申请海外研究生录取率影响因素的探究

2018010720 李一笑

2020.06.07

目录

一、 研究背景及问题提出 2
二、 数据集描述 2
三、 数据探索 3
3.1 相关性 3
3.2 PCA 数据可视化 3
四、 探索性因子分析 7
4.1 因子数选择，模型稳定性 7
4.2 模型解读 8
4.3 正交因子模型 8
4.4 斜交因子模型 9
五、 回归模型预测 11
六、 总结 12
6.1 结论 12
参考资料 13
6.2 MLE 方法下，旋转为 varimax 模型输出 14
6.3 Minimum residual 方法下，旋转为 promax 模型输出 15
6.4 回归模型 16
6.5 Scree plot 17

一、研究背景及问题提出

众所周知，申请国外研究生是一个“三费”的过程——费钱、费时、费力．与普通高等学校招生全国同一考试的排名录取不同，申请海外研究生时需要考量学生的多项指标，如 GRE、TOEFL 及 GPA、推荐信、科研经历等等．当然，如果面面俱到自然轻松优胜，但是考虑到精力有限，能力有限的情况，逐点提升这些指标得分是困难的．于是我们考虑这样几个问题，这些看似各异的指标之间是否存在内在的联系，是否反映了学生不同层面的能力？如果存在这样的层次，学生需要侧重这些方面来综合提升指标得分．同时，就单个指标而言，学生应该如何投入时间精力，才能对研究生录取率提升有最大帮助？

基于以上的考量，我们提出下列问题：

1. 是否存在潜在的因素 (latent variables)，影响学术在留学申请几项指标的得分？

2. 找出对录取率有显著帮助的指标．

3. 利用已知留学申请记录数据集，预测被录取的可能性．

二、数据集描述

原始数据来源于数据建模分析竞赛平台 Kaggle 上 Graduate Admission 2 项目，其包括 500 条印度学生的留学申请记录，包含变量如表1：

变量	变量说明
GRE 成绩	GRE Score，数值型，满分 340;
TOEFL 成绩	TOEFL Score，数值型，满分 120;
本科大学等级	University Rating，整数型，最高等级为 5;
申请文书	SOP，整数型，最高等级为 5;
推荐信力度	LOR，数值型，最高等级为 5;
本科阶段 GPA	CGPA，数值型，满分为 10；
研究经历	Research，逻辑型，没参与过科研为 0、参与过为 1;
被录取的可能性	Chance of Admit，数值型，0 至 1;

表 1 变量说明

数据没有缺失值，基本统计量描述如表2，由于 SOP，LOR 以及 Universities．Rating 的水平较多，我们可以将其看作连续的数值型变量，易于之后的分析．

	vars	n	mean	sd	median	trimmed	mad	min	max	range
GRE.Score	1	500.00	316.47	11.30	317.00	316.48	11.86	290.00	340.00	50.00
TOEFL.Score	2	500.00	107.19	6.08	107.00	107.09	7.41	92.00	120.00	28.00
CGPA	3	500.00	8.58	0.60	8.56	8.58	0.68	6.80	9.92	3.12
Chance.of.Admit	4	500.00	0.72	0.14	0.72	0.73	0.15	0.34	0.97	0.63
University.Rating	5	500.00	3.11	1.14	3.00	3.10	1.48	1.00	5.00	4.00
SOP	6	500.00	3.37	0.99	3.50	3.40	0.74	1.00	5.00	4.00
LOR	7	500.00	3.48	0.93	3.50	3.50	0.74	1.00	5.00	4.00
Research	8	500.00	0.56	0.50	1.00	0.57	0.00	0.00	1.00	1.00

表 2 数据的基本描述

三、数据探索

3.1 相关性

数据呈现较强的相关性（图1），注意到响应变量录取率与所有的解释变量的相关性都高于 0.64，其中与 CGPA 的相关性最强，为 0.882；同时解释变量间的相关性也较强，且均为正相关，散点图也反映了这一点

同时，从箱线图（图1），我们注意到研究经历对其他指标的得分分布有所影响，有研究经历下其他指标得分的均值高于无研究经历下得分均值.

3.2 PCA 数据可视化

为了对数据有更好的了解以及将多元数据降维后可视化，我们对取出分类变量（研究经历）的数据标准化后做主成分分析.（见注释3.2）画出 scree plot 后，注意到两个主成分即可保留大部分的方差信息，实际上 PC1,PC2 方差解释比例达到了 0.835. 我们将各变量的 PC1，PC2 的主成分系数，以及样本观测值的第一、第二主成分得分可视化. 如图3,2.

其中，颜色反映了该变量对 PC1,PC2 的相对贡献，反映了该变量与主成分的相关性；如果相关性强，则说明该变量对数据中方差的解释能力强，贡献较大；与之相对的，相关性弱，说明贡献小，可能考虑在降维和数据选择时舍去. 我们从图2中看出，本科大学等级的贡献最小，而 GRE 成绩与 CGPA 贡献较大. 这为之后的变量选择提供参考. 同时，我们注意到，根据变量的第二主成分的系数的正负（图中体现为二、三象限的区别），可以将变量分为两组，该系数正负性反映变量对 PC2 的作用不同.

