使用多元线性回归模型对中国人口增长率进行分析

张博强 大数据 2 班 32019070233

一、摘要

本文对计划生育之后的中国人口增长率进行了研究分析，研究影响人口增长率的因素。为了研究，从国家统计局官网中国统计年鉴中整理出了 1988-2006 年的连续的人口增长率及其相关数据。对数据用 R 语言建立了多元线性回归模型，并对建立的模型进行分析，并处理的模型的多重共线性问题。最终得出结论，揭露了一些因素对人口增长率的影响。

二、引言

从 1971 年到 2013 年为止，中国开展了长达四十二年的计划生育政策。在 1980 年 9 月，党中央发表《关于控制我国人口增长问题致全体共产党、共青团员的公开信》，提倡一对夫妇只生一个孩子。这项政策推出之后，使得中国自然增长率从 1970 年的 5.8 降到 1980 年的 2.24，效果显著。本文也将研究在此之后的自然增长率的影响因素。

自然增长率也很大程度上与经济发展等各方面因素相联系，与我们的经济生活息息相关，为了研究这时影响中国自然增长率的原因，分析人口增长，和猜测中国未来的增长趋势，需要建立研究模型。在此文中，采用了多元线性回归模型。

影响自然增长率的因素有很多，大致可以有一下几种：从国民经济上来看，经济增长是人口自然增长的基本源泉；居民消费水平也会一定程度上反应出人口增长率；居民的文化程度，会影响人民的思想，从而影响人口的自然增长率；人口分布，非农业和农业人口的占比和分布也会对人口自然增长率造成影响。

三、模型建立及数据收集

从国家统计局官网中国统计年鉴中整理收集获得 http://www.stats.gov.cn/tjsj/ndsj/

收集到的信息整理成如下数据信息：

https://github.com/13roky/The_study_of_natural_population_growth_rates/blob/master/%E4%B8%AD%E5%9B%BD%E8%87%AA%E7%84%B6%E5%A2%9E%E9%95%BF%E7%8E%87%E5%8F%8A%E7%9B%B8%E5%85%B3%E6%95%B0%E6%8D%AE.csv

年份	人口自然增长率（%）	国民总收入（亿元）	居民消费价格指数增长率（%）	人均GDP（元）
1988	15.73	15037	18.8	1366
1989	15.04	17001	18	1519
1990	14.39	18718	3.1	1644
1991	12.98	21826	3.4	1893
1992	11.6	26937	6.4	2311
1993	11.45	35260	14.7	2998
1994	11.21	48108	24.1	4044
1995	10.55	59811	17.1	5046
1996	10.42	70142	8.3	5846
1997	10.06	78061	2.8	6420
1998	9.14	83024	-0.8	6796
1999	8.18	88479	-1.4	7159
2000	7.58	98000	0.4	7858
2001	6.95	108068	0.7	8622
2002	6.45	119096	-0.8	9398
2003	6.01	135174	1.2	10542
2004	5.87	159587	3.9	12336
2005	5.89	184089	1.8	14040
2006	5.38	213132	1.5	16024

（表 1 中国人口自然增长率及相关数据）

根据搜集的数据，为了更加全面的了解人口增长率的影响因素，选择人口自然增长率为解释变量，以此来反映人口增长。选择国民总收入和人均 GDP 作为精致增长的指标。选择居民消费价格指数增长率作为居民消费水平的代表。鉴于搜集到的数据，暂且考虑这些影响因素。

