

## 多元统计分析 作业 3 因子分析

地信一班 罗皓文 15303096

## 实验环境：

OS: Windows 7 Pro SP1 x64

CPU: Intel Xeon E3-1241 v3 @ 3.50GHz 3.50GHz

RAM: 8.00 Gb

Soft: SPSS Statistics 19

## 因子分析

利用给定数据(作业 3Employee data.sav),用 spss 的 factor analysis 模块或 matlab 进行因子分析;

部分数据示例:

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority
1	1	m	02/03/1952	15	3	\$57,000	\$27,000	98	144	0
2	2	m	05/23/1958	16	1	\$40,200	\$18,750	98	36	0
3	3	f	07/26/1929	12	1	\$21,450	\$12,000	98	381	0
4	4	f	04/15/1947	8	1	\$21,900	\$13,200	98	190	0
5	5	m	02/09/1955	15	1	\$45,000	\$21,000	98	138	0
6	6	m	08/22/1958	15	1	\$32,100	\$13,500	98	67	0
7	7	m	04/26/1956	15	1	\$36,000	\$18,750	98	114	0
8	8	f	05/06/1966	12	1	\$21,900	\$9,750	98	0	0

数据变量视图:

	名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量标准	角色
1	id	数值(N)	4	0	Employee Code	无	无	4	右	度量(S)	输入
2	gender	字符串	3	0	Gender	{f, Female}...	无	4	左	名义(N)	输入
3	bdate	日期	10	0	Date of Birth	无	无	8	右	度量(S)	输入
4	educ	数值(N)	2	0	Educational Level (years)	{0, 0 (Missi...	0	4	右	序号(O)	输入
5	jobcat	数值(N)	1	0	Employment Category	{0, 0 (Missi...	0	4	右	名义(N)	输入
6	salary	美元	8	0	Current Salary	{\$0, missing...	\$0	8	右	度量(S)	输入
7	salbegin	美元	8	0	Beginning Salary	{\$0, missing...	\$0	8	右	度量(S)	输入
8	jobtime	数值(N)	2	0	Months since Hire	{0, missing}...	0	4	右	度量(S)	输入
9	prevexp	数值(N)	6	0	Previous Experience (months)	{0, missing}...	无	5	右	度量(S)	输入
10	minority	数值(N)	1	0	Minority Classification	{0, No}...	9	7	右	名义(N)	输入

(1) 使用 SPSS 线性回归模块: 分析(A) - 降维 - 因子分析(F)...

(2) 选取变量。设置参数如下:



(3) 设置描述、抽取、旋转、得分参数如下：

The figure shows four SPSS Factor Analysis dialog boxes with the following settings:

- 因子分析: 描述统计 (Descriptive Statistics):**
  - 统计量 (Statistics): ☐ 单变量描述性 (U) (Univariate Descriptive), ☒ 原始分析结果 (I) (Initial Eigenvalues).
  - 相关矩阵 (Correlation Matrix): ☐ 系数 (C) (Coefficients), ☐ 逆模型 (N) (Inverse Model), ☐ 显著性水平 (S) (Significance Level), ☐ 再生 (R) (Regeneration), ☐ 行列式 (D) (Determinant), ☐ 反映象 (A) (Reflection), ☒ KMO 和 Bartlett 的球形度检验 (K) (KMO and Bartlett's Sphericity Test).
- 因子分析: 抽取 (Extraction):**
  - 方法 (M): 主成份 (Principal Component).
  - 分析 (Analysis): ☒ 相关性矩阵 (R) (Correlation Matrix), ☐ 协方差矩阵 (V) (Covariance Matrix).
  - 输出 (Output): ☒ 未旋转的因子解 (E) (Unrotated Factor Solution), ☒ 碎石图 (S) (Scree Plot).
  - 抽取 (Extraction): ☒ 基于特征值 (E) (Eigenvalue-based), 特征值大于 (A): 1; ☐ 因子的固定数量 (N) (Fixed number of factors), 要提取的因子 (I): .
  - 最大收敛性迭代次数 (X): 25.
- 因子分析: 旋转 (Rotation):**
  - 方法 (Method): ☒ 无 (None), ☒ 最大四次方法法 (Q) (Quartimax), ☐ 最大方差法 (V) (Varimax), ☐ 最大平衡值法 (E) (Equamax), ☐ 直接 Oblimin 方法 (O) (Direct Oblimin), ☐ Promax (P) (Promax).
  - Delta (D): 0, Kappa (K): 4.
  - 输出 (Output): ☒ 旋转解 (R) (Rotated Solution), ☒ 载荷图 (L) (Loading Plot).
  - 最大收敛性迭代次数 (X): 25.
- 因子分析: 因子得分 (Factor Scores):**
  - ☒ 保存为变量 (S) (Save as variable).
  - 方法 (Method): ☒ 回归 (R) (Regression), ☐ Bartlett (B) (Bartlett), ☐ Anderson-Rubin (A) (Anderson-Rubin).
  - ☐ 显示因子得分系数矩阵 (D) (Display factor score coefficient matrix).