在观测值主成分得分图3中，我们有 cos2 衡量主成分对单个观测的重要性，取 PC1,PC2 主成分得分，相当于将高维空间的数据点投影到一个保留最多信息的二维平面上，这种投影难免会损失信息，而 cos2 即反映了原有向量与投影后向量夹角，夹角越小，cos2 值越大，用第一、第二主成分得分表示观测点损失的信息越少，第一第二主成分对该观测就越重要[1]. 图3中可以看出，保留第一、第二主成分可以有效表示大多数

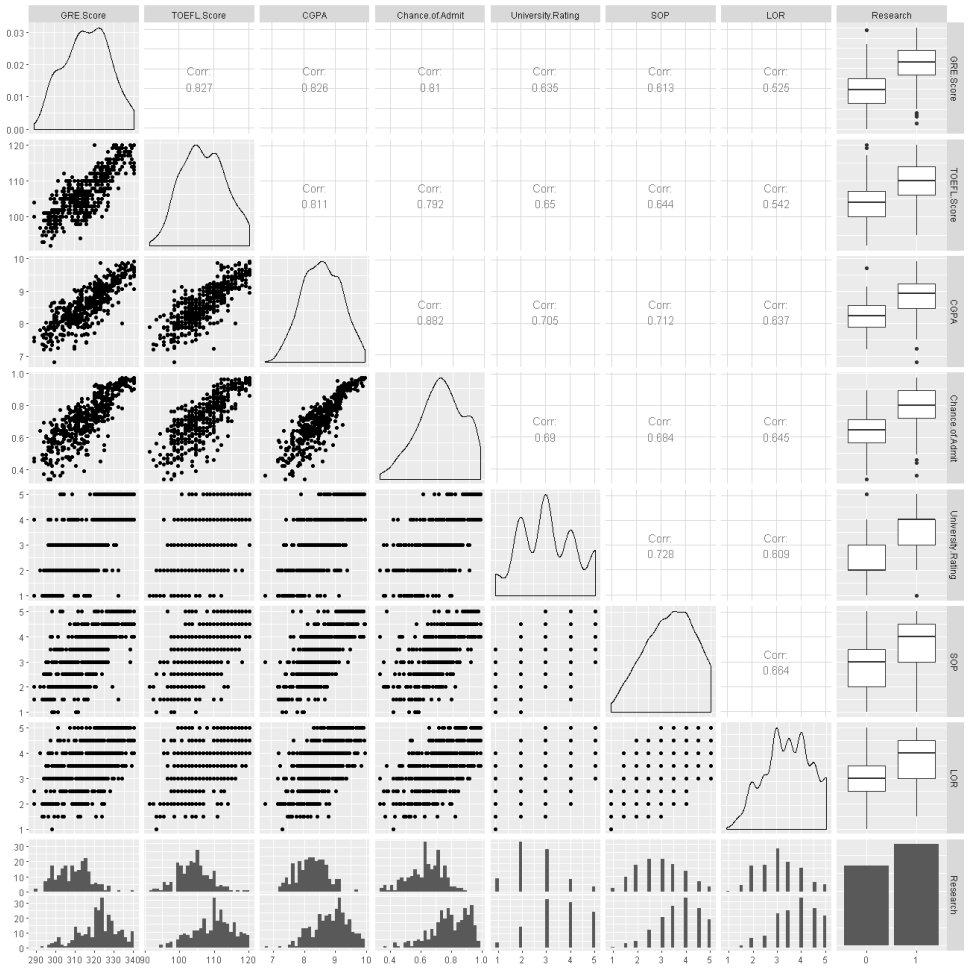

图 1　散点-分布-相关性图

观测.

同时，我们加入是否有研究经历这一变量，可以看到，该变量对观测的主成分得分有系统性的影响．很有可能是否有研究经历是一个比较显著的变量.

Remark. 在做 *PCA* 时，选择将 *0-1* 变量“研究经历”先剔除．实际上有资料指出可以将于 *0-1* 变量加入 *PCA*[2]，但我认为可解释性不好，而且我在做 *PCA* 所使用的矩阵为 *Pearson* 相关矩阵，在加入 *0-1* 变量时，*Pearson* 相关矩阵可能不如 *polychoric correlation matrix*.

