HDI 与国民生活状况关系的考察

问题背景

个人的发展与国家的发展息息相关. 一个国家的发展状况可以用 Social Progress Index 来描述. SPI 由众多指标决定, 包括营养, 安全, 公共卫生, 教育, 自由, 人权等, 是当下情况和发展预期的综合考虑. 本文想要考察, 国民的幸福感在多大程度上与 SPI 有关. 同时想要对比非参方法和参数方法, 体会非参方法的意义和优势.

数据集与变量意义描述.

数据集样本量为 205, 包括这些变量: country, SPI, HDIrank, HDIindex, HDI_cat, happiness, gendereq, infantmort, birth_MF, sixty_MF,logGDP. 变量意义如下:

- country: 国家.
- SPI: Social Progress Index, 在 0 到 100 间, 意义已描述.
- HDIindex: Human Development Index, 是人的一生发展的综合考量, 包括出生时的预期寿命, 健康, 在 25 岁前受学校教育的时长, 入学年纪, 生活水平, 平均收入, 机遇等. 在 0 到 1 之间.
- HDIrank: HDIindex 的排名.
- HDI_cat: 根据 HDI 的大小, 把 HDI 分为 4 个类别, 从低到高分别是 “Low”, “Medium”, “High”, “Very High”,
- gendereq: 0 到 1, 包括不同性别经济参与度, 政治参与度, 受教育情况, 健康, 寿命.
- infantmort: 婴儿死亡率.
- sixty_MF: 已经活到 60 岁的人的预期寿命. 主要与医疗水平和经济水平有关. birth_MF: 婴儿出生时的预期寿命. 相比 sixty_MF, 与安全, 稳定的环境也有明显关系.

