第六讲:典型相关分析

典型相关分析（Canonical Correlation analysis）

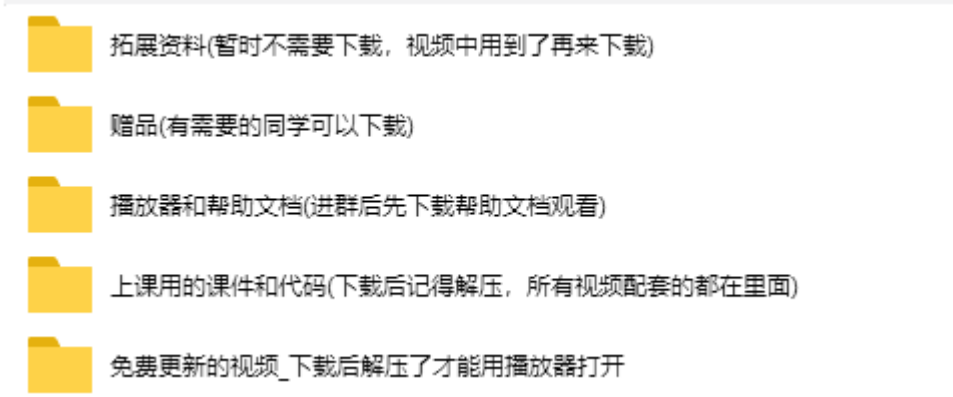
研究两组变量（每组变量中都可能有多个指标）之间相关关系的一种多元统计方法。它能够揭示出两组变量之间的内在联系。

温馨提示：这一讲涉及到多元统计的知识，如果对原理听不懂也没关系，等以后学完了主成分分析模型后再回过头来看比较合适。

温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、我视频中推荐的资料**等。



（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

一些例子

我们要探究观众和业内人士对于一些电视节目的观点有什么样的关系呢?

观众评分来自低学历（led）、高学历（hed）和网络（net）调查三种,它们形成第一组变量；而业内人士分评分来自包括演员和导演在内的艺术家（arti）、发行（com）与业内各部门主管（man）三种，形成第二组变量。

对电视节目的打分

#	led	hed	net	arti	com	man	#	led	hed	net	arti	com	man
1	86	43	85	43	93	71	16	39	80	71	76	52	81
2	99	74	99	78	99	89	17	65	5	53	11	67	41
3	37	22	10	27	24	33	18	28	11	31	12	23	35
4	5	19	56	13	11	38	19	50	32	68	23	49	58
5	45	43	55	39	54	58	20	69	98	69	97	81	99
6	21	32	21	34	35	32	21	55	99	78	97	60	90
7	36	78	48	75	42	78	22	36	11	5	15	26	5
8	69	31	85	32	70	52	23	77	18	61	27	68	54
9	40	98	36	99	64	86	24	67	33	95	34	59	61
10	26	14	40	8	25	21	25	45	87	46	85	67	80
11	51	68	38	68	48	72	26	61	72	63	63	62	75
12	63	86	79	87	76	95	27	41	63	74	55	50	76
13	39	80	57	80	55	68	28	6	5	13	5	5	13
14	78	40	72	42	75	58	29	28	53	35	51	31	59
15	56	49	54	48	52	61	30	66	20	79	18	67	55

案例来源：人大吴喜之《从数据到结论》 文件：tv.xlsx

一些小例子

对电视节目的打分													
#	led	hed	net	arti	com	man	#	led	hed	net	arti	com	man
1	86	43	85	43	93	71	16	39	80	71	76	52	81
2	99	74	99	78	99	89	17	65	5	53	11	67	41
3	37	22	10	27	24	33	18	28	11	31	12	23	35
4	5	19	56	13	11	38	19	50	32	68	23	49	58
5	45	43	55	39	54	58	20	69	98	69	97	81	99
6	21	32	21	34	35	32	21	55	99	78	97	60	90
7	36	78	48	75	42	78	22	36	11	5	15	26	5
8	69	31	85	32	70	52	23	77	18	61	27	68	54
9	40	98	36	99	64	86	24	67	33	95	34	59	61
10	26	14	40	8	25	21	25	45	87	46	85	67	80
11	51	68	38	68	48	72	26	61	72	63	63	62	75
12	63	86	79	87	76	95	27	41	63	74	55	50	76
13	39	80	57	80	55	68	28	6	5	13	5	5	13
14	78	40	72	42	75	58	29	28	53	35	51	31	59
15	56	49	54	48	52	61	30	66	20	79	18	67	55