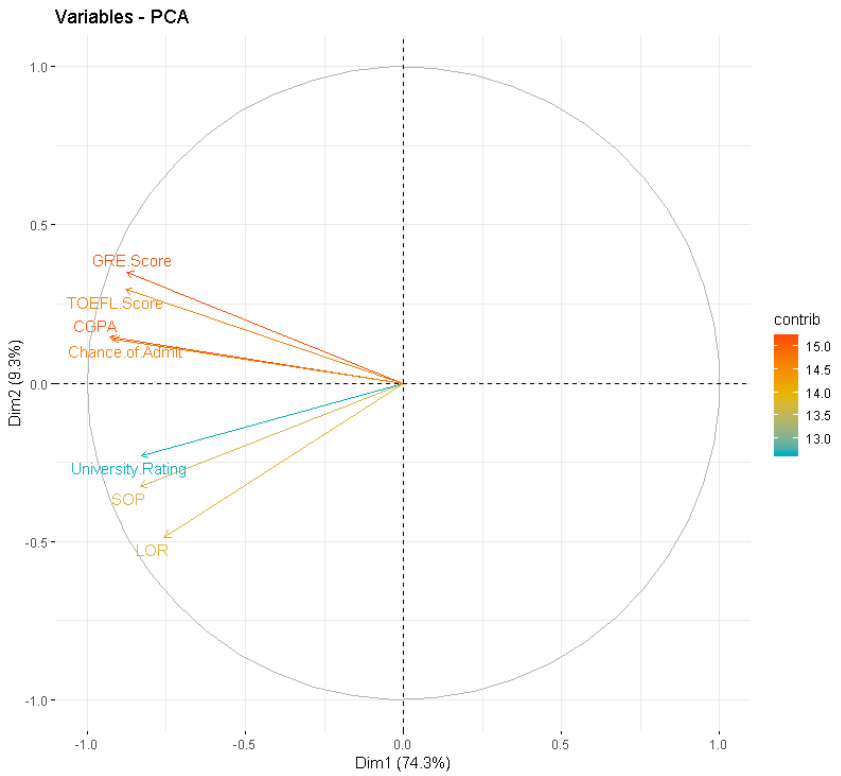

图 2 各变量的 PC1，PC2 的主成分系数

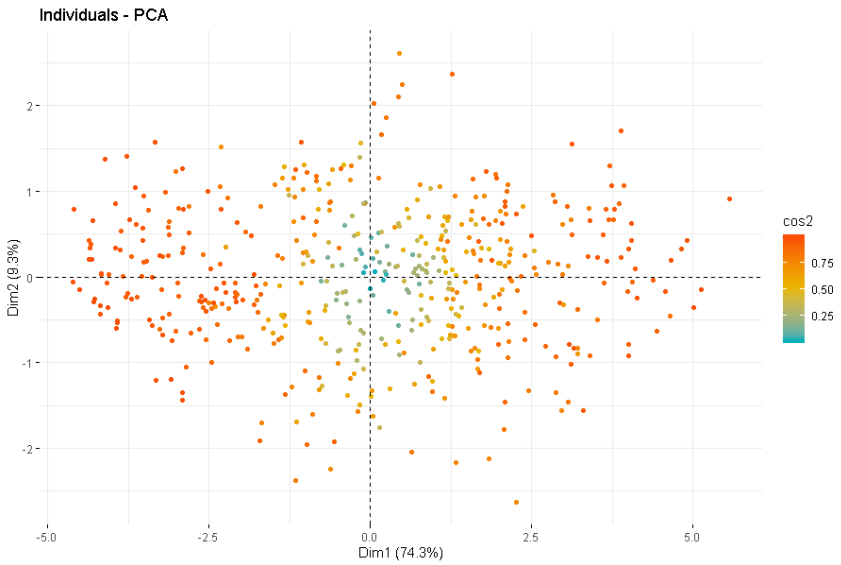

图 3 观测值第一、第二主成分得分图，颜色反映主成分对观测值的重要性

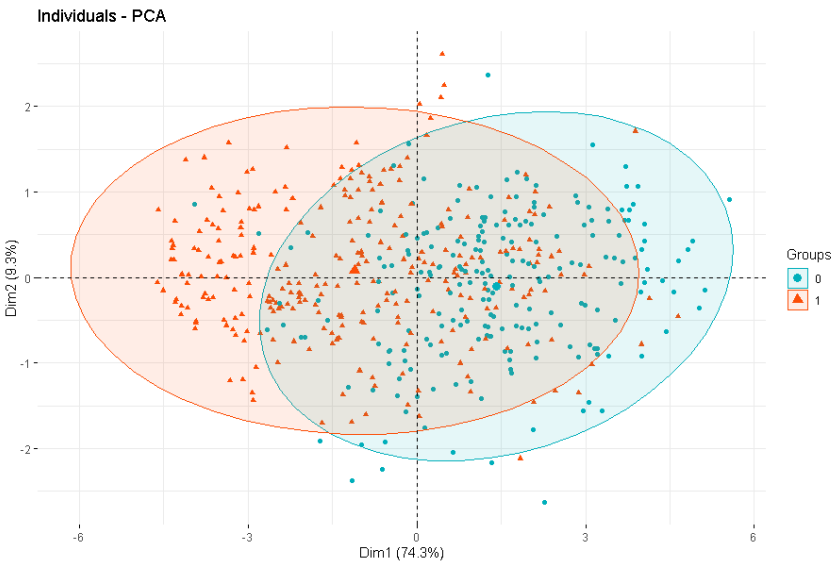

图 4 观测值第一、第二主成分得分图，颜色区分是否有研究经历

四、探索性因子分析

4.1 因子数选择，模型稳定性

在 PCA 中，我们注意到保留第一、第二主成分就可以解释 0.835 的方差，如果利用主成分方法求解因子模型，选择两个因子实际是理想的．我们选用令因子数 $m = 2, 3$ 分别利用主成分法，最大似然法 (旋转方法为 varimax) 求解，比较结果．由图5可以看出，