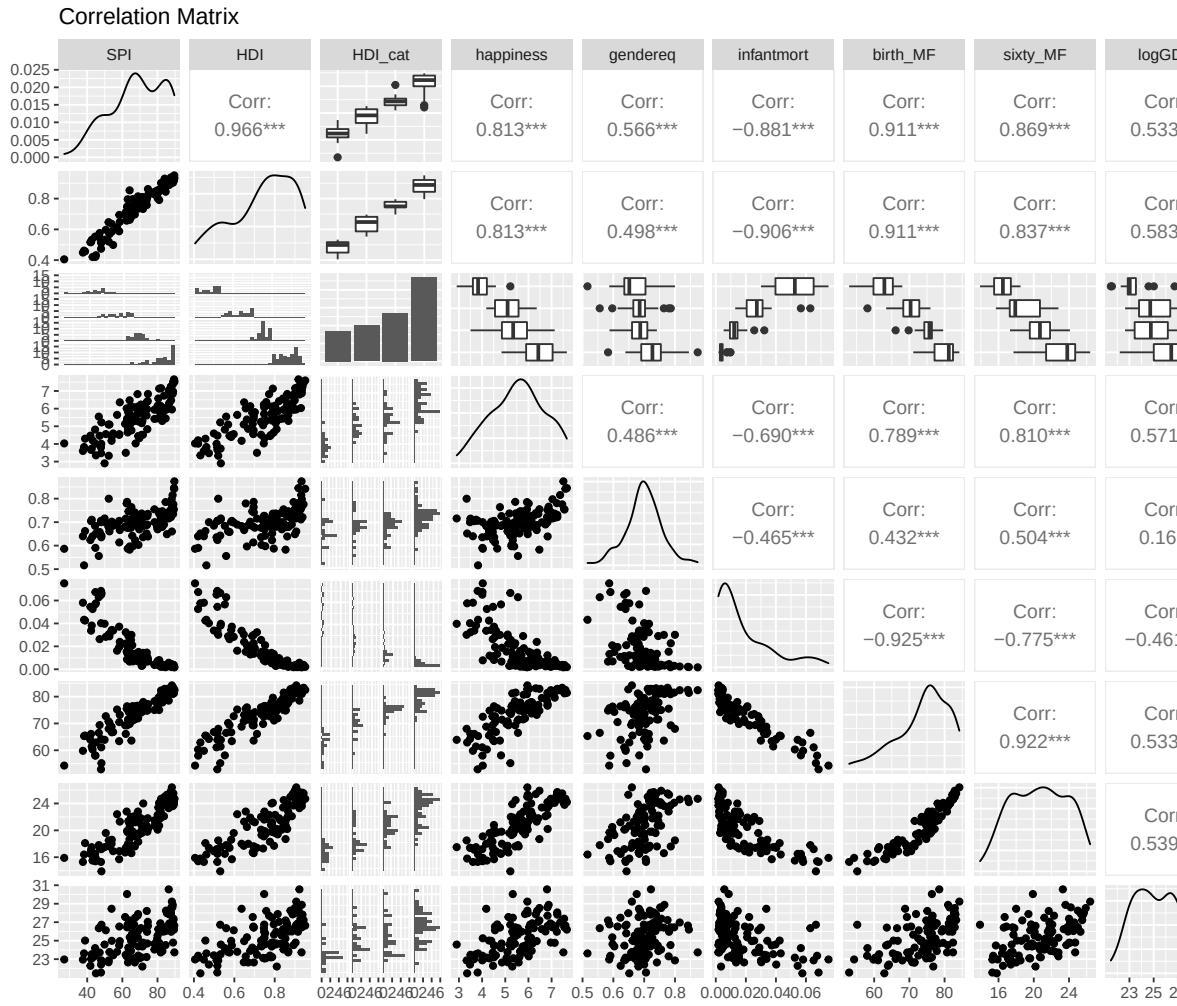
预处理

含有 NA 的有 94 个观测, 总观测数是 205 个. 去除含有 NA 的观测.

```
## [1] 205
```

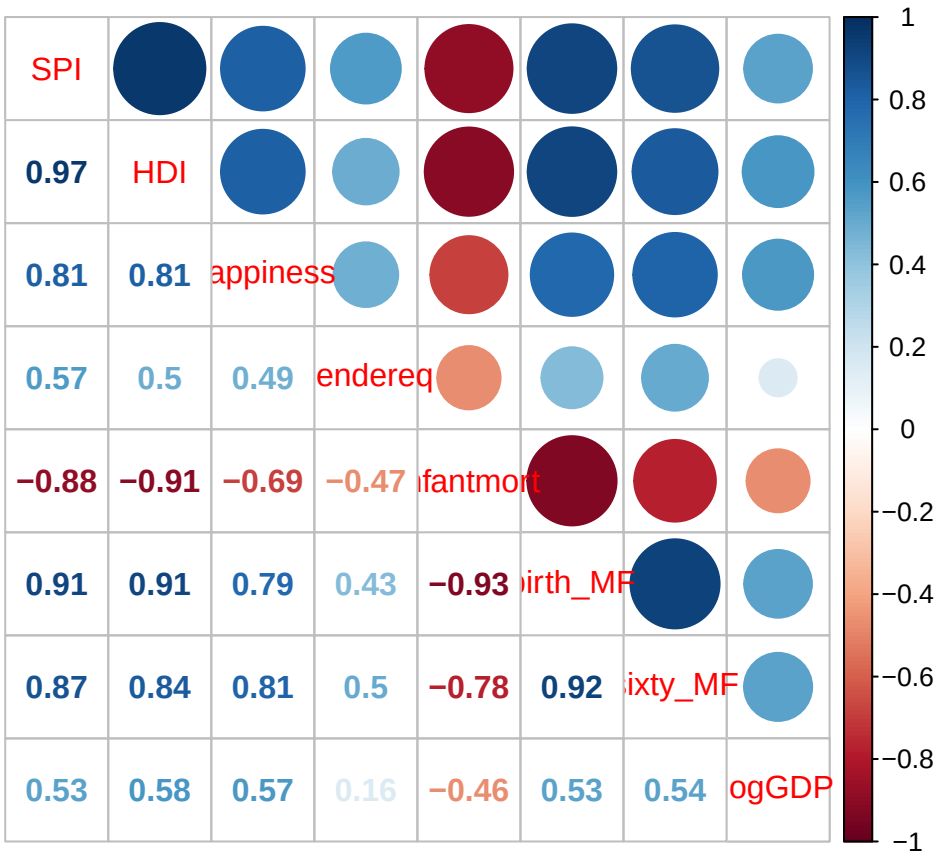
```
## [1] 94
```

EDA



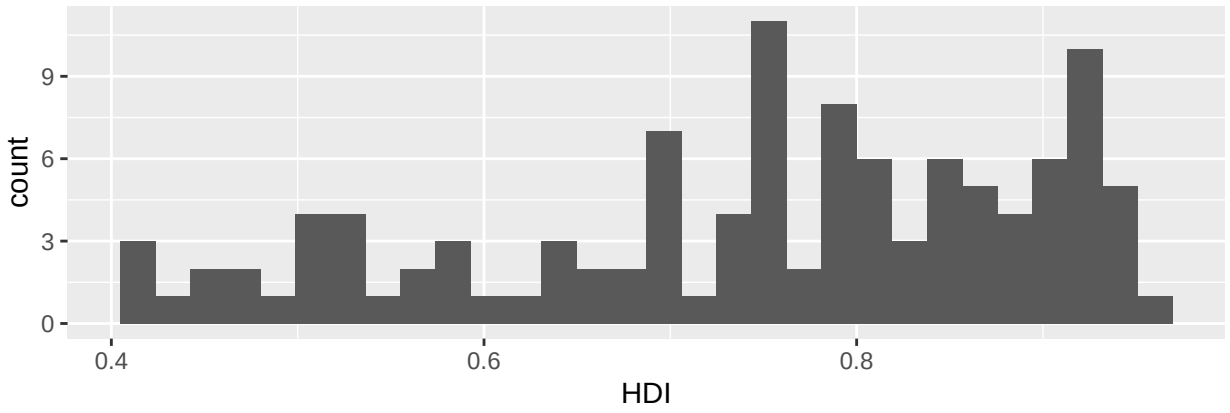
单个变量的分布情况

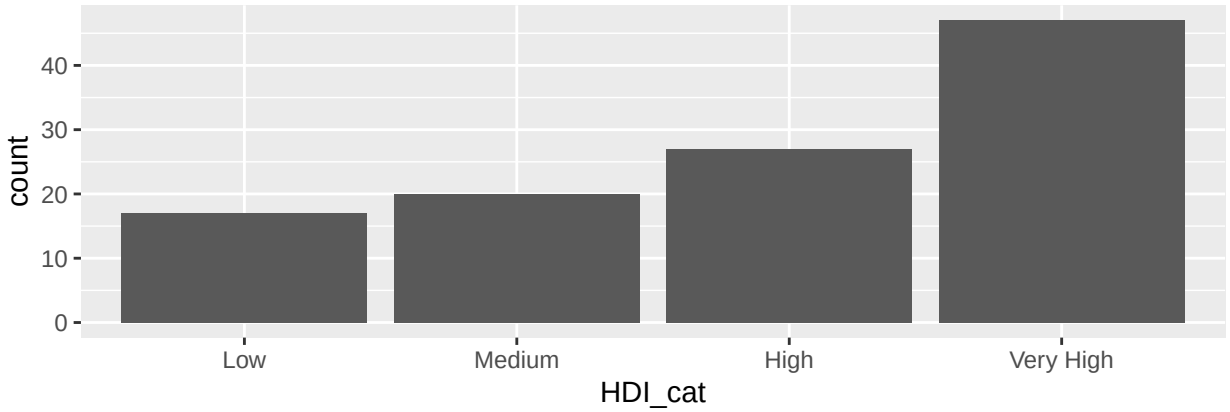
多数变量明显不满足正态性, SPI, HDIindex, infantmort, birth_MF 明显不具有对称性. sixty_MF, logGDP 则具有多峰的性质. 提示多数分析下, 参数方法的前提假设正态性并不成立, 非参的方法更加适用.



