多元-10-2020270026

2020270026 王姿文

5/10/2021

1. 数据

- 数据叙述：全国31个省、直辖市、自治区的城镇居民平均每人全年家庭收入来源及现金消费支出情况，其中**收入来源**- 工资性收入 (x1)、 经营净收入 (x2)、 财产性收入 (x3)、 转移性收入 (x4)；**现金消费性支出指标**- 食品 (y1)、 衣着 (y2)、 居住 (y3)、 交通和通信 (y4)
- 目标： 研究**收入来源**与**现金消费性支出指标**的关系。仅分析相关性，而非因果关系，进行典型相关分析

下表为数据，以及数据的结构：

```
df <- read_excel("ex9.5.xls")
a <- df$地区
df <- df[,-1]
rownames(df) <- a
colnames(df) <- c('工资性收入','经营净收入','财产性收入','转移性收入',
                  '食品','衣着','居住','交通和通信')
kbl(df) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size =
7)
```

	工资性收入	经营净收入	财产性收入	转移性收入	食品	衣着	居住	交通和通信
北京	27961.78	1430.22	717.56	10993.54	7535.29	2638.90	1970.94	3781.51
天津	21523.81	1200.10	515.49	9704.61	7343.64	1881.43	1854.22	3083.37
河北	13154.52	2257.48	338.47	6148.95	4211.16	1541.99	1502.41	1723.75
山西	14973.64	1041.43	301.84	5783.41	3855.56	1529.47	1438.88	1672.29
内蒙古	16872.58	2698.67	564.02	4655.51	5463.18	2730.23	1583.56	2572.93
辽宁	14846.05	2710.30	493.01	7866.35	5809.39	2042.40	1433.28	2323.29
吉林	13535.33	2168.82	324.03	5631.45	4635.27	2044.80	1594.14	1780.67
黑龙江	11700.50	1729.29	186.10	5751.95	4687.23	1806.92	1336.85	1462.61
上海	31109.30	2267.15	575.82	10802.23	9655.60	2111.17	1790.48	4563.80
江苏	20102.05	3421.90	689.96	8305.20	6658.37	1915.97	1437.08	2689.51
浙江	22385.09	4694.40	1465.32	9450.02	7552.02	2109.58	1551.69	4133.50
安徽	14812.54	2155.33	549.62	6007.07	5814.92	1540.66	1396.97	1809.72
福建	19976.01	3336.96	1795.21	5769.73	7317.42	1634.21	1753.86	2961.78
江西	13348.06	1946.82	527.63	5327.72	5071.61	1476.63	1173.91	1501.34
山东	19856.05	2621.41	704.90	4823.24	5201.32	2196.98	1572.35	2370.23
河南	13666.49	2545.14	333.81	5351.78	4607.47	1885.99	1190.81	1730.35
湖北	14191.04	2158.33	476.23	6078.25	5837.93	1783.41	1371.15	1476.98
湖南	13237.06	3008.33	867.76	5691.40	5441.63	1624.57	1301.60	2084.15
广东	23632.20	3603.89	1468.73	5339.56	8258.44	1520.59	2099.75	4176.66
广西	14693.47	2131.79	883.71	5500.43	5552.56	1146.46	1377.26	2088.64
海南	14672.28	2397.44	717.61	5022.54	6556.10	864.96	1521.04	2004.34
重庆	15415.44	2183.51	538.43	6673.59	6870.23	2228.76	1177.02	1903.24
四川	14249.32	2017.84	633.82	5427.34	6073.86	1651.14	1284.09	1946.72
贵州	12309.17	1982.45	355.70	5395.56	4992.85	1399.00	1013.53	1891.03