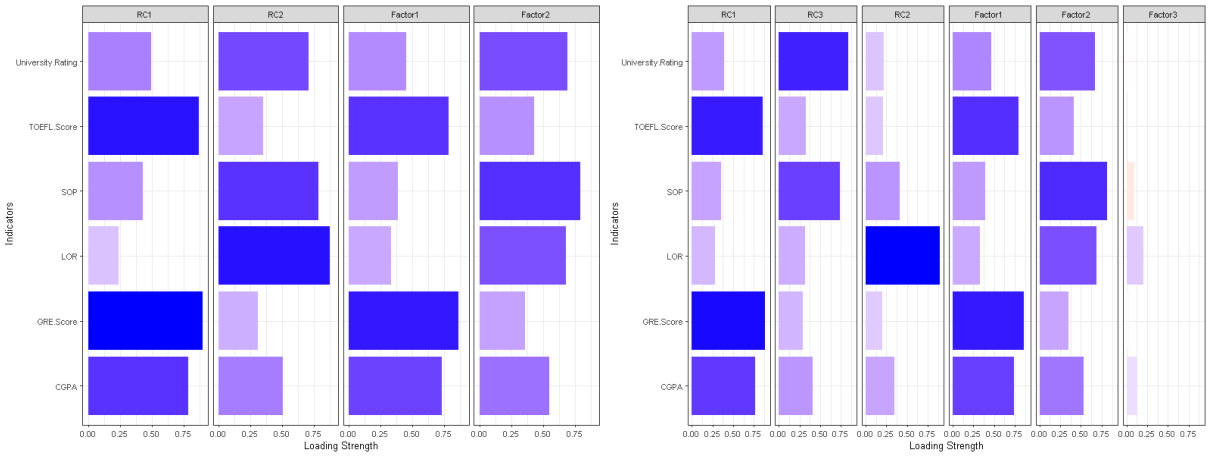

图 5 因子载荷可视化，其中 RC 为主成分得到载荷，Factor 为极大似然法得到载荷

左侧的 $m = 2$，两种方法得到结果比较相符，且载荷在两个因子上分布比较集中，因子对变量的贡献比较分明，可解释性好．而右侧的 $m = 3$，两种方法结果有一定差异，且 Factor3 载荷都比较小，解释性不好．

为了避免过拟合而产生结果的误读，我们将数据随机分为两组，分别进行两种方法的求解，将结果可视化．如图6．可以看出，两个子数据集得到的结果有比较一致的结果，利用全体数据得到的因子模型不是由于数据的特性．

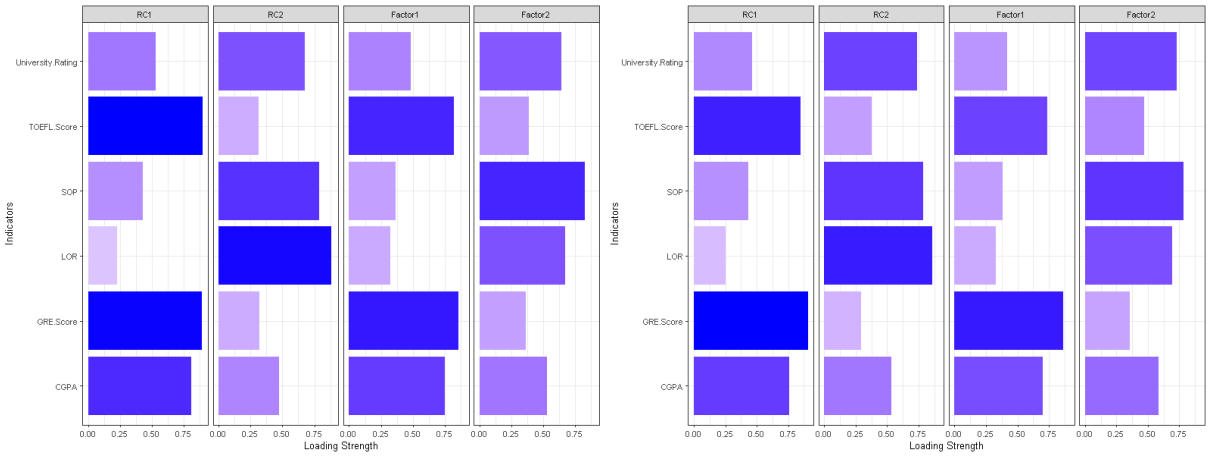

图 6 因子载荷可视化，其中 RC 为主成分得到载荷，Factor 为极大似然法得到载荷

我们最后基于样本，对模型进行假设检验，结果显示因子数为 2 是充分的．

Remark.

- 我们选择将 *0-1* 变量“研究经历”先剔除．在查阅相关资料后，与 *PCA* 的有争议的说法不同，几乎所有资料都建议不要将 *0-1* 加入 *FA*[3]，在实际操作中，将 *0-1* 变量加入后，两种方法得到的因子载荷差异较大，并且在两个子数据得到的结果差异也较大．
- 我们这里只考虑了解释变量，剔除了录取率，但是在加入录取率后，其在因子 *1* 的载荷远远大于因子 *2*，这一发现与 *PCA* 分析中的分类结果一致．

4.2 模型解读

4.3 正交因子模型

我们考察利用 mle 方法，在旋转方法为 vaimax 下得到的模型中（结果见附录），变量 GRE 成绩，TOEFL 成绩，本科阶段 GPA 在因子 1 上载荷显著，变量申请文书, 推荐信力度以及本科大学等级在因子 2 上载荷更为显著．我们将因子 1 解释为“硬实力”，其主要为学生可以经过标准量化的能力，如语言考试 GRE 以及 TOFEL 成绩；相对的，因子 2 可以解释为”软实力”，由一些不容易进行标准化考核的因素构成，如申请文书, 推荐信力度以及本科大学等级．将各变量因子上的载荷可视化 (图7)，我们可以清晰看出变量的分类．