变量之间的线性关系

SPI 与 HDIindex 具有非常强的线性关系 (0.97), 说明两个指标具有很大的相似性. SPI 与 happiness,infantmort,birth_MF,sixty_MF 均具有 0.8 以上或者-0.8 以下的相关系数, 但与 gendereq 和 logGDP 的线性关系相对较低 (0.5 到 0.6), 提示 GDP 并不是 SPI 的主要考虑因素. 除去 SPI 与 HDIindex, 其它变量之间普遍也有明显的线性关系 (绝对值 0.5 以上), 线性关系不明显的是 logGDP 和 gendereq(0.16), 具体是否有关系, 后文会继续考察. happiness 与 birth_MF, sixty_MF 具有很强的线性关系 (0.8 左右). infantmort, sixty_MF, birth_MF 这三个变量间彼此具有很强的线性关系, 这符合常识, 它们都是医疗水平和营养水平, 公共卫生状况的反应. infantmort 与其它所有变量都有不同程度的负相关关系, 最低的是 logGDP, 也达到-0.46. 说明婴儿死亡率就已经能体现出一个国家的很多方面. #### HDI

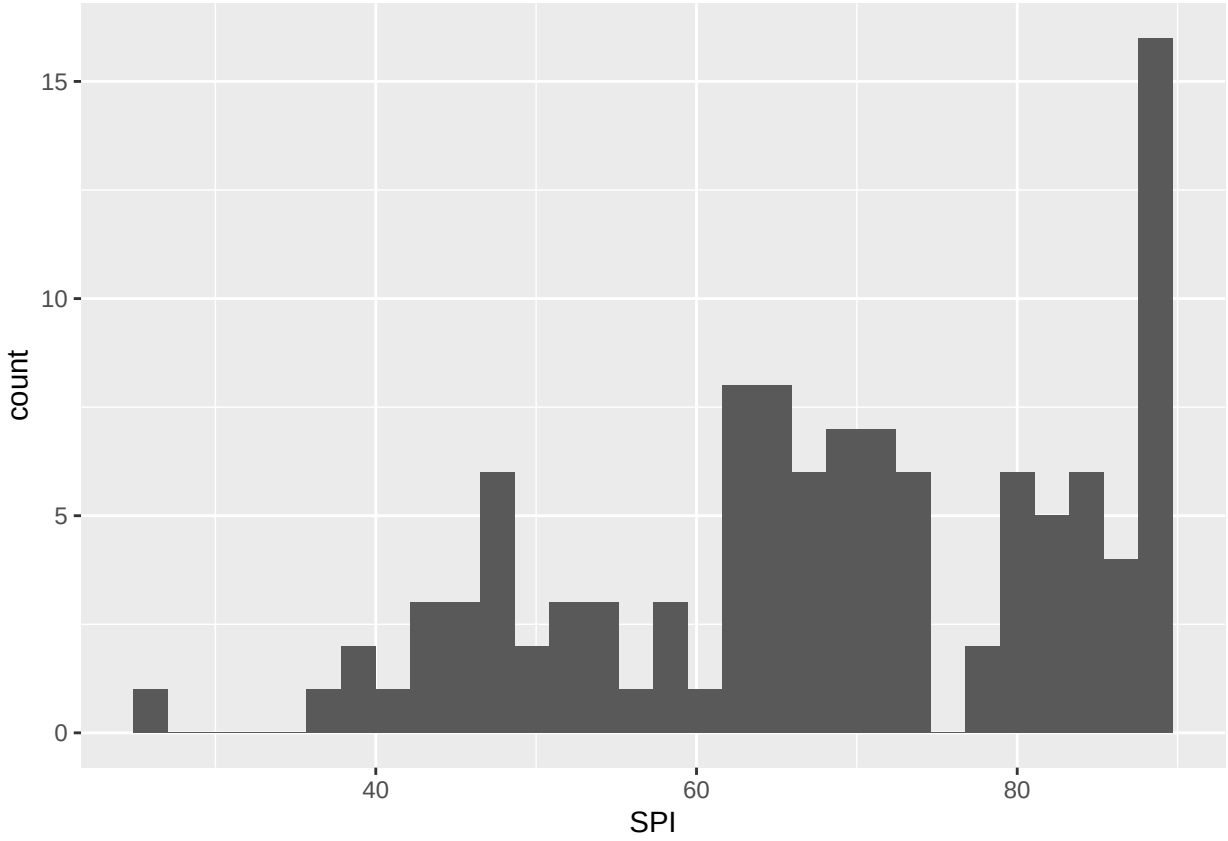





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4053  0.6527  0.7722  0.7446  0.8673  0.9512
```

```
## [1] 0.1499782
```

可以看到 HDI 分布不对称, HDI_cat 从低到高数量逐渐增多. 中位数大于均值也体现了这一点.



SPI

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.92   58.42   68.94   68.62   82.37   89.62
```

```
## [1] 15.38066
```