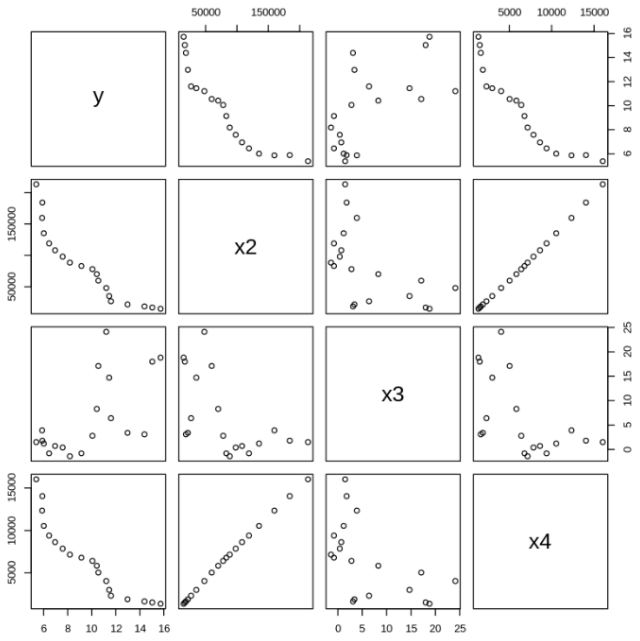
根据以上模型，建立多元线性回归模型：

$Y = \beta1+\beta2\mathbf{X2}+\beta3\mathbf{X3}+\beta4\mathbf{X4}$

四、方法介绍

对数据本身进行可视化：

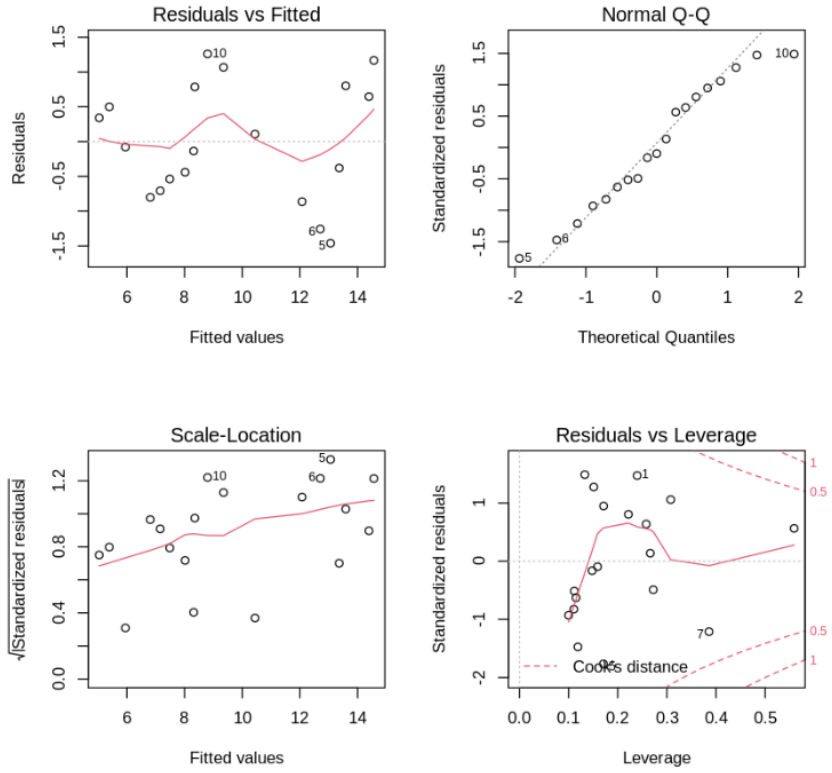
1. 使用 plot(data)，对数据进行散点图可视化



利用最小二乘估计模型，使用 R 语言估计步骤如下：

1. 建立相应的文件，将数据文件和 R 存储在同一目录下，然后使用，data = read.csv("中国增长率及相关数据.csv") 函数读取数据，使用 colnames(data) = c("y","x2","x3","x4")方法对列名做对应的参数更改。

2. 对于修改好的数据，使用 fit=lm(y~.,data=data)对数据进行最小二乘的多元线性回归拟合用 summary(fit)进行估计。拟合模型效果如下可见效果不错：



3. 估计结果如下：

```
Call:
lm(formula = y ~ ., data = data)

Coefficients:
(Intercept)           x2           x3           x4
 15.7197750    0.0003751    0.0497390   -0.0056601
```

（表 2）

```
[4]: summary(fit)


Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4613 -0.6229 -0.0797  0.7153  1.2592

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.7197750  0.8702058  18.064 1.37e-11 ***
x2           0.0003751  0.0001061   3.535  0.00300 **
x3           0.0497390  0.0329629   1.509  0.15209
x4          -0.0056601  0.0014259  -3.969  0.00123 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9091 on 15 degrees of freedom
Multiple R-squared:  0.9363,    Adjusted R-squared:  0.9236
F-statistic:  73.5 on 3 and 15 DF,  p-value: 3.379e-09
```

（表 3）

得到的多元回归模型为：

Y = 15.7197750+0.0003751 **X2**+0.0497390 **X3**-0.0056601 **X4**

S=（0.8702058）（0.0001061）（0.0329629）（0.0014259）

t=（18.064）　　(3.535)　　(1.509)　　(-3.969)