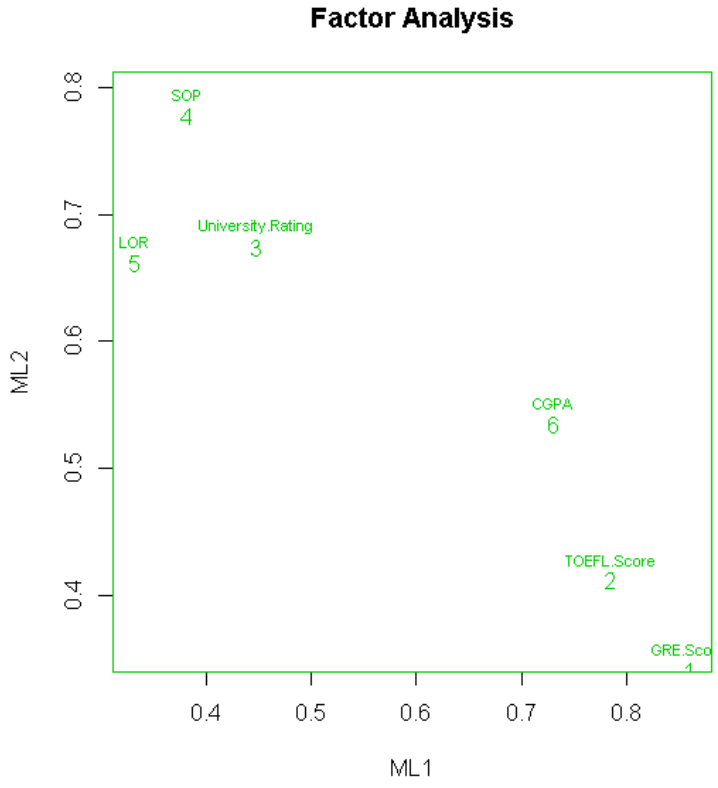

图 7 因子载荷坐标

其中 GRE 成绩在在因子 1 中最为重要，载荷 $l_{11} = 0.86$；申请文书对因子 2 最为重

要，载荷 $l_{32} = 0.78$．在方差的解释上，二者累计解释 0.76 的方差，对于因子模型来言以及较为显著‘因子 1 解释了 0.39 的方差，略高于因子 2 解释比例 0.37．我们基本认为因子 1、2 对在解释方差上同等重要．

我们考察各观测值的因子得分与录取率，研究经历的关系，如图8, 图中点的大小反映录取率的高低，颜色标准研究经历的有无，我们注意到：

- 从图斜上 45 度方向看：“硬实力”，“软实力”得分都高的个体，其录取率高（图中体现为右上角较大的点）；反之，其录取率低．（图中体现为左下角较小的点）
- 从图斜下 45 度方向看：“硬实力”，“软实力”对录取率并没有明显的侧重，即在加二者得分和相同情况下，录取率几乎相同．（图中体现为沿着斜下 45 度方向，点的大小没有明显变化）
- 研究经历对二者得分有着系统性的影响，有研究经历的个体第一、第二因子得分整体性高于没有研究经历的个体．（图中体现为蓝色点集中于右上，而橘色点集中于左下．）这一结果与 PCA 中的发现相符合．

“硬实力”与“软实力”对录取概率的作用基本与我们的常识相符，.

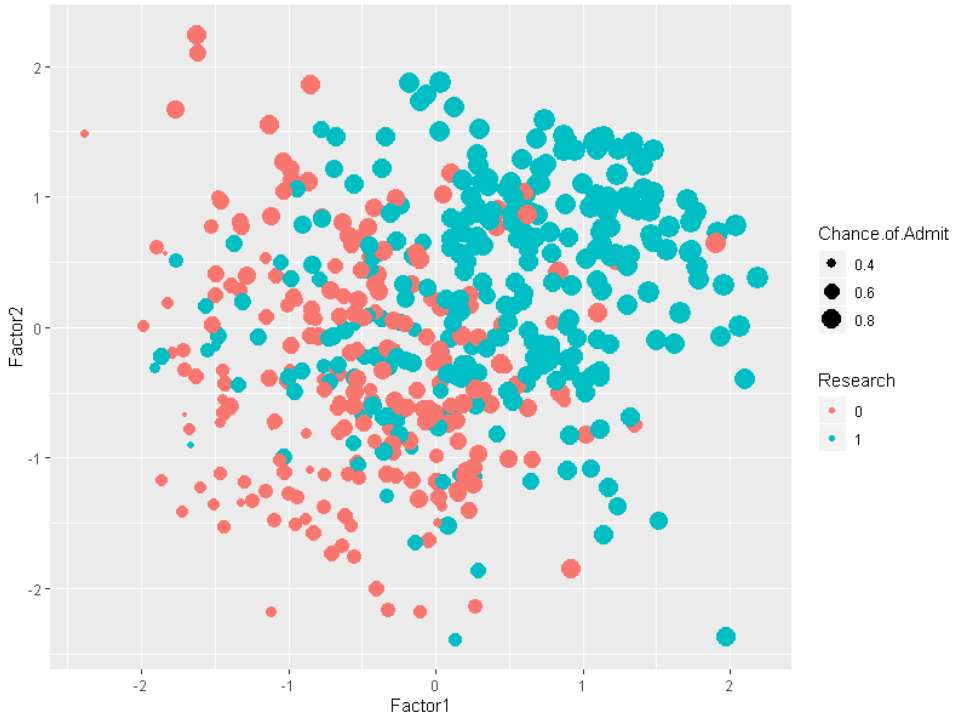

图 8 因子得分散点图

4.4 斜交因子模型

我们在提出因子的解释意义后，我们注意到这样一个问题：假设两个因子正交是否合理？我们直观上感觉，“硬实力”和“软实力”有一定的相关性，我们在相关性分析

时注意到变量间有很强的相关性，这一点的确可以利用公共因子给予解释，但如果引入斜交因子模型，是否有更好的分析效果？

我们对比利用 Minimunm residual 的方法4.4，在旋转方法为 promax 下得到的模型中（结果见附录）．在没有了因子正交的假设下，我们可以计算得到因子的相关性为 0.77，具有较强的相关性．因子模型示意图如图9．