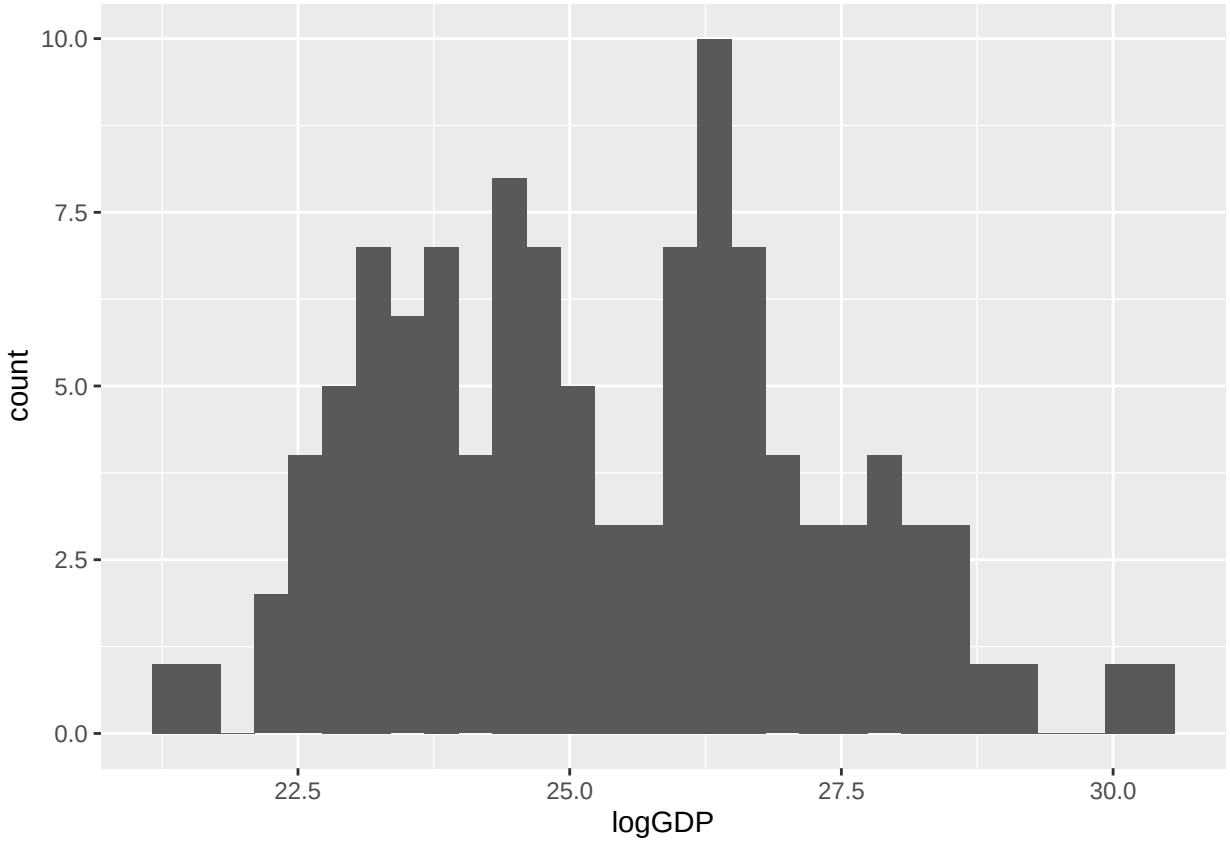
SPI 呈现出双峰的特点, 频数最多的是最高得分. 中位数和均值非常接近.

```
## [1] "Australia"   "Germany"     "Ireland"     "Norway"      "Switzerland"
```

```
## [1] "Denmark"     "Finland"     "Iceland"     "Netherlands" "Norway"
```

SPI 最高的 5 个国家与 HDI 最高的 5 个国家, 重合的有 Norway, Iceland. SPI 最高的 5 个国家全部是北欧国家.

