**32. 协方差分析**

**（一）原理**

**一、基本思想**

在实际问题中，有些随机因素是很难人为控制的，但它们又会对结果产生显著影响。如果忽略这些因素的影响，则有可能得到不正确的结论。这种影响的变量称为协变量（一般是连续变量）。

例如，研究3种不同的教学方法的教学效果的好坏。检查教学效果是通过学生的考试成绩来反映的，而学生现在考试成绩是受到他们自身知识基础的影响，在考察的时候必须排除这种影响。

协方差分析回归分析与方差分析的结合，在做两组和多组均值之间的比较前，用直线回归的方法找出各组因变量Y与协变量X之间的数量关系，求得在假定X相等时的修正均均值，然后用方差分析比较修正均值之间的差别。

简单来说，协方差分析就是扣除协变量的影响，或者将这些协变量处理成相等，再对修正的Y的均值作方差分析。

根据协变量的个数的不同，协方差分析分为一元协方差分析和多元协方差分析。

**二、协方差分析需要满足的条件**

（1）自变量是分类变量，协变量是定距变量，因变量是连续变量；对连续变量或定距变量的协变量的测量不能有误差；

（2）协变量与因变量之间的关系是线性关系，可以用协变量和因变量的散点图来检验是否违背这一假设；协变量的回归系数（即各回归线的斜率）是相同的，且不等于0，即各组的回归线是非水平的平行线。否则，就有可能犯第一类错误，即错误地接受虚无假设；

（3） 自变量与协变量相互独立，若协方差受自变量的影响，那么协方差分析在检验自变量的效应之前对因变量所作的控制调整将是偏倚的，自变量对因变量的间接效应就会被排除；

（4）各样本来自具有相同方差σ2的正态分布总体，即要求各组方差齐性。

三、基本理论

1. 观测值=均值+分组变量影响+协变量影响+随机误差. 即

 （1）

其中，为所有协变量的平均值。

注：在方差分析中，协变量影响是包含在随机误差中的，在协方差分析中需要分离出来。

用协变量进行修正，得到修正后的*yij*(adj)为



就可以对*yij*(adj)做方差分析了。关键问题是求出回归系数β.

2. 总离差=分组变量离差+协变量离差+随机误差，

（1）计算总离差平方和时，记





总离差平方和：

最终要检验分组自变量对因变量有无显著作用。原假设H0：无显著作用。假设检验是在H0为真条件下进行，可认为ti=0，则



按最小二乘法原理线性回归可得到β的估计值



记修正的总离差平方和（残差平方和）为Tyy(adj)，则

，自由度为n-2

注：为回归平方和，若（回归线为水平线），表示协变量*x*对*y*无作用，用方差分析就可以解决了。

（2）计算组内离差平方和时，记





组内总离差平方和：

根据协方差分析的基本假设：各组内回归系数相等（做协方差分析时需要检验这一点），得到组内回归系数*βw*的估计值



记修正的组内总离差平方和（组内残差平方和）为Eyy(adj), 则

， 自由度为n-k-1

其中，为组内回归平方和，当时，组内总离差平方和认为完全是由随机因素引起的，Eyy(adj)就是随机为误差。这里的是的加权平均值。

（3）计算分组变量离差平方和Byy(adj)，它反映的是各个水平之间的差异。



即，分组变量离差=总离差-协变量离差-随机误差。

于是，就可以进行组间无差异检验了：



3. 因此，在做协方差分析前，需要依次做两个假设检验：

（1）协变量对因变量的影响对与各组来说都是相同的，即各组回归系数相等：;

步骤：

① 先按回归系数相等和不相等分别表示模型





并计算出误差平方和





其中，.

② 计算F值



若F值小于临界值Fα，则说明各组回归系数无显著差异（相等）。

（2）这些相等的回归系数.