云南	14408.29	2425.03	999.98	5167.14	5468.17	1759.89	973.76	2264.23
西藏	17672.12	570.88	417.86	1563.31	5517.69	1361.57	845.18	1387.45
陕西	15547.32	881.96	269.58	5907.14	5550.71	1789.06	1322.22	1788.38
甘肃	12514.92	1125.68	259.63	4598.23	4602.33	1631.40	1287.93	1575.67
青海	12614.39	1191.42	92.98	5847.84	4667.34	1512.24	1232.39	1549.76
宁夏	13965.62	2522.84	160.88	5252.90	4768.91	1875.70	1193.37	2110.41
新疆	14432.12	1633.22	145.50	3983.71	5238.89	2031.14	1166.59	1660.27

```
str(df)
```

```
## tibble[,8] [31 × 8] (S3: tbl_df/tbl/data.frame)
##  $ 工资性收入: num [1:31] 27962 21524 13155 14974 16873 ...
##  $ 经营净收入: num [1:31] 1430 1200 2257 1041 2699 ...
##  $ 财产性收入: num [1:31] 718 515 338 302 564 ...
##  $ 转移性收入: num [1:31] 10994 9705 6149 5783 4656 ...
##  $ 食品       : num [1:31] 7535 7344 4211 3856 5463 ...
##  $ 衣着       : num [1:31] 2639 1881 1542 1529 2730 ...
##  $ 居住       : num [1:31] 1971 1854 1502 1439 1584 ...
##  $ 交通和通信: num [1:31] 3782 3083 1724 1672 2573 ...
```

---

2. 典型相关分析

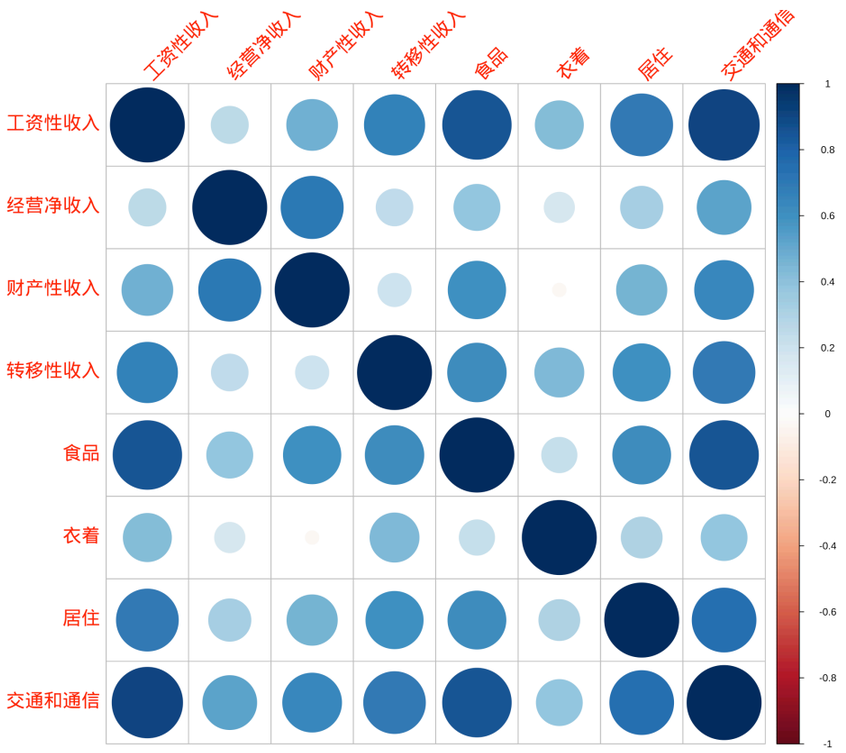
首先来看8个变量的相关系数阵，可以看到相关系数>0.7的只有下面五项，并没很多项：

- 食品 vs 工资性收入
- 交通和通信 vs 工资性收入
- 财产性收入 vs 经营净收入
- 交通和通信 vs 食品
- 交通和通信 vs 居住

```
# corr
cor(df)
```

```
##              工资性收入 经营净收入  财产性收入 转移性收入       食品         衣着
## 工资性收入   1.0000000  0.2514160  0.47145061  0.6625290 0.8558322  0.42234913
## 经营净收入   0.2514160  1.0000000  0.70402929  0.2461835 0.3871062  0.16709554
## 财产性收入   0.4714506  0.7040293  1.00000000  0.2012875 0.6018076 -0.03220643
## 转移性收入   0.6625290  0.2461835  0.20128753  1.0000000 0.6254785  0.43888810
## 食品         0.8558322  0.3871062  0.60180762  0.6254785 1.0000000  0.22712753
## 衣着         0.4223491  0.1670955 -0.03220643  0.4388881 0.2271275  1.00000000
## 居住         0.6985302  0.3258861  0.46182699  0.5972373 0.6118066  0.30534466
## 交通和通信   0.9038415  0.5291950  0.63214441  0.6951497 0.8589257  0.38533608
##                    居住 交通和通信
## 工资性收入 0.6985302  0.9038415
## 经营净收入 0.3258861  0.5291950
## 财产性收入 0.4618270  0.6321444
## 转移性收入 0.5972373  0.6951497
## 食品       0.6118066  0.8589257
## 衣着       0.3053447  0.3853361
## 居住       1.0000000  0.7420448
## 交通和通信 0.7420448  1.0000000
```

```
# 食品 vs 工资性收入; 交通和通信 vs 工资性收入;
# 财产性收入 vs 经营净收入; 交通和通信 vs 食品;
# 交通和通信 vs 居住
corrplot(cor(df),tl.cex=1.5,tl.srt=45)
```