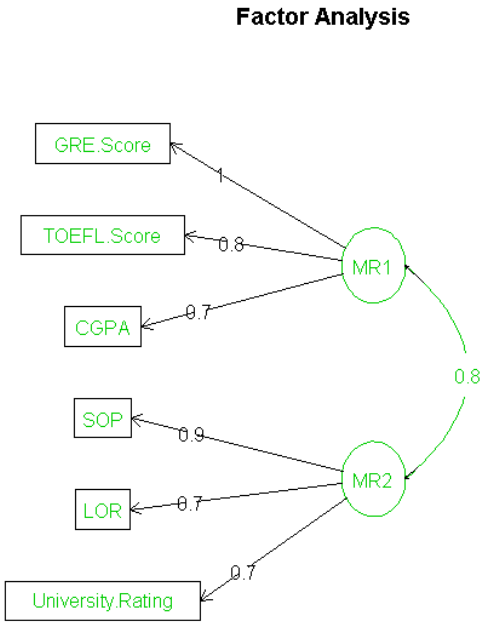

图 9 斜交因子模型示意图

变量在因子上的载荷分布更加集中，可解释性更好，代表的“软，硬实力”意义更加明晰

我们计算在该模型下因子得分，并可视化，如图10:

在斜交因子模型下的结果，与正交因子模型的结果有较好的一致性．

Remark. 我们采取的 *minimum residual* 方法与 *MLE* 方法思路相近，只是使用了最小二乘而非极大似然方法．其结果与 *MLE* 得到的相近，但是对正态性不好的数据有更优的表现[4]．

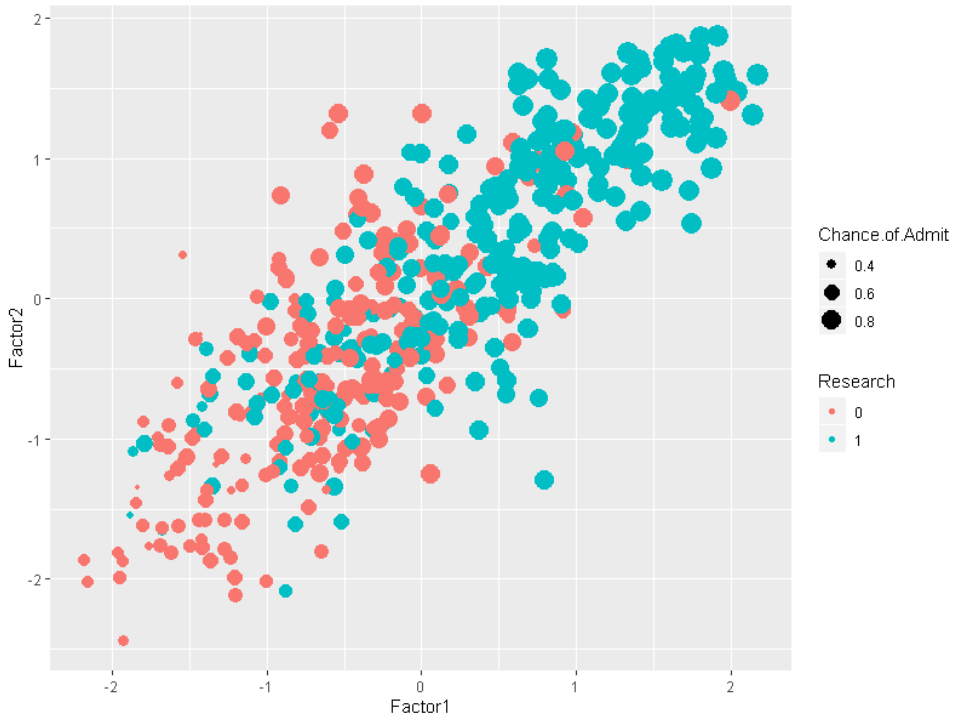

图 10 斜交因子得分散点图

五、回归模型预测

利用因子分析，我们得到对留学指标深层的解读，但是此时没有建立预测的模型，这里回归预测模型可以弥补这一缺失.

我们利用回归方法，基于赤池信息量逐步挑选变量. 我们将数据随机分成两个部分，90% 数据为训练集，其余为校验集，该方法下得到的最优模型为：

$$Chance.of.Admit \sim TOFEL + LOR + CGPA + Research$$

系数为：

$$\hat{Chance.of.Admit} = -1.30 + 0.003TOFEL.Score + 0.02LOR + 0.14CGPA + 0.12Research$$

残差分析结果显示残差正态性满足. 该模型的各变量均是显著的 ($p.value < 0.001$), 调整 R 方达到 0.8144，解释 81.4% 的方差，在加入其他变量，调整 R 方均无明显提升，变量显著性不高（$p.value > 0.01$），甚至降低其他变量显著性.（结果见附录）

利用该模型预测校验集中数据的录取率，与真实数据比较，如图11, 预测结果对录取率较高的群体的准确性要好于较低的群体.