即采用一元线性回归的显著性检验，





4. 协方差分析的步骤

（1）检验数据是否满足假设条件：正态分布性、方差齐性、线性相关性、平行性；

（2）检验效应因子的显著性；

（3）估计校正的组均值；

（4）检验校正的组均值之间的差异。

**（二）PROC GLM过程步**

用PROC GLM过程步进行协方差分析，需要注意是不同试验设计时class语句和model语句的写法。设分类变量为A、B，协变量为X，因变量为Y，则有：

（1）单因素k水平设计的协方差分析模型

class A;

model A X ;

（2）随机区组设计的协方差分析模型

class A B;

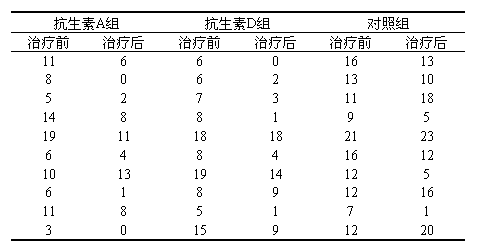
model A B X ;

（3）两因素析因设计的协方差分析模型

class A B;

model A B A\*B X ;

**例1** 为了研究两种药物对癫疯病菌的治疗效果，将30名病人随机分成3组，一组使用抗生素A，一组使用抗生素D，另一组作为对照组使用安慰剂。治疗前和治疗后分别对病人身体的癫疯病菌数量进行了检测，病菌的数量是由每一个病人身体上六个部位病菌感染的程度而定的。试对该试验研究进行统计分析。



这是完全随机设计资料。用x表示治疗前病菌数；y表示治疗后病菌数；drug表示用药方式，取值为’A’，’D’，’F’(安慰剂).

**代码：**

**data** Treatments;

do id=**1** to **10**;

do drug='A', 'D', 'F';

input x y @@;

output;

end;

end;

drop id;

datalines;

11 6 6 0 16 13

8 0 6 2 13 10

5 2 7 3 11 18

14 8 8 1 9 5

19 11 18 18 21 23

6 4 8 4 16 12

10 13 19 14 12 5

6 1 8 9 12 16

11 8 5 1 7 1

3 0 15 9 12 20

;

**run**;

**proc** **sort** data = Treatments;

by drug;

**run**;

**proc** **print** data = Treatments;

**run**;

**proc** **univariate** data = Treatments normal; \*检验正态性;

var y;

by drug;

**run**;

**proc** **discrim** data = Treatments pool=test; \*检验方差齐性;

class drug;

var y;

**run**;

**proc** **reg** data = Treatments; \*检验线性相关性;

model y = x;

by drug;

**run**;

**proc** **glm** data = Treatments; \* 用glm过程，选项drug\*x检验平行性;

class drug;

model y = drug x drug\*x;

**run**;

**proc** **glm** data = Treatments plot = meanplot(cl);

class drug;

model y = drug x/ solution; \*选项solution输出回归系数的估计值及其标准误差和假设检验等;

lsmeans drug / stderr pdiff cov out=adjmeans;

**run**;

**proc** **print** data = adjmeans;

**run**;

**运行结果及说明**（部分）：

| **Obs** | **drug** | **x** | **y** |
| --- | --- | --- | --- |
| **1** | A | 11 | 6 |
| **2** | A | 8 | 0 |
| **3** | A | 5 | 2 |
| **4** | A | 14 | 8 |
| **5** | A | 19 | 11 |
| **6** | A | 6 | 4 |

部分原始数据

drug=A

| **正态性检验** | | | | |
| --- | --- | --- | --- | --- |
| **检验** | **统计量** | | **p 值** | |
| **Shapiro-Wilk** | **W** | 0.924307 | **Pr < W** | 0.3943 |

drug=D

| **正态性检验** | | | | |
| --- | --- | --- | --- | --- |
| **检验** | **统计量** | | **p 值** | |
| **Shapiro-Wilk** | **W** | 0.869627 | **Pr < W** | 0.0990 |

drug=F

| **正态性检验** | | | | |
| --- | --- | --- | --- | --- |
| **检验** | **统计量** | | **p 值** | |
| **Shapiro-Wilk** | **W** | 0.96944 | **Pr < W** | 0.8856 |

正态性检验，样本量<2000，SAS建议采用W检验，三个P值都大于α=0.05，故接受原假设H0，即都满足正态性。

DISCRIM 过程

分类内协方差矩阵的齐性检验

| **卡方** | **自由度** | **Pr > 卡方** |
| --- | --- | --- |
| 1.551005 | 2 | 0.4605 |

用DISCRIM过程步进行方差齐性检验，P值=0.4605>α=0.05, 故接受原假设H0（方差相等），即满足方差齐性。

drug=A

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 1 | 113.34841 | 113.34841 | 11.23 | 0.0101 |

drug=D

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 1 | 283.16532 | 283.16532 | 39.24 | 0.0002 |