接着来看研究**收入来源**与**现金消费性支出指标**的典型相关分析：

首先能看出第一对典型相关系数很大，高达0.9717098

```
ds <- scale(df)
cc1 <- cancor(ds[,1:4], ds[,5:8])
cc1$cor# 第一对典型相关系数很大
```

```
## [1] 0.9717098 0.4925284 0.3122749 0.1216499
```

仔细看第一对典型相关维度内各个变量的系数，可以看出**收入来源**是 工资性收入 的对比，而**现金消费性支出指标**是 交通和通信 的对比，并且 工资性收入 和 交通和通信 呈正相关

```
cc1$xcoef
```

```
##                  [,1]         [,2]        [,3]         [,4]
## 工资性收入 0.13286255 -0.098195311 -0.00126420 -0.236350584
## 经营净收入 0.03893537 -0.187154119  0.19492511 -0.002965217
## 财产性收入 0.01910854  0.287829101 -0.03043755  0.084578882
## 转移性收入 0.03289586  0.009524422 -0.10835869  0.234323514
```

```
cc1$ycoef
```

```
##                  [,1]        [,2]        [,3]        [,4]
## 食品        0.04657712  0.10643415 -0.27505098 -0.21225000
## 衣着        0.01603639 -0.17450230 -0.09523197 -0.03694245
## 居住        0.01028582  0.05179583 -0.16585735  0.21050385
## 交通和通信 0.12693326 -0.06103539  0.41985724  0.04990190
```

且也能看到各个省份的xscores, yscores…等

```
cc2 <- cc(ds[,1:4], ds[,5:8])
cc2$scores$xscores
```