直接对这些变量的相关进行两两分析，很难得到关于这两组变量（观众和业内人士）之间关系的一个清楚的印象。

解决思路

对电视节目的打分													
#	led	hed	net	arti	com	man	#	led	hed	net	arti	com	man
1	86	43	85	43	93	71	16	39	80	71	76	52	81
2	99	74	99	78	99	89	17	65	5	53	11	67	41
3	37	22	10	27	24	33	18	28	11	31	12	23	35
4	5	19	56	13	11	38	19	50	32	68	23	49	58
5	45	43	55	39	54	58	20	69	98	69	97	81	99
6	21	32	21	34	35	32	21	55	99	78	97	60	90
7	36	78	48	75	42	78	22	36	11	5	15	26	5
8	69	31	85	32	70	52	23	77	18	61	27	68	54
9	40	98	36	99	64	86	24	67	33	95	34	59	61
10	26	14	40	8	25	21	25	45	87	46	85	67	80
11	51	68	38	68	48	72	26	61	72	63	63	62	75
12	63	86	79	87	76	95	27	41	63	74	55	50	76
13	39	80	57	80	55	68	28	6	5	13	5	5	13
14	78	40	72	42	75	58	29	28	53	35	51	31	59
15	56	49	54	48	52	61	30	66	20	79	18	67	55

把多个变量与多个变量之间的相关化为两个具有代表性的变量之间的相关。

选谁做代表？

对电视节目的打分													
#	led	hed	net	arti	com	man	#	led	hed	net	arti	com	man
1	86	43	85	43	93	71	16	39	80	71	76	52	81
2	99	74	99	78	99	89	17	65	5	53	11	67	41
3	37	22	10	27	24	33	18	28	11	31	12	23	35
4	5	19	56	13	11	38	19	50	32	68	23	49	58
5	45	43	55	39	54	58	20	69	98	69	97	81	99
6	21	32	21	34	35	32	21	55	99	78	97	60	90
7	36	78	48	75	42	78	22	36	11	5	15	26	5
8	69	31	85	32	70	52	23	77	18	61	27	68	54
9	40	98	36	99	64	86	24	67	33	95	34	59	61
10	26	14	40	8	25	21	25	45	87	46	85	67	80
11	51	68	38	68	48	72	26	61	72	63	63	62	75
12	63	86	79	87	76	95	27	41	63	74	55	50	76
13	39	80	57	80	55	68	28	6	5	13	5	5	13
14	78	40	72	42	75	58	29	28	53	35	51	31	59
15	56	49	54	48	52	61	30	66	20	79	18	67	55

代表：能较为综合、全面的衡量所在组的内在规律。

一组变量最简单的综合形式就是该组变量的线性组合。

典型相关分析的定义

典型相关分析由Hotelling提出，其基本思想和主成分分析非常相似。

首先在每组变量中找出变量的线性组合，使得两组的线性组合之间具有最大的相关系数；

然后选取和最初挑选的这对线性组合不相关的线性组合，使其配对，并选取相关系数最大的一对；

如此继续下去，直到两组变量之间的相关性被提取完毕为止。

被选出的线性组合配对称为**典型变量**，它们的相关系数称为**典型相关系数**。典型相关系数度量了这两组变量之间联系的强度。

典型相关分析的思路

假设两组变量分别为：$X^{(1)}=(X_1^{(1)},X_2^{(1)},\cdots,X_p^{(1)}),X^{(2)}=(X_1^{(2)},X_2^{(2)},\cdots,X_q^{(2)})$