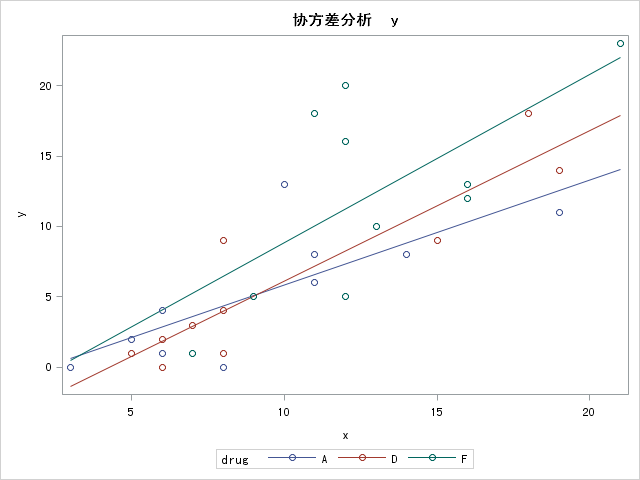
drug=F

| **方差分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| **模型** | 1 | 201.02832 | 201.02832 | 6.21 | 0.0374 |

用REG过程步（线性回归），检验线性相关性，三个P值都小于α=0.05，故拒绝原假设H0(线性无关)，即都线性相关。

| **源** | **自由度** | **I 型 SS** | **均方** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **drug** | 2 | 293.6000000 | 146.8000000 | 8.86 | 0.0013 |
| **x** | 1 | 577.8974030 | 577.8974030 | 34.89 | <.0001 |
| **x\*drug** | 2 | 19.6446451 | 9.8223226 | 0.59 | 0.5606 |

| **源** | **自由度** | **III 型 SS** | **均方** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **drug** | 2 | 8.5025845 | 4.2512923 | 0.26 | 0.7757 |
| **x** | 1 | 564.5675283 | 564.5675283 | 34.08 | <.0001 |
| **x\*drug** | 2 | 19.6446451 | 9.8223226 | 0.59 | 0.5606 |



在GLM过程步的MODEL语句自变量位置加上drug\*x,检验平行性（注意前面要有class drug; 否则会报错）。P值=0.5606>α=0.05, 故接受原假设H0（斜率相等）。

GLM 过程（因变量: y ）

| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **模型** | 3 | 871.497403 | 290.499134 | 18.10 | <.0001 |
| **误差** | 26 | 417.202597 | 16.046254 |  |  |
| **校正合计** | 29 | 1288.700000 |  |  |  |

| **R 方** | **变异系数** | **根 MSE** | **y 均值** |
| --- | --- | --- | --- |
| 0.676261 | 50.70604 | 4.005778 | 7.900000 |

模型的总体检验P值<0.0001<α=0.05，R2=0.676261，说明模型有统计学意义，即x和drug对y的联合作用是显著的。

| **源** | **自由度** | **I 型 SS** | **均方** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **drug** | 2 | 293.6000000 | 146.8000000 | 9.15 | 0.0010 |
| **x** | 1 | 577.8974030 | 577.8974030 | 36.01 | <.0001 |

I型检验结果（只考虑排第一位的drug对y的影响效应，而没有用x进行修正）P值=0.0010<α=0.05, 表明不考虑用药前后的病情情况，三种治疗方法是有显著区别的。

| **源** | **自由度** | **III 型 SS** | **均方** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **drug** | 2 | 68.5537106 | 34.2768553 | 2.14 | 0.1384 |
| **x** | 1 | 577.8974030 | 577.8974030 | 36.01 | <.0001 |

III型检验结果（检验drug对y的影响效应时，考虑其它变量x对y的影响，即用x对y做修正）P值=0.1384>α=0.05, 表明校正了治疗前的病情状况后，这三种治疗方法是没有显著性区别。

| **参数** | **估计值** |  | **标准误差** | **t 值** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | -0.434671164 | B | 2.47135356 | -0.18 | 0.8617 |
| **drug A** | -3.446138280 | B | 1.88678065 | -1.83 | 0.0793 |
| **drug D** | -3.337166948 | B | 1.85386642 | -1.80 | 0.0835 |
| **drug F** | 0.000000000 | B | . | . | . |
| **x** | 0.987183811 |  | 0.16449757 | 6.00 | <.0001 |