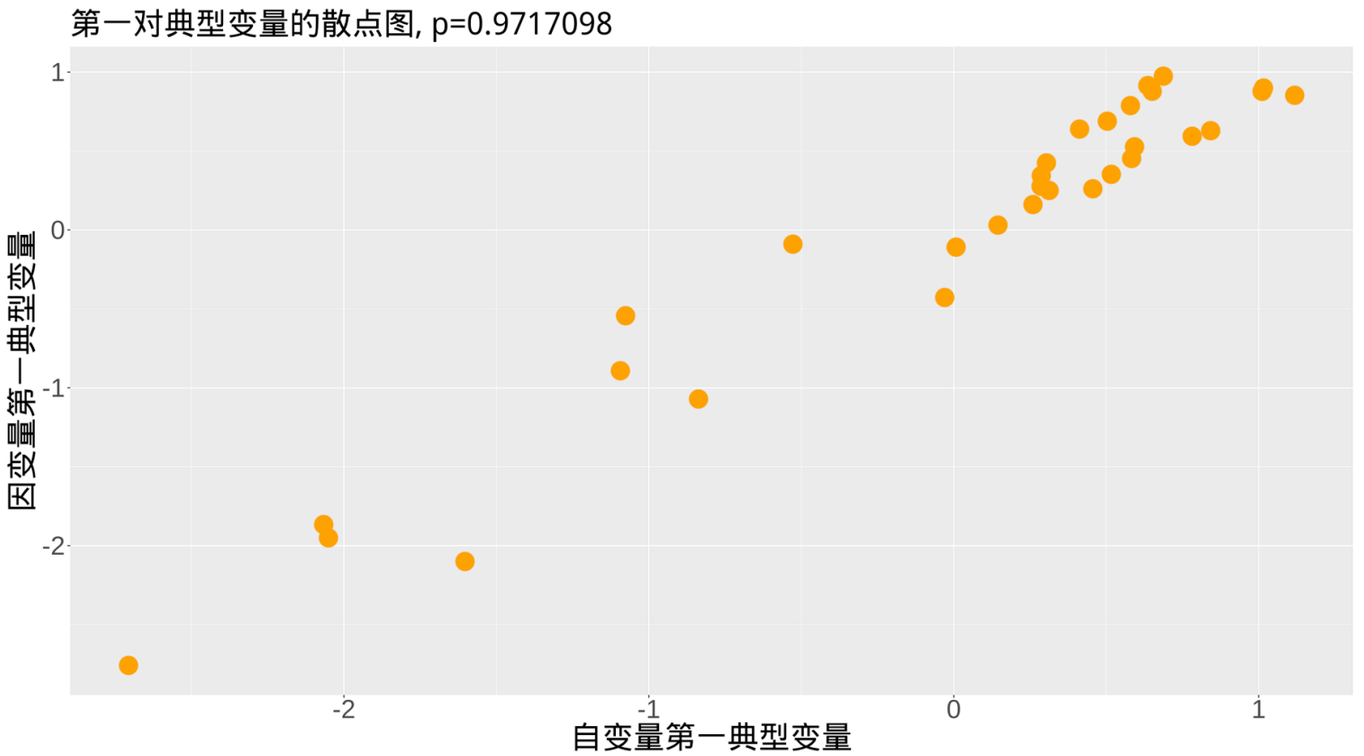
```
##                 [,1]        [,2]         [,3]         [,4]
## 北京    -2.066447310  0.20950623 -2.474116478 -0.177406396
## 天津    -0.836994554  0.38905361 -2.273208869 -0.889840090
## 河北     0.579342094 -0.68321421  0.179235020 -0.664153418
## 山西     0.636456444  0.38209223 -1.186134172  0.097834670
## 内蒙古 -0.029715928 -0.77890119  1.069851999  1.086021009
## 辽宁     0.007611411 -0.75452122  0.148530790 -1.486340144
## 吉林     0.592885741 -0.69355567  0.232241921 -0.206150460
## 黑龙江   1.011080583 -0.50711567 -0.281705385 -0.641102026
## 上海    -2.706392800 -1.70045979 -1.337717876  0.998969837
## 江苏    -1.076367460 -1.40654524  0.797070441 -0.532032399
## 浙江    -2.050785897 -0.07425535  1.682177969 -1.521705041
## 安徽     0.303858669  0.07554522  0.006434678 -0.359219036
## 福建    -1.093445591  3.00222993  0.996607879 -0.196898246
## 江西     0.650769778  0.38426160 -0.032541068 -0.300193968
## 山东    -0.527563183 -0.47099541  0.861514112  1.638092283
## 河南     0.503516118 -1.11963958  0.772832479  0.008202098
## 湖北     0.412592604 -0.14420389  0.020190195 -0.492592045
## 湖南     0.286166406  0.50166197  1.015078771 -0.942742760
## 广东    -1.602981678  0.96811134  1.583815617  1.479722421
## 广西     0.287000242  1.42190257 -0.008847579 -0.449689190
## 海南     0.312625655  0.44405269  0.529700989  0.053619838
## 重庆     0.145027375 -0.05342946 -0.156283285 -0.613366795
## 四川     0.456160058  0.61881444 -0.021317045 -0.237374846
## 贵州     0.842629425 -0.21476170  0.063894538 -0.432104725
## 云南     0.259764507  1.56016390  0.402415818 -0.440710597
## 西藏     0.687297416  0.97050191 -0.541881687  3.453656605
## 陕西     0.583252425  0.37943272 -1.405987715  0.210471458
## 甘肃     1.118189994  0.36801208 -0.704396516  0.240246464
## 青海     1.015903744 -0.34521137 -0.931350720 -0.352562325
## 宁夏     0.516756016 -1.81291276  0.847081822  0.355640243
## 新疆     0.781807694 -0.91561996  0.146813357  1.313707583
```

```
cc2$scores$yscores
```

```
##                [,1]         [,2]        [,3]         [,4]
## 北京    -1.86636378 -1.369859227 -0.85715111 -0.793789156
## 天津    -1.07110827  0.559753932 -1.10788620 -0.662520251
## 河北     0.78745428  0.164104919  0.47799003 -1.776775409
## 山西     0.91423996 -0.008301942  0.97342706 -1.826291612
## 内蒙古 -0.42800832 -2.442893849 -0.56077989 -0.638593903
## 辽宁    -0.10926576 -0.650853717 -0.22357924  0.004356629
## 吉林     0.52646870 -0.811617409 -0.83036939 -1.527457991
## 黑龙江  0.87783997 -0.338077455 -0.59488553 -0.456082238
## 上海    -2.75864554  0.393747666  0.06016300  1.314742037
## 江苏    -0.54359931 -0.097645920 -0.07273619  0.566241218
## 浙江    -1.95069320 -0.618109318  2.11809966  0.541444221
## 安徽     0.42459946  0.747642064 -0.81123876  0.059029242
## 福建    -0.89204317  1.102204259 -0.74841530 -0.367327274
## 江西     0.87816896  0.467776726  0.02880278  0.367114636
## 山东    -0.09038372 -1.182344457 -0.04709003 -1.040296976
## 河南     0.68865463 -0.818400197  0.57469286  0.023516022
## 湖北     0.63888388  0.264580984 -1.96873652  0.416613758
## 湖南     0.27645062  0.171954341  0.54732825  0.070148302
## 广东    -2.10019505  1.678714388  0.44600435 -1.378786261
## 广西     0.34398571  1.470836773  0.82924742 -0.388938775
## 海南     0.25001412  2.789547242 -0.64264278 -0.196634972
## 重庆     0.03036768 -0.726894084 -2.00193847  2.226951938
## 四川     0.26074045  0.425981888 -0.53169267  0.764332715
## 贵州     0.62863073  0.311341376  1.77889484  0.785788044
## 云南     0.16060669 -0.547097163  1.86762263  1.442249397
## 西藏     0.97391670  0.665418157  0.41691663  2.081050924
## 陕西     0.45244570 -0.047973781 -0.65529506  0.263145505
## 甘肃     0.85265153 -0.037662452  0.19765754 -0.459984571
## 青海     0.89876420  0.238756123  0.39118845 -0.229441863
## 宁夏     0.35211103 -0.865602502  1.40755033  0.031562571
## 新疆     0.59331114 -0.889027367 -0.46114869  0.784634096
```