分别在两组变量中选取若干有代表性的综合变量U_i、V_i，

使得每一个综合变量是原变量的线性组合，即

$$U_i=a_1^{(i)}X_1^{(1)}+a_2^{(i)}X_2^{(1)}+\cdots+a_p^{(i)}X_p^{(1)}\triangleq\mathbf{a}^{(i)'}\mathbf{X}^{(1)}$$

$$V_i=b_1^{(i)}X_1^{(2)}+b_2^{(i)}X_2^{(2)}+\cdots+b_q^{(i)}X_q^{(2)}\triangleq\mathbf{b}^{(i)'}\mathbf{X}^{(2)}$$

注意：综合变量的组数是不确定的，如果第一组就能代表原样本数据大部分的信息，

那么一组就足够了。假设第一组反映的信息不够，我们就需要找第二组了。

并且为了让第二组的信息更有效，需要保证两组的信息不相关。

不相关：$cov(U_1,U_2)=cov(V_1,V_2)=0$

第一组要满足的条件：

在$var(U_1)=var(V_1)=1$满足的条件下，找到$\mathbf{a}^{(1)}$和$\mathbf{b}^{(1)}$两组系数，使得$\rho(U_1,V_1)$最大

（为什么要固定这个条件：因为相关系数与量纲无关：$\rho(U_1,V_1)=\rho(aU_1,bV_1)$）

后面的证明见一个很全面的PPT

厦门大学多元统计分析
第九章典型相关分析.ppt

典型相关分析关键步骤

（1）**数据的分布有假设：**两组数据服从联合正态分布。

- 在实际分析应用中，总体的协差阵通常是未知的，往往需要从研究的总体中随机抽取一个样本，根据样本估计出总体的协差阵，并在此基础上进行典型相关分析。

- 设 $\mathbf{X}=\begin{bmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{bmatrix}$ 服从正态分布 $N_{p+q}(\boldsymbol{\mu},\boldsymbol{\Sigma})$，从该总体中抽取样本容量为 n 的样本，得到下列数据矩阵：

$$\mathbf{X}^{(1)}=\begin{bmatrix}X_{11}^{(1)} & X_{12}^{(1)} & \cdots & X_{1p}^{(1)}\\ X_{21}^{(1)} & X_{22}^{(1)} & \cdots & X_{2p}^{(1)}\\ \vdots & \vdots & \ddots & \vdots\\ X_{n1}^{(1)} & X_{n2}^{(1)} & \cdots & X_{np}^{(1)}\end{bmatrix}$$

典型相关分析关键步骤

（2）首先要对两组变量的相关性进行检验（构造似然比统计量）。

p值小于0.05（0.1）表示在95%（90%）的置信水平下拒绝原假设，即认为两组变量有关。

- 在利用样本进行两组变量的典型相关分析时，应就两组变量的相关性进行检验。这是因为，如果两个随机向量$\mathbf{X}^{(1)}$、$\mathbf{X}^{(2)}$互不相关，则两组变量协差阵 $Cov(\mathbf{X}^{(1)},\mathbf{X}^{(2)})=0$。但是有可能得到的两组变量的样本协差阵不为零，因此，在用样本数据进行典型相关分析时应就两组变量的协差阵是否为零进行检验。即检验假设

$$H_0:\boldsymbol{\Sigma}_{12}=0,\qquad H_1:\boldsymbol{\Sigma}_{12}\neq 0$$

根据随机向量的检验理论可知，用于检验的似然比统计量为

$$\Lambda_0=\frac{\left|\hat{\boldsymbol{\Sigma}}\right|}{\left|\hat{\boldsymbol{\Sigma}}_{11}\right|\left|\hat{\boldsymbol{\Sigma}}_{22}\right|}=\prod_{i=1}^{r}(1-\hat{\lambda}_i^2)\qquad (9.13)$$

典型相关分析关键步骤

（3）确定典型相关变量的个数（直接看典型相关系数对应的P值即可）

- 若总体典型相关系数 $\lambda_k = 0$，则相应的典型变量 U_k，V_k 之间无相关关系，因此对分析 $\mathbf{X}^{(1)}$ 对 $\mathbf{X}^{(2)}$ 的影响不起作用．这样的典型变量可以不予考虑，于是提出如何根据样本资料来判断总体典型相关系数是否为零，以便确定应该取几个典型变量的问题。巴特莱特（**Bartlett**）提出了一个根据样本数据检验总体典型相关系数 $\lambda_1, \lambda_2, \cdots, \lambda_r$ 是否等于零的方法。检验假设为