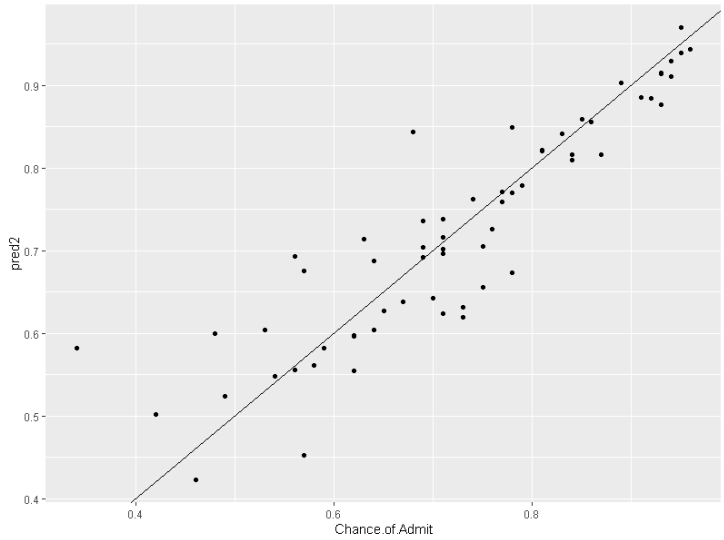

图 11 预测值与实际值比较

六、总结

6.1 结论

根据以上分析，我们回答初始提出的问题：

- 按斜交因子模型解释，存在两个因子，现实意义可以解释为可以标准化考核的“硬实力”与考察相对自由“软实力”；二者有较强的相关性 (0.77); 同时，观测样本的因子得分较高，则更有可能有较高的录取率．结合因子得分与研究经历的数据，可以比较好的判断学生的录取可能．

- 根据回归模型的结果，“硬实力”里，TODFEL.Score, CGPA 对录取率的影响最为显著；“软实力”中为推荐信，实际在模型筛选中，考虑到多重共线性的问题而舍去了 GRE 成绩，实际上在单独具有 TODFEL 成绩与 GRE 成绩的模型中，二者的调整 R 方是相近的，GRE 成绩对于录取率的影响也是显著的．

- 我们在 PCA，EFA 以及回归中都考察了研究经历对其余变量的影响，三者共同表面，研究经历可能是一个非常重要的指标：在 PCA 中，我们注意到了根据有无研究经历，在第一、第二主成分平面上，我们可以划分出两个群体，而这划分的意义在 EFA 因子得分的部分得到了展示，具有研究经历的学生普遍在因子得分较高．这一点与相关性分析的结果相符合．在回归模型中，Research 为显著的变量而得到保留，在其余变量不变，有研究经历可以提高 0.12 的录取率．这一点可能对希望进行海外留学的有所帮助，在已有的基础上，参与科研可能对于录取率的提升有所帮助．

参考文献

[1] Abdi, Hervé, and Lynne J. Williams. ”Principal component analysis.” Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

[2] https://www.researchgate.net/post/Should_I_use_PCA_with_categorical_data

[3] https://stats.idre.ucla.edu/stata/faq/how-can-i-perform-a-factor-analysis-with-categorical-or-categorical-and-continuous-variables/

[4] Revelle, William. ”How to: Use the psych package for factor analysis and data reduction.” Evanston, IL: Northwestern University, Department of Psychology (2016).

附录：因子分析模型结果

6.2 MLE 方法下，旋转为 varimax 模型输出

Listing 1: R output

```
Factor Analysis using method = minres
Call: fa(r = fa, nfactors = 2, n.obs = 500, rotate = "promax", fm = "minres")
Standardized loadings (pattern matrix) based upon correlation matrix
                   MR1   MR2   h2   u2 com
GRE.Score        0.975 -0.058 0.867 0.133 1.01
TOEFL.Score      0.822 0.091 0.799 0.201 1.02
University.Rating 0.167 0.687 0.676 0.324 1.12
SOP             -0.013 0.896 0.785 0.215 1.00
LOR              0.014 0.740 0.565 0.435 1.00
CGPA             0.651 0.315 0.842 0.158 1.44

                       MR1  MR2
SS loadings          2.339 2.194
Proportion Var       0.390 0.366
Cumulative Var       0.390 0.756
Proportion Explained 0.516 0.484
Cumulative Proportion 0.516 1.000

 With factor correlations of
     MR1  MR2
MR1 1.000 0.773
MR2 0.773 1.000

Mean item complexity = 1.1
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are 15 and the objective function was 4.856 with Chi Square of 2409.39
The degrees of freedom for the model are 4 and the objective function was 0.011

The root mean square of the residuals (RMSR) is 0.007
The df corrected root mean square of the residuals is 0.013

The harmonic number of observations is 500 with the empirical chi square 0.685 with prob < 0.953
The total number of observations was 500 with Likelihood Chi Square = 5.364 with prob < 0.252

Tucker Lewis Index of factoring reliability = 0.9979
RMSEA index = 0.026 and the 90 % confidence intervals are 0 0.0765
BIC = -19.495
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                               MR1  MR2
Correlation of (regression) scores with factors 0.968 0.947
Multiple R square of scores with factors  0.936 0.897
Minimum correlation of possible factor scores 0.873 0.793
```