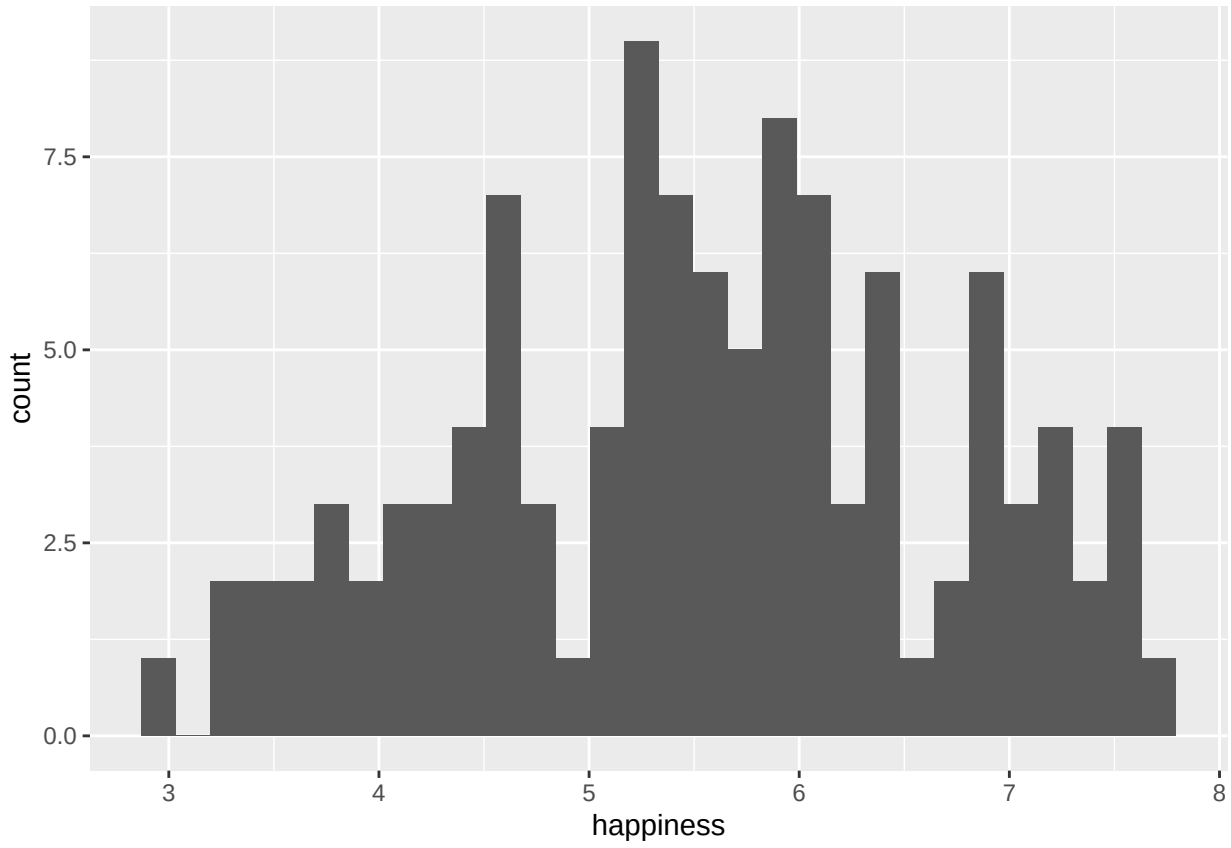
GDP



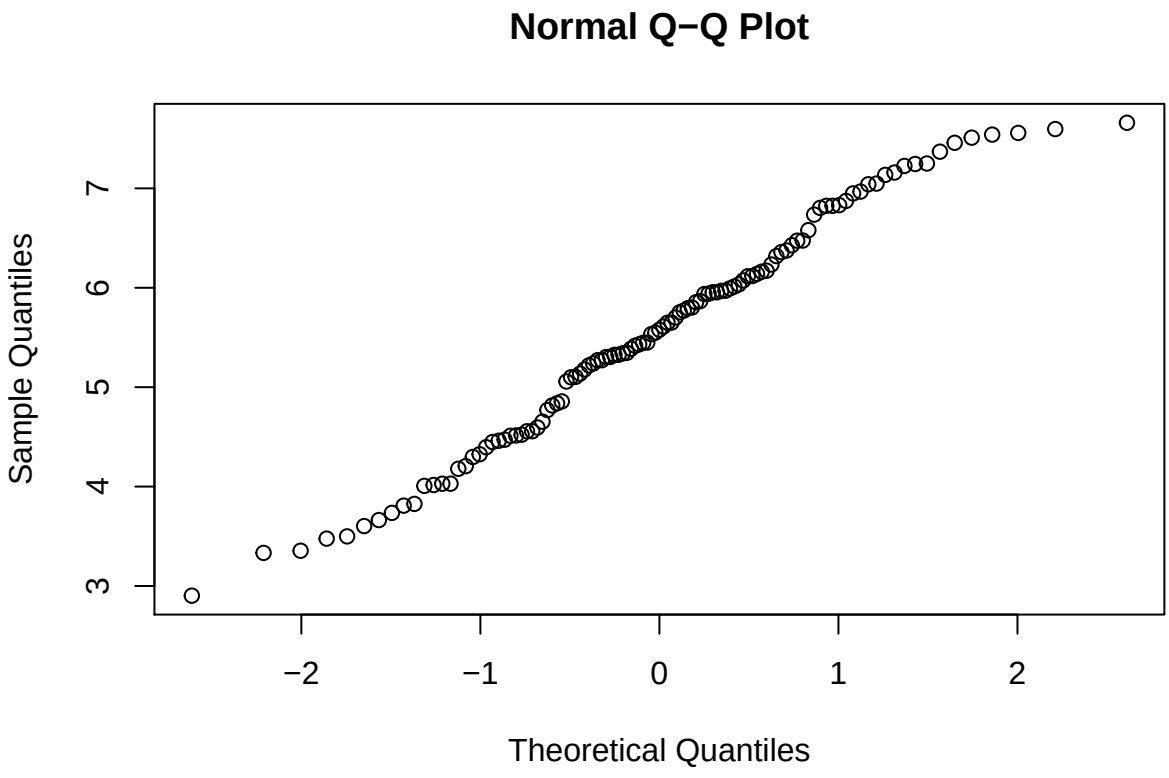

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.47   23.73   25.01   25.32   26.67   30.56
```