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_r = 0$$

$$H_1 : \lambda_{k+1} \neq 0$$

用于检验的似然比统计量为：

$$\Lambda_k = \prod_{i=k+1}^{r} (1 - \hat{\lambda}_i^2) \qquad (9.14)$$

典型相关分析关键步骤

（4）利用标准化后的典型相关变量分析问题

- 典型相关分析涉及多个变量，不同的变量往往具有不同的量纲及不同的数量级别。在进行典型相关分析时，由于典型变量是原始变量的线性组合，具有不同量纲变量的线性组合显然失去了实际意义。其次，不同的数量级别会导致“以大吃小”，即数量级别小的变量的影响会被忽略，从而影响了分析结果的合理性。因此，为了消除量纲和数量级别的影响，必须对数据先做标准化变换处理，然后再做典型相关分析。显然，经标准化变换之后的协差阵就是相关系数矩阵，因而，也即通常应从相关矩阵出发进行典型相关分析。

可得到标准化的第一对典型变量：

$$U_1^* = 0.7751Z_1^{(1)} - 1.579Z_2^{(1)} + 0.059Z_3^{(1)}$$

$$V_1^* = 0.349Z_1^{(2)} + 1.054Z_2^{(2)} - 0.716Z_3^{(2)}$$

其中，$Z_i^{(1)}$ 和 $Z_j^{(2)}$ 分别为原始变量 X_i 和 Y_j 标准化后的结果。

典型相关分析关键步骤

（5）进行典型载荷分析

集合 1 典型载荷

变量	1	2	3
体重x1	-.621	-.772	-.135
腰围x2	-.925	-.378	-.031
脉搏x3	.333	.041	.942

- 以上结果说明生理指标的第一典型变量与体重的相关系数为-0.621，与腰围的相关系数为-0.925，与脉搏的相关系数为0.333。从另一方面说明生理指标的第一对典型变量与体重、腰围负相关，而与脉搏正相关。其中与腰围的相关性最强。第一对典型变量主要反映了体形的胖瘦。

典型相关分析关键步骤

(6)计算前 r个典型变量对样本总方差的贡献

已解释的方差比例

典型变量	集合 1 * 自身	集合 1 * 集合 2	集合 2 * 自身	集合 2 * 集合 1
1	.451	.285	.408	.258
2	.247	.010	.434	.017
3	.302	.002	.157	.001

求得生理指标样本方差由自身 3 个典型变量解释的方差比例分别为：

第一典型变量解释的方差比例$=(0.621^2+0.925^2+0.333^2)/3=0.451$

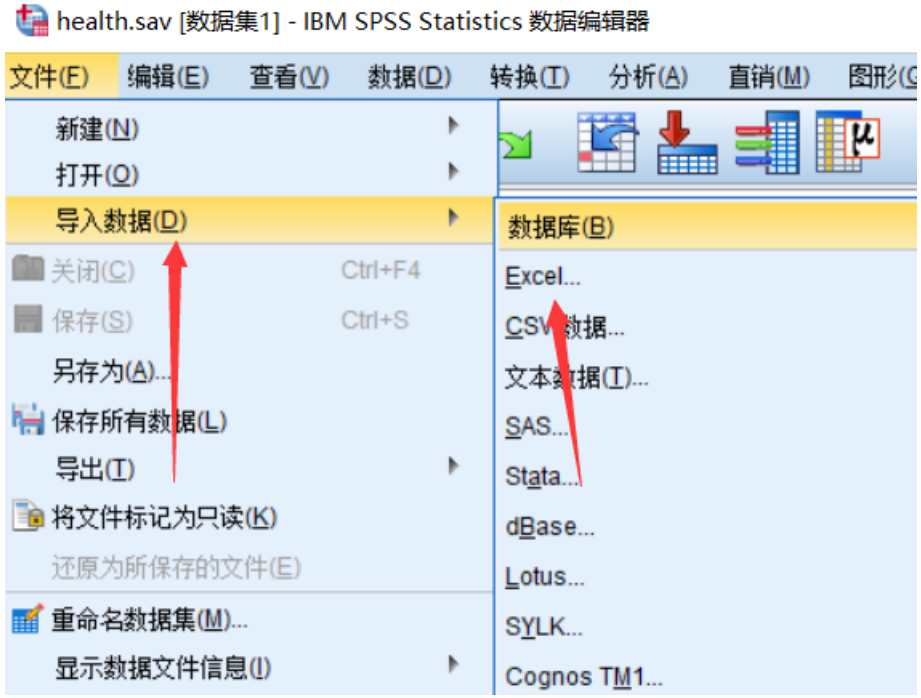
第二典型变量解释的方差比例$=(0.772^2+0.377^2+0.041^2)/3=0.246$

