性暂时还未可知。

4.2 三型方差分析我们首先对原模型进行三型的方差分析,

结果如下图

Response: Life expectancy at birth (historical)

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	6891.5	1	1210.2251	< 2.2e-16	***
`Birth rate`	294.6	1	51.7324	3.590e-11	***
`Death rate`	644.2	1	113.1260	< 2.2e-16	***
Population	2.1	1	0.3624	0.5481695	
migration	4.1	1	0.7274	0.3952049	
`Population density`	0.1	1	0.0207	0.8858923	
`Urban population rate`	8.9	1	1.5593	0.2138662	
`Air pollution Deaths(Rate)`	4.3	1	0.7595	0.3849918	
`Primary energy consumption per capita (kWh/person)`	8.6	1	1.5069	0.2216824	
`Transport (per capita)`	0.0	1	0.0003	0.9853968	
`Total tax revenue (% of GDP) (ICTD (2021))`	108.6	1	19.0758	2.438e-05	***
`Average Total Years of Schooling for Adult Population`	113.7	1	19.9700	1.618e-05	***
as.factor(Aging)	77.8	1	13.6633	0.0003138	***
as.factor(`Living condition`)	148.3	2	13.0188	6.569e-06	***
as.factor(`Medical Care`)	23.0	4	1.0103	0.4043968	
Residuals	791.5	139			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

而后再删除全模型中不显著的变量，对其重新进行拟合。我们可以看到，是否老龄化和城市宜居程度确实展现了较高的显著性，保留在了模型中，而国家所处地形未被选中。重新拟合的结果如下

```
Response: Life expectancy at birth (historical)
                                                  Sum Sq  Df  F value    Pr(>F)
(Intercept)                                      11566.1   1 2038.467 < 2.2e-16 ***
`Death rate`                                      1158.3   1  204.149 < 2.2e-16 ***
`Birth rate`                                       377.0   1   66.446 1.315e-13 ***
`Total tax revenue (% of GDP) (ICTD (2021))`       154.6   1   27.255 5.852e-07 ***
`Average Total Years of Schooling for Adult Population`   175.4   1   30.914 1.203e-07 ***
as.factor(Aging)                                    65.2   1   11.487 0.0008958 ***
as.factor(`Living condition`)                      229.5   2   20.222 1.676e-08 ***
Residuals                                          851.1 150
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

我们对两个模型的 AIC 值进行计算，可以看到，不论是根据 AIC 标准还是 BIC 标准，我们都认为简化后的模型优于全模型。

	全模型	筛选后模型
AIC	742.9806	732.4438
BIC	804.2325	760.0072

4.3 模型选择

借助 R 自动搜索比较众多线性模型后，分别利用 AIC 和 BIC 准则，在筛选模型中选择最优模型。

```
Response: Life expectancy at birth (historical)
                                                  Sum Sq  Df   F value    Pr(>F)
(Intercept)                                      10343.6   1 1848.1438 < 2.2e-16 ***
`Birth rate`                                       373.4   1   66.7114 1.236e-13 ***
`Death rate`                                       982.4   1  175.5317 < 2.2e-16 ***
`Urban population rate`                             17.2   1    3.0677 0.0819220 .
`Total tax revenue (% of GDP) (ICTD (2021))`       131.7   1   23.5317 3.064e-06 ***
`Average Total Years of Schooling for Adult Population`   135.1   1   24.1352 2.338e-06 ***
as.factor(Aging)                                    69.5   1   12.4230 0.0005636 ***
as.factor(`Living condition`)                      173.9   2   15.5399 7.417e-07 ***
Residuals                                          833.9 149
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Response: Life expectancy at birth (historical)
                                                  Sum Sq  Df  F value    Pr(>F)
(Intercept)                                      11566.1   1 2038.467 < 2.2e-16 ***
`Death rate`                                      1158.3   1  204.149 < 2.2e-16 ***
`Birth rate`                                       377.0   1   66.446 1.315e-13 ***
`Total tax revenue (% of GDP) (ICTD (2021))`       154.6   1   27.255 5.852e-07 ***
`Average Total Years of Schooling for Adult Population`  175.4   1   30.914 1.203e-07 ***
as.factor(Aging)                                    65.2   1   11.487 0.0008958 ***
as.factor(`Living condition`)                      229.5   2   20.222 1.676e-08 ***
Residuals                                          851.1 150
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

比较二者的模型选择，可以看到，在 0.05 的显著性标准下，AIC 准则和 BIC 准则的选择相一致，同时也与利用假设检验方法所得到的模型结果相一致。最终与人们的期望寿命相关的是出生率、死亡率、税收、平均受教育年限，以及离散变量老龄化程度和城市宜居程度。

4.4 模型预测

我们利用上述生成的模型来评估各个国家的寿命。如下图所示,est.Life 反映了剔除了所有模型中所有重要的相关因素后各个国家的期望寿命，即为模型中的残差项，而现在的 Life expectancy 即为模型所估计的期望寿命，可以看到大部分国家的估计偏差较小。

Description: df [20 × 18]

Average Total Years of Schooling for Adult Population <dbl>	dependency ratio <dbl>	Life expectancy at birth (historical) <dbl>	Aging <dbl>	Living condition <dbl>	Topography <dbl>	est.life <dbl>
9.3	30.11	79.12015	1	1	5	0.27984680
5.2	53.87	71.22082	1	1	5	-0.72082240
10.5	46.71	77.69234	1	1	5	-0.99234305
12.2	45.50	73.94821	1	1	2	0.05178575
11.7	54.31	80.23186	1	1	5	0.66813542
10.5	56.37	74.19808	1	2	1	-0.99807674
3.1	48.59	70.65938	1	1	1	-0.35938270
8.7	61.16	69.70828	1	2	1	-2.40828368
9.0	43.62	75.77956	1	1	3	0.42044009
9.2	59.85	67.16270	1	2	1	-3.36269554

11–20 of 20 rows | 13–19 of 18 columns

Previous 1 2 Next