```
## [1] 1.941864
```

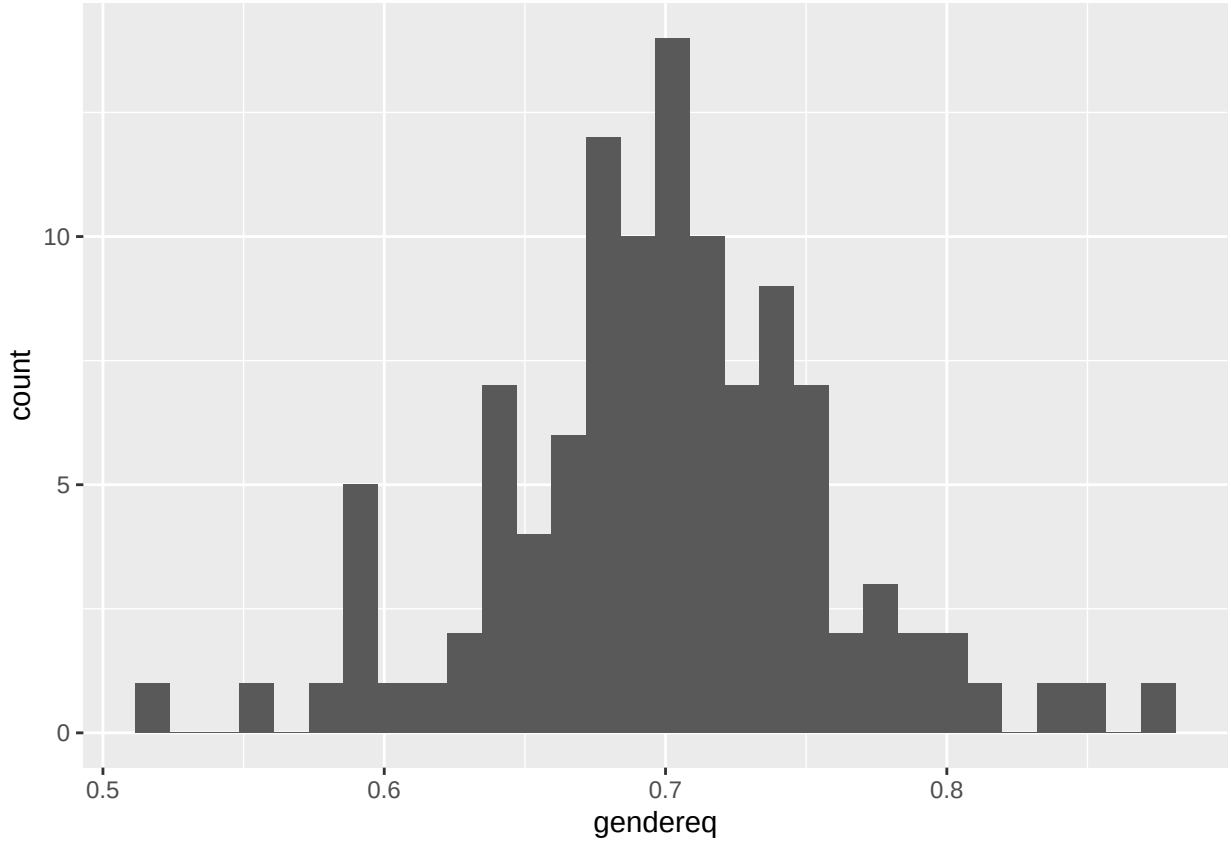
logGDP 呈现出双峰的特点.



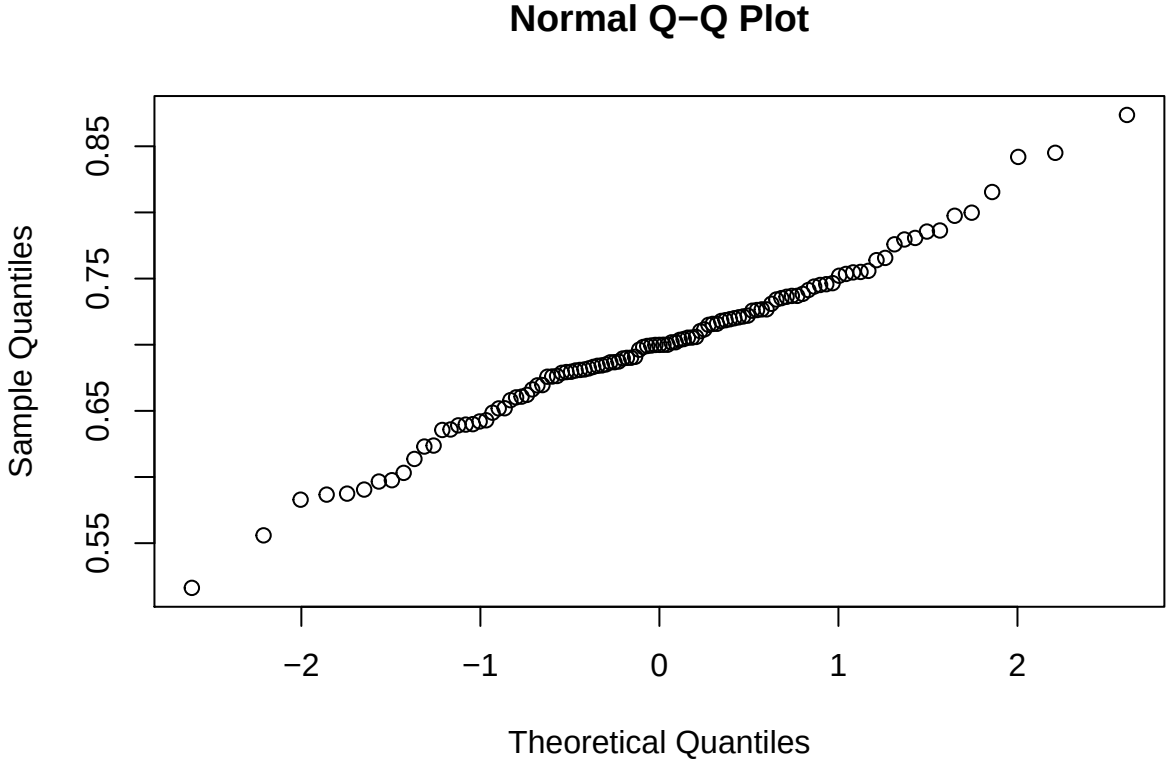
happiness



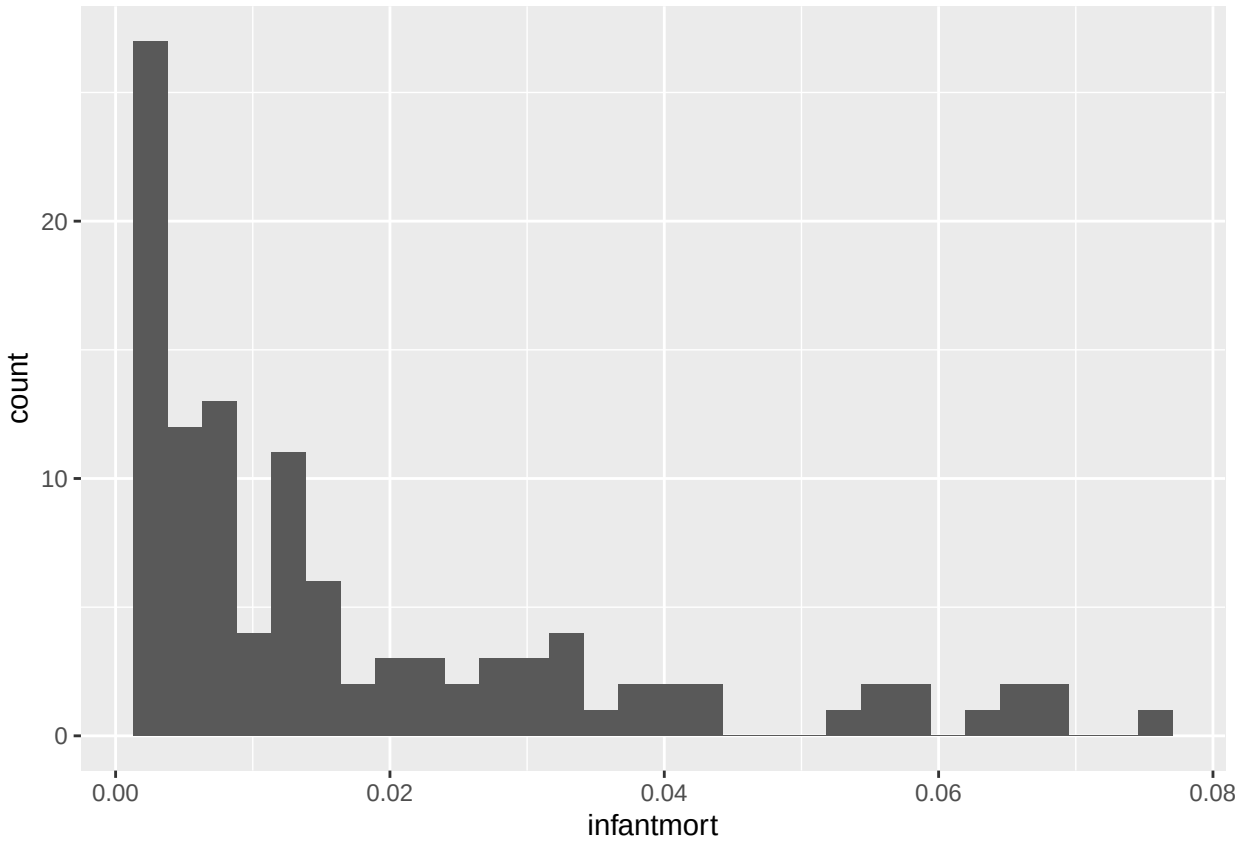
happiness 大致有对称性和正态性.



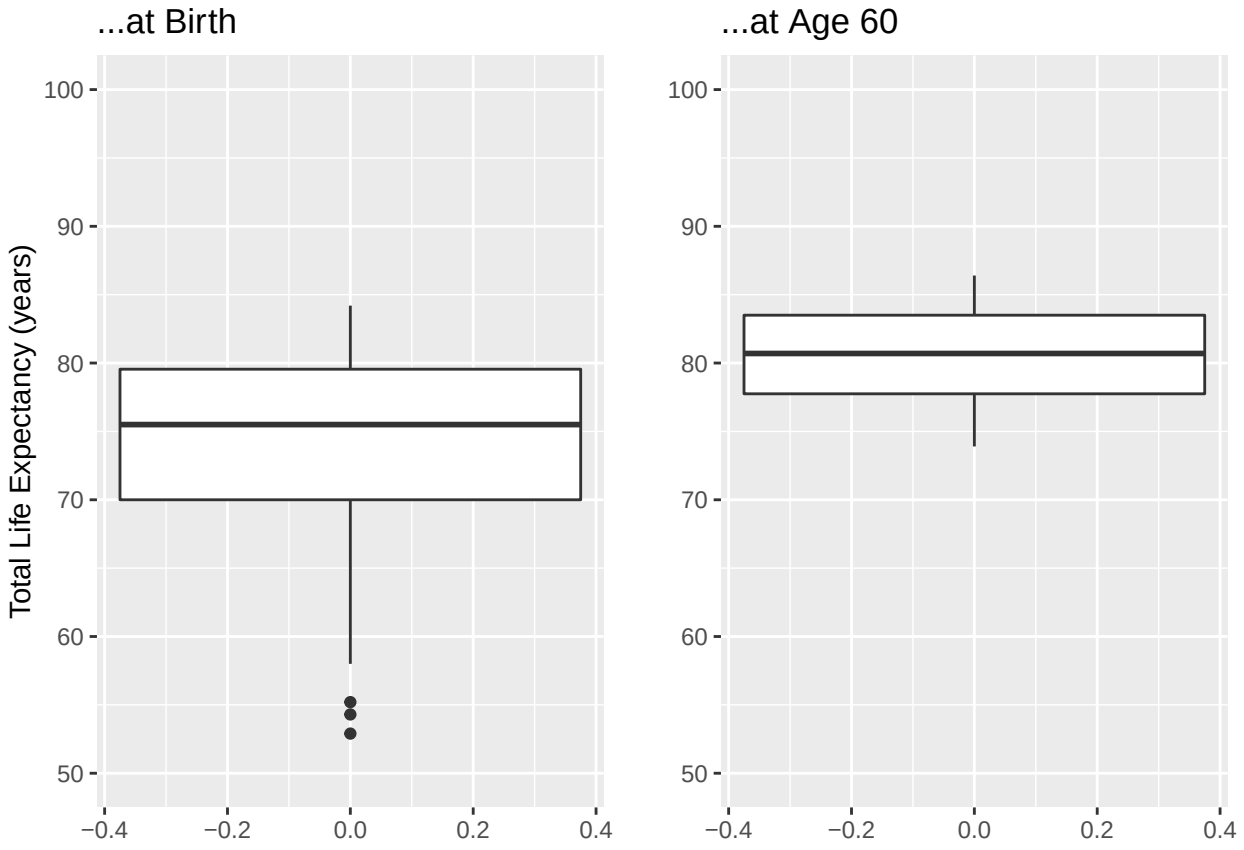
gendereq



gendereq 也大致具有对称性和正态性.



infantmort 明显右偏.