第三典型变量解释的方差比例$=(0.135^2+0.031^2+0.942^2)/3=0.302$

SPSS操作步骤

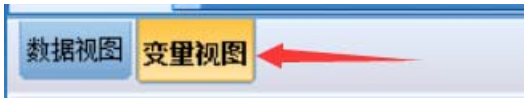
注意：Spss至少需要24版本，低版本不能直接进行典型相关分析的操作，需要编程。（如果新版本仍不能运行，则检查电脑用户名是否为中文，如果是中文的话就需要在电脑上新建一个用户，在新用户上面重新安装SPSS）

第一步：导入数据



SPSS操作步骤

第二步：检验数据的类型

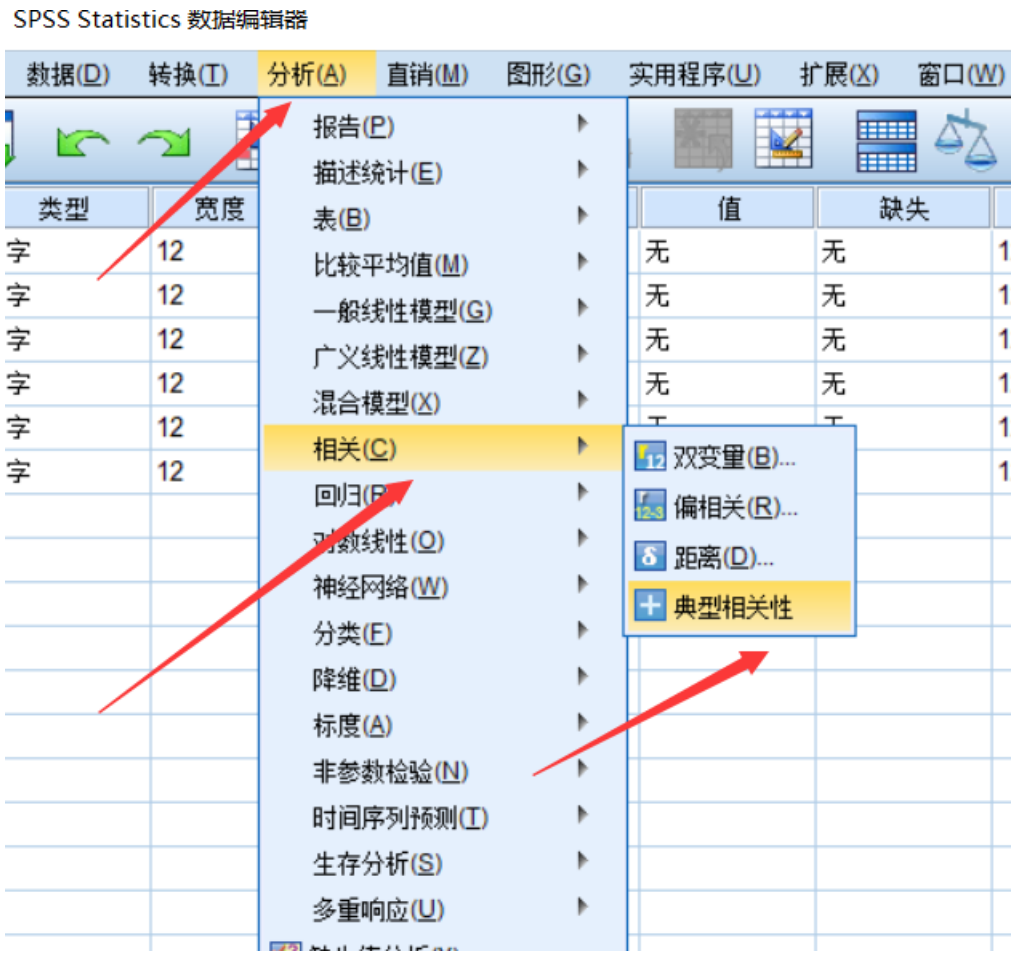


	名称	类型	宽度	小数位数	标签	值	缺失	列	对齐	测量
1	体重x1	数字	12	0	体重(x1)	无	无	12	右	标度
2	腰围x2	数字	12	0	腰围(x2)	无	无	12	右	标度
3	脉搏x3	数字	12	0	脉搏(x3)	无	无	12	右	标度
4	引体向上次...	数字	12	0	引体向上次数(y1)	无	无	12	右	标度
5	起坐次数y2	数字	12	0	起坐次数(y2)	无	无	12	右	标度
6	跳跃次数y3	数字	12	0	跳跃次数(y3)	无	无	12	右	标度
7										
8										