参数估计表给出了广义回归模型（GLM）的系数，得到回归方程如下：



最小二乘均值

| **drug** | **y LSMEAN** | **标准误差** | **Pr > |t|** | **LSMEAN 号** |
| --- | --- | --- | --- | --- |
| **A** | 6.7149635 | 1.2884943 | <.0001 | 1 |
| **D** | 6.8239348 | 1.2724690 | <.0001 | 2 |
| **F** | 10.1611017 | 1.3159234 | <.0001 | 3 |

经过协变量*x*修正的均值及标准误，以及T检验（H0: LSMEANS=0）的P值。

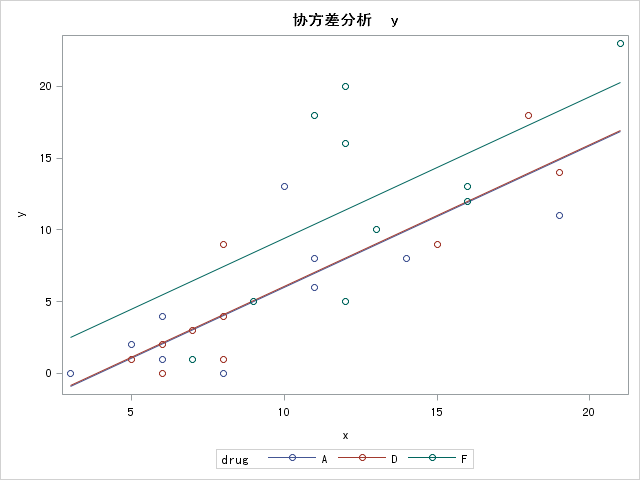
| **效应“drug”的最小二乘均值 Pr > |t|（针对 H0）: LSMean(i)=LSMean(j) 因变量: y** | | | |
| --- | --- | --- | --- |
| **i/j** | **1** | **2** | **3** |
| **1** |  | 0.9521 | 0.0793 |
| **2** | 0.9521 |  | 0.0835 |
| **3** | 0.0793 | 0.0835 |  |

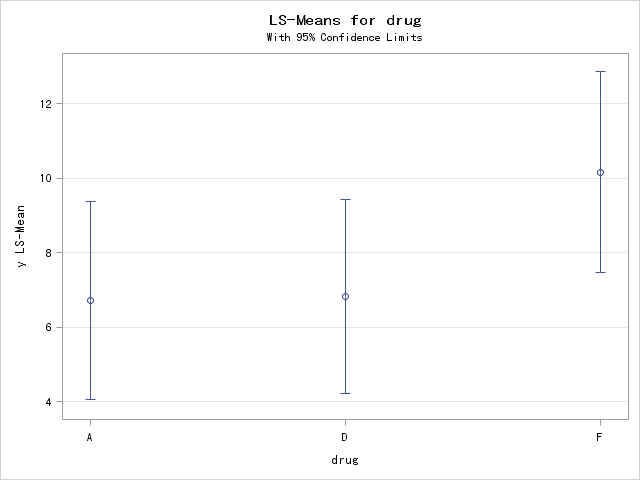
PDIFF选项，输出LSMEANS均值比较的P值。

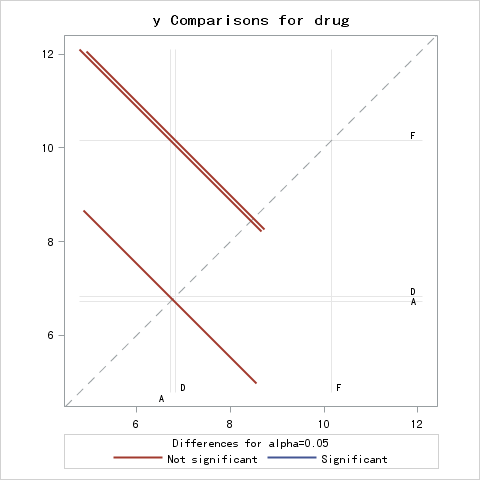
| **Obs** | **\_NAME\_** | **drug** | **LSMEAN** | **STDERR** | **NUMBER** | **COV1** | **COV2** | **COV3** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | y | A | 6.7150 | 1.28849 | 1 | 1.66022 | 0.02844 | -0.08403 |
| **2** | y | D | 6.8239 | 1.27247 | 2 | 0.02844 | 1.61918 | -0.04299 |
| **3** | y | F | 10.1611 | 1.31592 | 3 | -0.08403 | -0.04299 | 1.73165 |