```
## [1] 6.703604
```

可以看到, 60 岁的期望寿命比出生时的期望寿命高 6.7 岁, 因为前者已经是条件期望. sixty_MF 具有良好的对称性. birth_MF 则左偏.

非参方法与参数方法的对比

HDI 的区间估计

```
##
##  Shapiro-Wilk normality test
##
## data:  .data$HDI
## W = 0.93074, p-value = 2.201e-05
```

正态性检验说明 HDI 不服从正态分布. 用 bootstrapping 给出区间估计. 4 种区间估计都很接近.

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_obj, type = "all")
```

```
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.7172,  0.7720 )   ( 0.7166,  0.7717 )
##
## Level     Percentile            BCa
## 95%   ( 0.7174,  0.7725 )   ( 0.7175,  0.7725 )
## Calculations and Intervals on Original Scale
```

```
##
##  One Sample t-test
##
## data:  .data$HDI
## t = 52.303, df = 110, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.7163408 0.7727629
## sample estimates:
## mean of x
## 0.7445518
```

t 区间为 (0.7163408,0.7727629). 尽管不满足正态性, 参数估计给出的 t 区间估计也是很接近的.

性别平等与经济水平毫无关系?

在 corrplot 中, 我们看到 gendereq 与 logGDP 的相关系数仅为 0.16, 那么两者是否不存在关系?

```
## An object of class "testforDEP_result"
## Slot "TS":
## [1] 0.009986915
##
## Slot "p_value":
## [1] 0.0369963
##
## Slot "CI":
## list()
```

Hoeffding's Test 给出的 p 值为 0.03949605, 表明 logGDP 与 gendereq 存在关系, 但不一定是线性关系.

```
## [[1]]
## [1] "Permutation correlation test.  Method is pearson"
##
## [[2]]
```

```
## [1] "p-value was estimated based on 20000 simulations."
##
## $alternative
## [1] "two.sided"
##
## $p.value
## [1] 0.0939
```

用 permutation test, p 值为 0.09315, 显著性水平为 0.05 时无法拒绝原假设, 说明两者存在线性关系的证据不足. 综合两个检验可以得出的结论是, logGDP 与 gendereq 明显存在关系, 但不是线性关系.