全部设置为标度

SPSS操作步骤

第三步：点击菜单功能

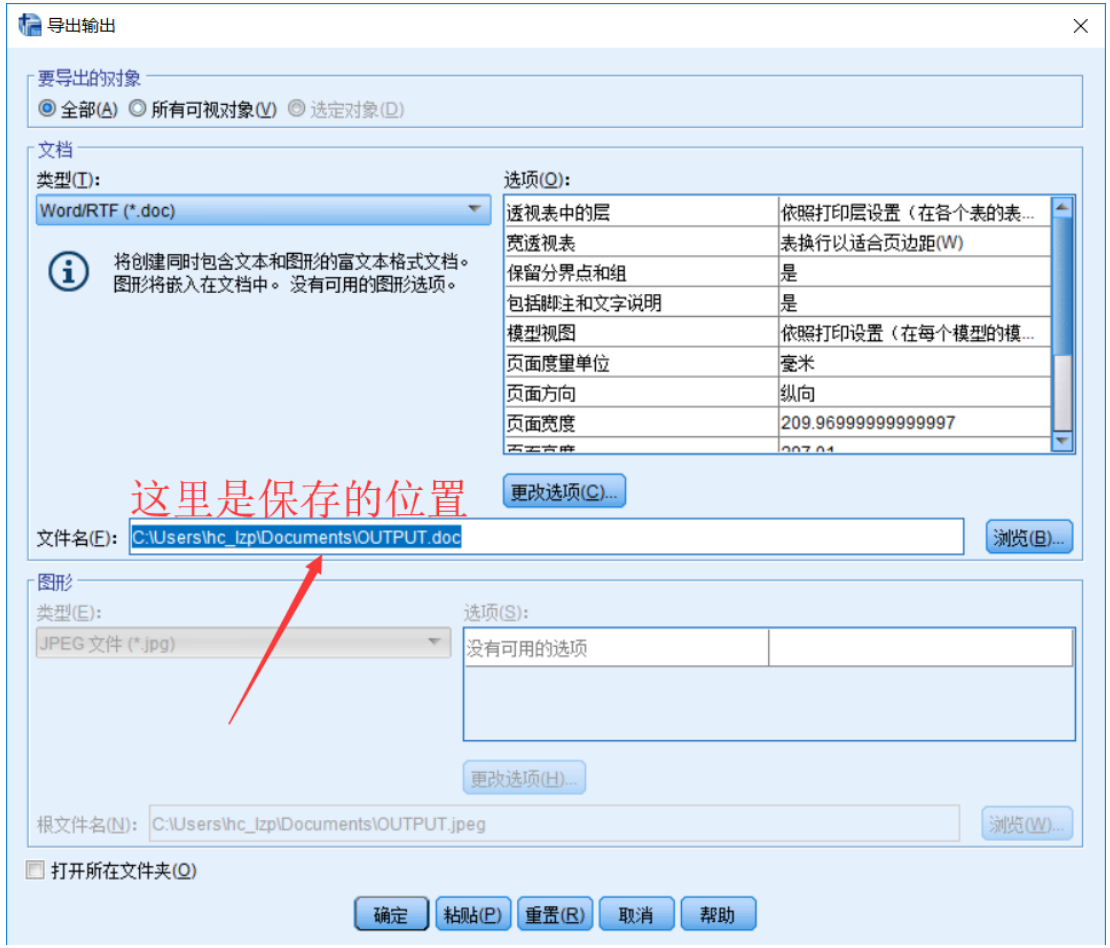


SPSS操作步骤

第四步：将数据移动到对应的集合

SPSS操作步骤

第五步：导出分析结果



SPSS操作步骤

第六步：对结果进行分析

典型相关性

	相关性	特征值	威尔克统计	F	分子自由度	分母自由度	显著性
1	.796	1.725	.350	2.048	9.000	34.223	.064
2	.201	.042	.955	.176	4.000	30.000	.949
3	.073	.005	.995	.085	1.000	16.000	.775

首先看这张表格的最后一列，这一列代表着检验统计量所对应的p值，我们要通过它确定典型相关系数的个数。

SPSS操作步骤

第六步：对结果进行分析

集合 1 标准化典型相关系数

变量	1	2	3
体重x1	.775	-1.884	-.191
腰围x2	-1.579	1.181	.506
脉搏x3	.059	-.231	1.051

集合 2 标准化典型相关系数

变量	1	2	3
引体向上次数y1	.349	-.376	-1.297
起坐次数y2	1.054	.123	1.237
跳跃次数y3	-.716	1.062	-.419

写出标准化后的典型变量

（根据上一步确定个数来写，有几个显著的典型相关性系数就要写几对出来）

可得到标准化的第一对典型变量：

$$U_1^* = 0.7751Z_1^{(1)} - 1.579Z_2^{(1)} + 0.059Z_3^{(1)}$$

$$V_1^* = 0.349Z_1^{(2)} + 1.054Z_2^{(2)} - 0.716Z_3^{(2)}$$

其中，$Z_i^{(1)}$ 和 $Z_j^{(2)}$ 分别为原始变量 X_i 和 Y_j 标准化后的结果。

SPSS操作步骤

第六步：对结果进行分析

典型相关性