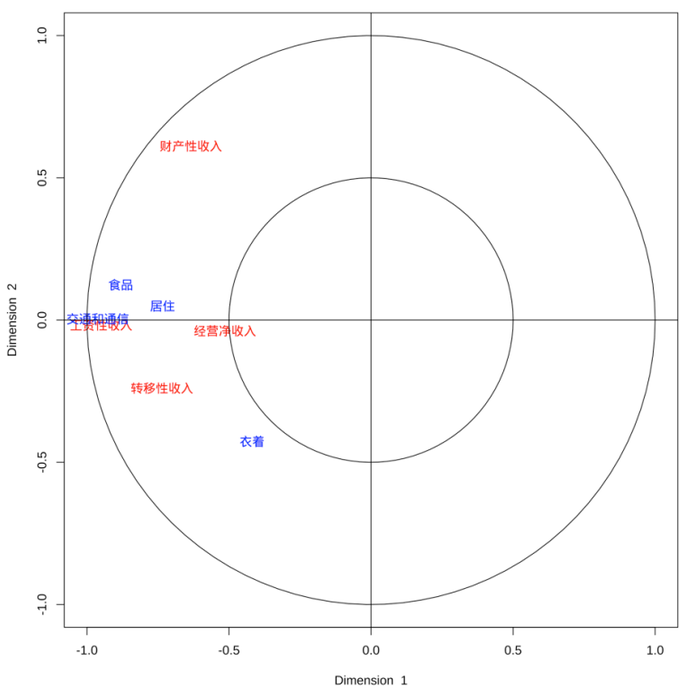
接著來看第一对典型变量的散点图, 其相關係數是0.9717098

```
dfc <- data.frame(cc2$scores$xscores[, 1], cc2$scores$yscores[, 1])
colnames(dfc) <- c('x','y')
ggplot(dfc, aes(x, y)) +
  geom_point(color = 'orange',size=8) +
  labs(title ='第一对典型变量的散点图, p=0.9717098',
       x='自变量第一典型变量',y='因变量第一典型变量') +
  theme(plot.title = element_text(size=30, face="bold"),
        axis.title = element_text(size=30, face="bold"),
        axis.text = element_text(size=25))
```