R^2=0.9363　　A-R^2=0.9236

F=73.5

五、数据分析

1. 经济意义检验

模型估计结果可以知道，在其他变量不变的情况下，当年国民总收入每增长 1 亿元，人口增长率增长 0.0003751%；在其他变量不变的情况下，当年居民消费价格指数增长率每增长 1%，人口增长率增长 0.0497390%；在其他条件不变的情况下，当人均 GDP 每增加 1 元，人口增长率就会降低 0.0056601%。这些理论分析和经验判断一致。

2. 统计检验

拟合优度：由表 3 中数据可以得到 R^2=0.9363 ，修正的可决系数为 A-R^2=0.9236 ，这说明模型样本拟合的非常好。

F 检验：针对 H0：β2=β3=β4=0，给定显著性水平α=0.05，在 F 分布表中查出自由度为 k-1=3 和 n-k=14 的临界值 Fα(3,14)=3.34。由表 3 中得到 F=73.5>Fα(3,21)=3.075，应拒绝原假设 H0：β2=β3=β4=0，说明回归方程显著，即“国民总收入”、“居民消费价格指数增长

率”、“人均 GDP”等变量联合起来确实对“人口自然增长率”有显著影响。

T 检验：分别针对 H0：βj=0（j=1，2，3，4），给定显著性水平 a=0.05，查看 t 分布表得自由度为 n-k=14 临界值 tα/2(n-k)=2.145。

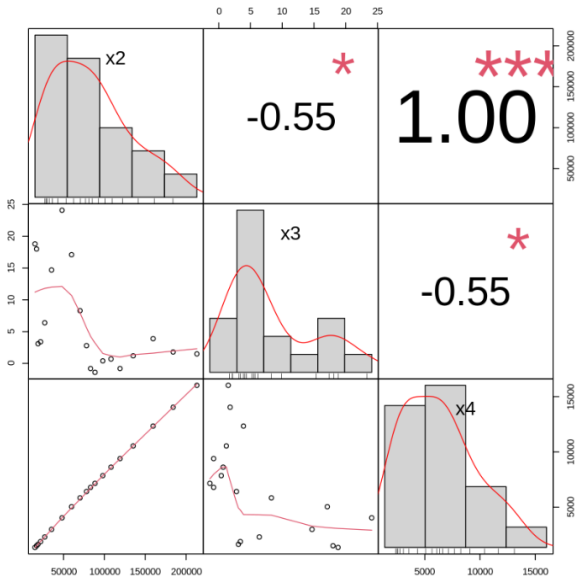
由表 3 中的数据可以得到，与β1，β2，β3，β4 对应的 t 统计量分别为 17.08010、2.482857、1.412721、-2.884953。除了β3，其绝对值均大于 tα/2(n-k)=2.145，这说明分别都应当拒绝 H0，也就是说，当在其他解释变量不变的情况下，解释变量“国民总收入”、“人均 GDP”分别对被解释变量“人口自然增长率”Y 都有显著的影响。

β3 的绝对值小于 tα/2(n-k)=2.145，这说明接受 H0，X3 系数对 t 检验不显著，这说明又存在多重共线性的可能。

所以计算各解释变量的相关系数，选择 X2、X3、X4 数据，使用 cor(dat)来获取相关系数矩阵；使用 PerformanceAnalytics 包中的 chart.Correlation(data[,-1], histogram=TRUE, pch=19)函数来得到相关系数矩阵的可视化数据。

A matrix: 4 × 4 of type dbl

	y	x2	x3	x4
y	1.0000000	-0.9193826	0.6323100	-0.9287967
x2	-0.9193826	1.0000000	-0.5461875	0.9993883
x3	0.6323100	-0.5461875	1.0000000	-0.5519322
x4	-0.9287967	0.9993883	-0.5519322	1.0000000



（表 4）

由相关系数矩阵可以看出：各解释变量相互之间的相关系数较高，由此可知确实存在多

重共线性的问题。

六、消除多重共线性

采用逐步回归的方法，去解决多重共线性问题，分别做 y 对 X2，X3，X4 的一元回归

Y 对 x2 的一元回归

```
Call:
lm(formula = y ~ x2)

Coefficients:
(Intercept)           x2
  1.401e+01   -5.145e-05
```

```
Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4976 -1.0325 -0.3228  0.7200  2.4956

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.401e+01  5.388e-01  26.000 3.96e-15 ***
x2          -5.145e-05  5.339e-06  -9.637 2.66e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.331 on 17 degrees of freedom
Multiple R-squared:  0.8453,    Adjusted R-squared:  0.8362
F-statistic: 92.86 on 1 and 17 DF,  p-value: 2.656e-08
```