	相关性	特征值	威尔克统计	F	分子自由度	分母自由度	显著性
1	.796	1.725	.350	2.048	9.000	34.223	.064
2	.201	.042	.955	.176	4.000	30.000	.949
3	.073	.005	.995	.085	1.000	16.000	.775

可得到标准化的第一对典型变量：

$$U_1^* = 0.7751Z_1^{(1)} - 1.579Z_2^{(1)} + 0.059Z_3^{(1)}$$

$$V_1^* = 0.349Z_1^{(2)} + 1.054Z_2^{(2)} - 0.716Z_3^{(2)}$$

其中，$Z_i^{(1)}$ 和 $Z_j^{(2)}$ 分别为原始变量 X_i 和 Y_j 标准化后的结果。

典型变量每个分量前面的系数代表着重要程度，可结合典型相关系数进行分析。

SPSS操作步骤

第六步：对结果进行分析

后面选择性的分析典型载荷和方差解释程度。

集合 1 典型载荷

变量	1	2	3
体重x1	-.621	-.772	-.135
腰围x2	-.925	-.378	-.031
脉搏x3	.333	.041	.942

已解释的方差比例

典型变量	集合 1 * 自身	集合 1 * 集合 2	集合 2 * 自身	集合 2 * 集合 1
1	.451	.285	.408	.258
2	.247	.010	.434	.017
3	.302	.002	.157	.001

课后作业

我们要探究观众和业内人士对于一些电视节目的观点有什么样的关系呢？
观众评分来自低学历(led)、高学历(hed)和网络(net)调查三种，它们形成第一组变量；而业内人士分评分来自包括演员和导演在内的艺术家(arti)、发行(com)与业内各部门主管(man)三种，形成第二组变量。

利用典型相关分析完成这道题，写一篇小论文。

B站搜索：av50504101

另外，读一下“2012年数学建模A题一等奖论文葡萄酒的评价”这篇文章
（文章和题目在拓展资料中）