再来看所有原始变量与前两对典型变量得分的相关系数的散点图， 横坐标是第一典型变量权重， 纵坐标是第二 典型权重。 自变量为红色，因变量为蓝。由于变量的点较为集中，说明集中的点内的自变量、因变量与前两对典型变量的相关性很类似。举例来看： 工资性收入 和 交通和通信 第一对相关系数绝对值均很高，且第二对相关系数绝对值均很低

```
plt.cc(cc2, d1 = 1, d2 = 2, type = "v", var.label = TRUE)
```




```
cor(ds[,1:4], cc2$scores$xscores[ , 1:2])
```

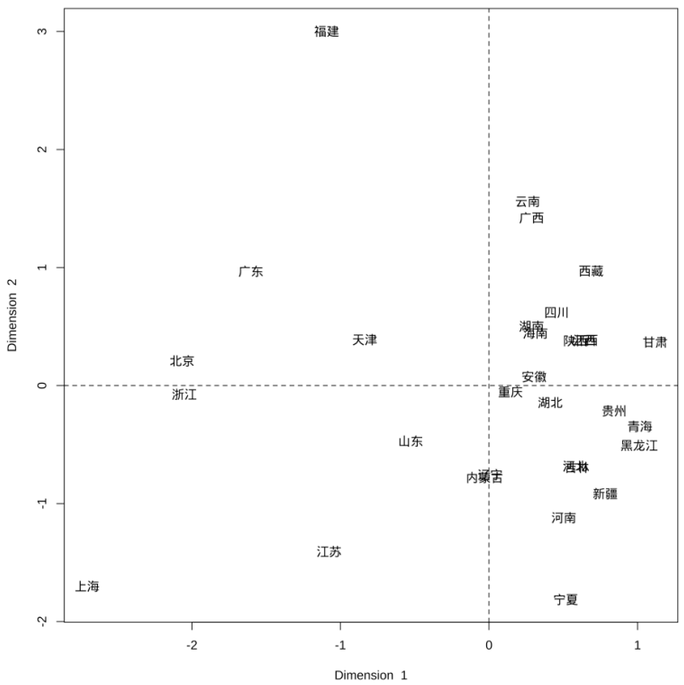
```
##                   [,1]        [,2]
## 工资性收入 -0.9500507 -0.01775406
## 经营净收入 -0.5142596 -0.03755796
## 财产性收入 -0.6341523  0.61175148
## 转移性收入 -0.7358802 -0.23919412
```

```
cor(ds[,5:8], cc2$scores$yscores[ , 1:2])
```

```
##                   [,1]         [,2]
## 食品         -0.9066923  0.252303108
## 衣着         -0.4308825 -0.865575474
## 居住         -0.7551386  0.100444627
## 交通和通信 -0.9900167  0.008634464
```

也可将每个观测用第一和第二典型变量表示，可以看出重点直辖市均在Dimension 1<0的位置上

```
plt.cc(cc2, d1 = 1, d2 = 2, type = "i", var.label = TRUE)
```



最后可以用Pillai-Bartlett Trace方法来检验此典型相关系数，可以看出第一典型相关的p-value<0.05，非常显著，所以由第一典型相关得出的结论均有一定可靠性

```
p.asym(cc1$cor, N=nrow(ds),p=4, q=4, tstat='Pillai')
```

```
##  Pillai-Bartlett Trace, using F-approximation:
##                stat    approx df1 df2      p.value
## 1 to 4:  1.2991185 3.1264869  16 104 0.0002377982
## 2 to 4:  0.3548985 1.2116302   9 112 0.2949570767
## 3 to 4:  0.1123143 0.8666929   4 120 0.4861467141
## 4 to 4:  0.0147987 0.4753168   1 128 0.4917992345
```

---

3. 結論

整体看来研究**收入来源**与**现金消费性支出指标**具有相关性，其相关性主要发生在 工资性收入 和 交通和通信 层面。