Y 对 x3 的一元回归

```
Call:
lm(formula = y ~ x3)

Coefficients:
(Intercept)           x3
     8.0310       0.2621
```

```
Call:
lm(formula = y ~ x3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1832 -2.1492 -0.4339  1.6051  5.5465

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.03098    0.78533  10.226 1.11e-08 ***
x3           0.26211    0.07789   3.365  0.00367 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.621 on 17 degrees of freedom
Multiple R-squared:  0.3998,    Adjusted R-squared:  0.3645
F-statistic: 11.32 on 1 and 17 DF,  p-value: 0.003674
```

Y 对 x4 的一元回归

```
Call:
lm(formula = y ~ x4)

Coefficients:
(Intercept)           x4
 14.3366203   -0.0006953
```

```
Call:
lm(formula = y ~ x4)

Residuals:
   Min     1Q Median     3Q    Max
-1.391 -1.063 -0.278  0.692  2.343

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.434e+01  5.305e-01   27.02 2.08e-15 ***
x4          -6.953e-04  6.729e-05  -10.33 9.55e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.254 on 17 degrees of freedom
Multiple R-squared:  0.8627,    Adjusted R-squared:  0.8546
F-statistic: 106.8 on 1 and 17 DF,  p-value: 9.55e-09
```

按 R^2 的大小排序为 X4、X2、X3

以 X2 为基础，顺次加入其他变量逐步回归。首先加入 X4 的结果为：

```
Call:
lm(formula = y ~ x2 + x4)

Coefficients:
(Intercept)           x2           x4
 16.5317284    0.0004048   -0.0061066
```

```
Call:
lm(formula = y ~ x2 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7229 -0.7503  0.0721  0.6010  1.4532

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.5317284  0.7106604  23.262  9.2e-14 ***
x2           0.0004048  0.0001084   3.736  0.00180 **
x4          -0.0061066  0.0014495  -4.213  0.00066 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9447 on 16 degrees of freedom
Multiple R-squared:  0.9266,    Adjusted R-squared:  0.9175
F-statistic: 101.1 on 2 and 16 DF,  p-value: 8.387e-10
```

Y=16.5317284+020004048 **X2** – 0.0061066 **X4**
T=（3.736）　　（-4.213）
R^2=0.9266

当取α=0.05 时，tα/2(n-k)=t0.025(18-3)=2.131，X2 参数的 t 检验显著，加入 X3 回归得

```
Call:
lm(formula = y ~ x2 + x3 + x4)

Coefficients:
(Intercept)           x2           x3           x4
 15.7197750    0.0003751    0.0497390   -0.0056601
```

```
Call:
lm(formula = y ~ x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4613 -0.6229 -0.0797  0.7153  1.2592

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.7197750  0.8702058  18.064 1.37e-11 ***
x2           0.0003751  0.0001061   3.535  0.00300 **
x3           0.0497390  0.0329629   1.509  0.15209
x4          -0.0056601  0.0014259  -3.969  0.00123 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9091 on 15 degrees of freedom
Multiple R-squared:  0.9363,    Adjusted R-squared:  0.9236
F-statistic:  73.5 on 3 and 15 DF,  p-value: 3.379e-09
```

Y=15.7197750 + 0.0003751 **X2** + 0.0497390 **X3** – 0.0056601 **X4**

T=(18.064)　(3.535)　(1.509)　(-3.969)

R^2=0.9363　A-R^2=0.9236

F = 73.5

当α=0.05 时，tα/2(18-4)=2.145，X3 的参数的 t 检验不显著，予以剔除

所以 Y=16.5317284+020004048 **X2** – 0.0061066 **X4** 时消除多重共线性之后的结果。

七、结论

因此根据研究模型得到结果，可知人口自然增长率确实受到这些研究变量的影响，在其他变量不变化的情况下，当国民总收入增长 1 亿元，人口增长率增长 0.0003751%；在其他变量不变化的情况下，当人均 GDP 每增加 1 元，人口自然增长率就会下降 0.0056601%。

八、参考文献及项目说明

参考文献：

开源项目 https://github.com/u6141461/Regression-model 中的多元线性回归部分的文章

部分 CSDN 对 R 语言的代码说明

项目说明:

本次项目的所有文件信息均托管于 Github。

项目地址：https://github.com/13roky/The_study_of_natural_population_growth_rates

可使用以下命令直接克隆所有项目

git clone git@github.com:13roky/The_study_of_natural_population_growth_rates.git