6.3 Minimum residual 方法下，旋转为 promax 模型输出

Listing 2: R output

```
Factor Analysis using method = ml
Call: fa(r = fa, nfactors = 2, n.obs = 500, rotate = "varimax", fm = "mle")
Standardized loadings (pattern matrix) based upon correlation matrix
                 ML1  ML2   h2   u2 com
GRE.Score      0.860 0.358 0.867 0.133 1.34
TOEFL.Score    0.784 0.429 0.799 0.201 1.55
University.Rating 0.449 0.691 0.680 0.320 1.72
SOP            0.382 0.794 0.777 0.223 1.44
LOR            0.333 0.679 0.572 0.428 1.46
CGPA           0.731 0.552 0.839 0.161 1.86

                       ML1  ML2
SS loadings          2.348 2.186
Proportion Var       0.391 0.364
Cumulative Var       0.391 0.756
Proportion Explained 0.518 0.482
Cumulative Proportion 0.518 1.000

Mean item complexity = 1.6
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are 15 and the objective function was 4.856 with Chi Square of 2409.39
The degrees of freedom for the model are 4 and the objective function was 0.01

The root mean square of the residuals (RMSR) is 0.007
The df corrected root mean square of the residuals is 0.014

The harmonic number of observations is 500 with the empirical chi square 0.746 with prob < 0.946
The total number of observations was 500 with Likelihood Chi Square = 5.129 with prob < 0.274

Tucker Lewis Index of factoring reliability = 0.9982
RMSEA index = 0.0237 and the 90 % confidence intervals are 0 0.0751
BIC = -19.73
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                             ML1  ML2
Correlation of (regression) scores with factors 0.911 0.871
Multiple R square of scores with factors  0.830 0.758
Minimum correlation of possible factor scores 0.660 0.516
```

6.4 回归模型

Listing 3: R output

```
Call:
lm(formula = Chance.of.Admit ~ TOEFL.Score + LOR + CGPA + Research,
   data = train_l)

Residuals:
     Min       1Q   Median       3Q      Max
-0.252491 -0.025929 0.008495 0.037991 0.168431

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.008296 0.059203 -17.031 < 2e-16 ***
TOEFL.Score 0.003788 0.000827 4.580 6.08e-06 ***
LOR         0.021431  0.004046  5.297 1.87e-07 ***
CGPA        0.143883  0.009440 15.242 < 2e-16 ***
Research    0.027280  0.006803  4.010 7.15e-05 ***
---
Signif. codes: 0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯ 1

Residual standard error: 0.06059 on 433 degrees of freedom
Multiple R-squared: 0.8157, Adjusted R-squared: 0.814
F-statistic: 479.2 on 4 and 433 DF, p-value: < 2.2e-16
```

6.5 Scree plot

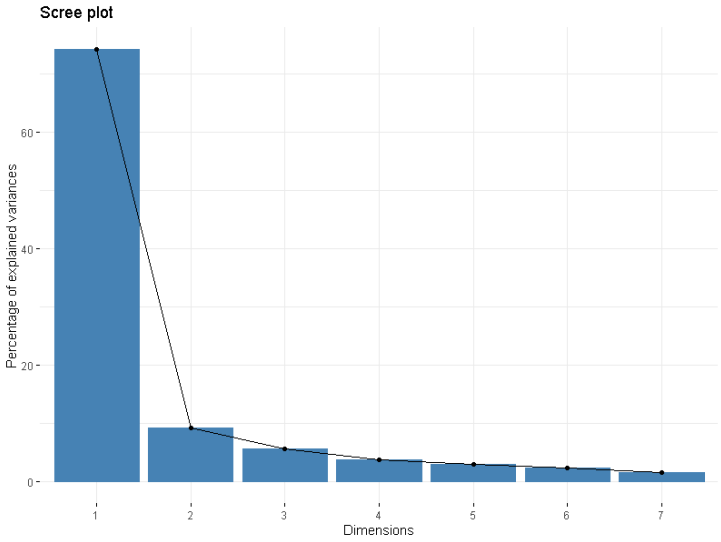

图 12 PCA scree plot

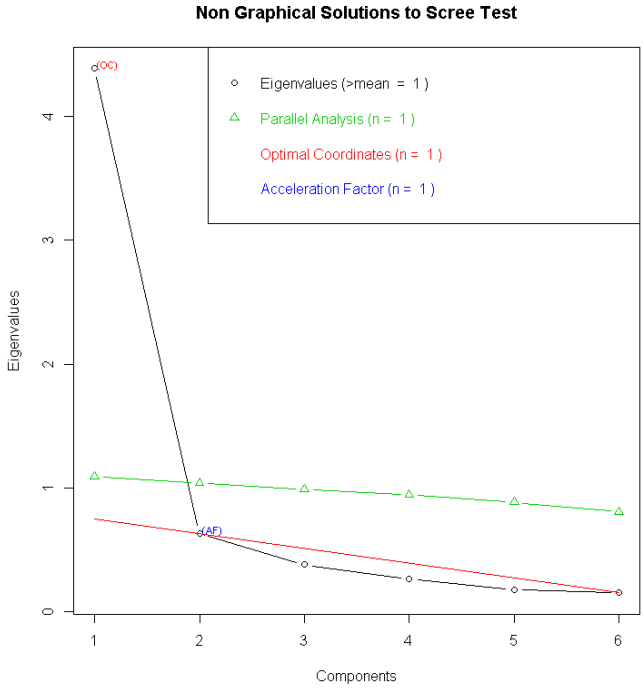

图 13 在”崖底碎石图”的基础上进行模拟