Shapiro-Wilk test 与 Lilliefors test 的对比

Shapiro-Wilk test 是参数方法, 用于检验正态性应当比 Shapiro-Wilk test 是参数方法, 用于检验正态性应当比 Lilliefors test 更有效. 对 happiness 和 gendereq 考察两种检验的效果.

```
##
##  Shapiro-Wilk normality test
##
## data:  .data$happiness
## W = 0.98157, p-value = 0.1289

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  .data$happiness
## D = 0.052943, p-value = 0.6239
```

对于 happiness, 两种检验都无法拒绝原假设, 说明 happiness 具有正态性. 但 shapiro.test 给出的 p 值要小很多.

```
##
##  Shapiro-Wilk normality test
##
## data:  .data$gendereq
## W = 0.98546, p-value = 0.2732

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  .data$gendereq
## D = 0.087421, p-value = 0.03634
```

但对于 gendereq, 与 happiness 中两种检验的 p 值对比关系正好相反, 0.05 的显著性水平下, lillie.test

拒绝原假设, 而 shapiro.test 无法拒绝原假设. 说明, 参数方数 Shapiro-Wilk 的 p 值并不总是小于非参方法 lillie.test, 两种检验都需要做.

infantmort 在不同的 HDI 水平下没有明显差别?

非参方法使用抽样做 permutation test, 检验统计量是各组的均值与总的均值的平方和.

```
## [1] 0
```

在抽样 2000 次的情况下, p 值为 0. 即没有出现任何一个样本比当前的统计量更大. 说明应该拒绝原假设, infantmort 在 HDI 不同的国家有很明显的差别.

```
##        Df         F value           Pr(>F)
##  Min.   :  3   Min.   :0.9225   Min.   :0.4327
##  1st Qu.: 29   1st Qu.:0.9225   1st Qu.:0.4327
##  Median : 55   Median :0.9225   Median :0.4327
##  Mean   : 55   Mean   :0.9225   Mean   :0.4327
##  3rd Qu.: 81   3rd Qu.:0.9225   3rd Qu.:0.4327
##  Max.   :107   Max.   :0.9225   Max.   :0.4327
##                NA's   :1        NA's   :1
```

在试图使用参数方法 ANOVA 之前, 先进行等方差检验, p 值为 0.4327, 无法拒绝原假设.

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  .data$infantmort
## D = 0.20109, p-value = 5.856e-12
```

但 infantmort 的正态性是严重违背的. infantmort 不具有正态性.

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## log(infantmort)   1 113.40  113.40   584.3 <2e-16 ***
## Residuals       109  21.16    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA 同样给出了拒绝原假设的结论, 尽管它的前提条件正态性并不满足.

happiness 与 HDI_cat 是否有明显区别?

先以 Medium 和 High 对比, 参数方法为两样本的 t 检验, 先检查等方差性. p 值为 0.1371, 无法拒绝原假设, 因此 t 检验的等方差假设可以认为是满足的. 并且由前面的检验可知, happiness 的正态性也是满足的.

```
##        Df         F value          Pr(>F)
##  Min.   :  3   Min.   :1.882   Min.   :0.1371
##  1st Qu.: 29   1st Qu.:1.882   1st Qu.:0.1371
##  Median : 55   Median :1.882   Median :0.1371
##  Mean   : 55   Mean   :1.882   Mean   :0.1371
##  3rd Qu.: 81   3rd Qu.:1.882   3rd Qu.:0.1371
##  Max.   :107   Max.   :1.882   Max.   :0.1371
##                NA's   :1       NA's   :1
```

```
##
##  Welch Two Sample t-test
##
## data:  happiness.med and happiness.high
## t = -1.1683, df = 44.989, p-value = 0.2489
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6839918  0.1818027
## sample estimates:
## mean of x mean of y
##  5.137048  5.388142
```

t 检验中, p 值为 0.2489, 无法拒绝原假设.

```
##
##  Wilcoxon rank sum exact test
##
## data:  happiness.med$happiness and happiness.high$happiness
## W = 219, p-value = 0.2802
## alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon 秩和检验的 p 值为 0.2802, 无法拒绝原假设. 即 HDI_cat 为 “Medium” 和 “High”, 幸福感没有显著差异.

```
##
##  Wilcoxon rank sum exact test
##
## data:  happiness.low$happiness and happiness.veryhigh$happiness
## W = 2, p-value = 5.8e-15
## alternative hypothesis: true location shift is not equal to 0
```

但对于 “Low” 与 “Very High”, Wilcoxon 秩和检验的 p 值为 5.8e-15. 说明 HDI_cat 与 happiness 有明显关系, 那么多大程度上有关系? 我们进行两两比较, 并且用 Bonferroni 进行修正.

```
## [[1]]
```

```
## [1] 7.181223e-07
##
## [[2]]
## [1] 9.963957e-08
##
## [[3]]
## [1] 5.799748e-15
##
## [[4]]
## [1] 0.2801743
##
## [[5]]
## [1] 3.167979e-08
##
## [[6]]
## [1] 3.101678e-06
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    1    2    2    3
## [2,]    2    3    4    3    4    4
```

因此, 一开始比较 “Medium” 与 “High” 的结果具有欺骗性. 仅有这两组没有显著差异, 其它所有组两两检验的 p 值都非常小, 远小于 Bonferroni 修正后的结果.

对 “Medium” 与 “High” 求均值, 发现两者的均值差距本来就不大, 与其它各组之间的差距小得多. 这部分解释了为什么仅 “Medium” 与 “High” 两组没有明显差异.

```
##       Low    Medium      High Very High
##  3.914385  5.137048  5.388142  6.417831
```

这说明 HDI_cat 很大程度上决定了国民幸福感和婴儿死亡率.

总结

本数据集的特点有 2 个, 样本量不大并且绝大多数变量没有正态性. 因此用非参方法更合适. 不过 t 检验, t 区间估计, ANOVA 等, 在样本量比较大的情况下 (本数据集去除含有 NA 的观测后样本量为 111), 也得出了与非参方法很接近的结论, 这说明在样本量大的情况下, 即使部分条件不成立, 参数方法的结果也还是可靠的.