(4) 输出结果与分析：

表 1 KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.622
Bartlett's Test of Sphericity Approx. Chi-Square	1235.362
df	21
Sig.	.000

表 1 为 KMO 与 Bartlett's 球形检验度。KMO 为 Kaiser-Meyer-Olkin 的取样适当性量，KMO 测度越高（越接近 1）时，表明变量间共同因子越多，数据越适合进行因子分析。一般来说 KMO 达到 0.9 以上为极好，0.8-0.9 为好，0.7-0.8 为一般，0.6-0.7 为较差，0.5-0.6 为差。若 KMO 小于 0.5 表明样本偏小，需要扩大样本。本例中，KMO=0.622，表明变量间共同因子并不多，但仍可以使用因子分析进行分析。

Bartlett's 球形检验度目的是检验相关矩阵是否为单位阵，其零假设为相关矩阵为单位阵，若拒绝零假设则表明数据适合用于因子分析，否则表明数据相关矩阵很有可能是单位阵，不宜使用因子分析。本例中卡方值为 1235.362，自由度为 21，显著性  $0.000 < 0.001$ ，表明可以拒绝零假设，相关矩阵不是单位阵，共同因子存在，适合进行因子分析。

表 2 为共同因子方差表，表明每个变量被解释的方差量。初始共同因子方差 (Initial Communalities) 为每个变量被所有成分或因子解释的方差估计量。对于主成分分析，该值恒为 1，因为原始变量数和成分数相同。

抽取共同因子 (Extraction Communalities) 指因子解中用于预测因子的变量的多重相关的平方.该值越小说明该变量不适合作因子，可在分析中排除之。

表 2 Communalities

	Initial	Extraction
Date of Birth	1.000	.875
Educational Level (years)	1.000	.728
Employment Category	1.000	.788
Beginning Salary	1.000	.859
Months since Hire	1.000	.814
Previous Experience (months)	1.000	.893
Minority Classification	1.000	.307

Extraction Method: Principal Component Analysis.

表 3 Total Variance Explained

Comp onent	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
				Loadings			Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.392	34.177	34.177	2.392	34.177	34.177	2.306	32.942	32.942
2	1.844	26.339	60.516	1.844	26.339	60.516	1.907	27.240	60.182
3	1.029	14.705	75.220	1.029	14.705	75.220	1.053	15.039	75.220
4	.911	13.009	88.230						
5	.414	5.907	94.137						
6	.218	3.112	97.249						
7	.193	2.751	100.000						

Extraction Method: Principal Component Analysis.

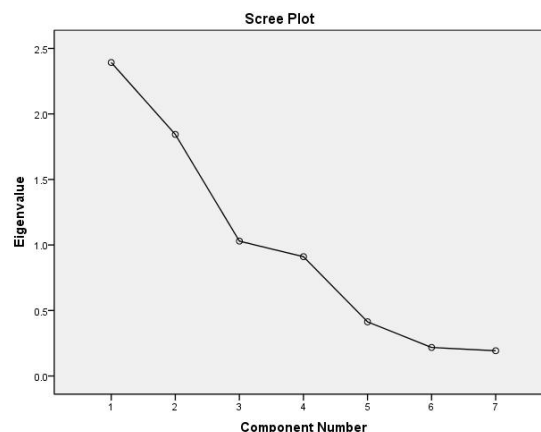


图 1 碎石图

表 3 所示旋转前后总的解释方差。如表中，共 7 个成分都有初始特征值，本例中取特征值大于 1 的 3 个成分，这 3 个成分能够解释原来 75.220% 的方差。旋转后的特征值与原来的特征值有所不同，旋转后各成分间的特征值的差异会变小。但累计方差百分比不变。

如图 1 为碎石图，其作用于表 3 类似，都是为了确定因子数目。可以看出在第三个因子之后曲线变得平缓，因此可以认为抽取 3 个因子是较为合理的。

表 4 为旋转前后成分矩阵，表示各个变量在未旋转成分或因子上的载荷。图 2 显示了各变量在旋转后的因子空间中的分布情况。

表 4 Component Matrix

	Component <sup>a</sup>			Rotated Component <sup>b</sup>		
	1	2	3	1	2	3
Date of Birth	.399	-.845	.036	.053	-.931	-.076
Educational Level (years)	.841	.012	.145	.792	-.314	.046
Employment Category	.777	.429	-.005	.879	.111	-.063
Beginning Salary	.832	.409	-.014	.920	.074	-.079
Months since Hire	-.005	.055	.901	.087	-.051	.897
Previous Experience (months)	-.354	.873	-.070	-.003	.944	.039
Minority Classification	-.323	.112	.437	-.221	.172	.479

a. 3 components extracted.

b. Rotation converged in 4 iterations.

最终结果为：

$$X = AF,$$

其中，X 为变量向量，F 为因子向量，A 为成分矩阵：

$$A = \begin{bmatrix} 0.053 & -0.931 & -0.076 \\ 0.792 & -0.314 & 0.046 \\ 0.879 & 0.111 & -0.063 \\ 0.920 & 0.074 & -0.079 \\ 0.087 & -0.051 & 0.897 \\ -0.003 & 0.944 & 0.039 \\ -0.221 & 0.172 & 0.479 \end{bmatrix}.$$

SPSS 可以自动计算各行对应各因子的值。

FAC1_1	FAC2_1	FAC3_1
1.67398	.49502	1.18041
.32212	-.66023	1.32099
-.33297	2.30664	1.21878
-.83051	.86560	1.07134
.36016	.03319	1.26852
-.05333	-.52683	1.30664

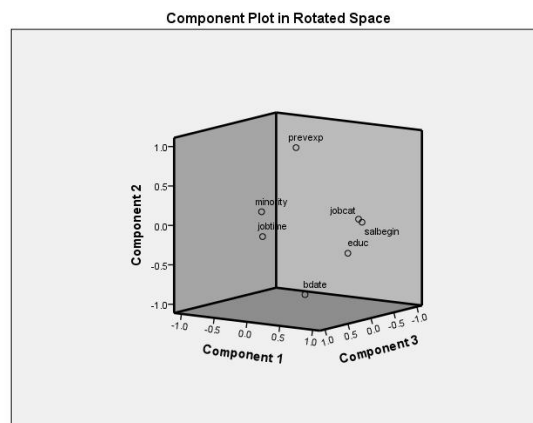


图 2 各变量在旋转后空间的位置