输出LSMEANS均值的估计、标准误、方差、协方差到数据集。

|  |
| --- |
|  |







**例2** **（多元协方差分析）**研究男女儿童的体表面积是否相同。考虑到儿童的身高和体重对表面积可能有影响，在某地测量了男女各15名初生至3周岁儿童的身高、体重和体表面积，得到测量数据如下：

**表26.3 3周岁男女儿童的身高、体重和体表面积**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 男（*male*） | | | 女（*female*） | | |
| 身高（*x*1） | 体重（*x*2） | 表面积（*y*） | 身高（*x*1） | 体重（*x*2） | 表面积（*y*） |
| 54.0 | 3.00 | 2446.2 | 54.0 | 3.00 | 2117.3 |
| 50.5 | 2.25 | 1928.4 | 53.0 | 2.25 | 2200.2 |
| 51.0 | 2.50 | 2094.5 | 51.5 | 2.50 | 1906.2 |
| 56.5 | 3.50 | 2506.7 | 51.0 | 3.00 | 1850.3 |
| 52.0 | 3.00 | 2121.0 | 51.0 | 3.00 | 1632.5 |
| 76.0 | 9.50 | 3845.9 | 77.0 | 7.50 | 3934.0 |
| 80.0 | 9.00 | 4380.8 | 77.0 | 10.0 | 4180.4 |
| 74.0 | 9.50 | 4314.2 | 77.0 | 9.50 | 4246.1 |
| 80.0 | 9.00 | 4078.4 | 74.0 | 9.00 | 3358.8 |
| 76.0 | 8.00 | 4134.5 | 73.0 | 7.50 | 3809.7 |
| 96.0 | 13.5 | 5830.2 | 91.0 | 12.0 | 5358.4 |
| 97.0 | 14.0 | 6013.6 | 91.0 | 13.0 | 5601.7 |
| 99.0 | 16.0 | 6410.6 | 94.0 | 15.0 | 6074.9 |
| 92.0 | 11.0 | 5283.3 | 92.0 | 12.0 | 5299.4 |
| 94.0 | 15.0 | 6101.6 | 91.0 | 12.5 | 5291.5 |

**代码**：

**proc** **format**;

value sexname **1** = 'male'

**2** = 'female';

**data** child;

do i=**1** to **15**;

do sex=**1** to **2**;

input x1 x2 y @@;

format sex sexname.;

output;

end;

end;

datalines;

54.0 3.00 2446.2 54.0 3.00 2117.3

50.5 2.25 1928.4 53.0 2.25 2200.2

51.0 2.50 2094.5 51.5 2.50 1906.2

56.5 3.50 2506.7 51.0 3.00 1850.3

52.0 3.00 2121.0 51.0 3.00 1632.5

76.0 9.50 3845.9 77.0 7.50 3934.0

80.0 9.00 4380.8 77.0 10.0 4180.4

74.0 9.50 4314.2 77.0 9.50 4246.1

80.0 9.00 4078.4 74.0 9.00 3358.8

76.0 8.00 4134.5 73.0 7.50 3809.7

96.0 13.5 5830.2 91.0 12.0 5358.4

97.0 14.0 6013.6 91.0 13.0 5601.7

99.0 16.0 6410.6 94.0 15.0 6074.9

92.0 11.0 5283.3 92.0 12.0 5299.4

94.0 15.0 6101.6 91.0 12.5 5291.5

;

**proc** **glm** data = child;

class sex;

model y=sex x1 x2 /solution;

lsmeans sex /stderr tdiff;

**run**;

**运行结果及说明**（部分）：

| **源** | **自由度** | **平方和** | **均方** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **模型** | 3 | 68523072.11 | 22841024.04 | 557.41 | <.0001 |
| **误差** | 26 | 1065399.76 | 40976.91 |  |  |
| **校正合计** | 29 | 69588471.87 |  |  |  |

| **R 方** | **变异系数** | **根 MSE** | **y 均值** |
| --- | --- | --- | --- |
| 0.984690 | 5.131187 | 202.4276 | 3945.043 |

| **源** | **自由度** | **I 型 SS** | **均方** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **sex** | 1 | 714100.41 | 714100.41 | 17.43 | 0.0003 |
| **x1** | 1 | 67440016.92 | 67440016.92 | 1645.81 | <.0001 |
| **x2** | 1 | 368954.79 | 368954.79 | 9.00 | 0.0059 |

| **源** | **自由度** | **III 型 SS** | **均方** | **F 值** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **sex** | 1 | 139769.3397 | 139769.3397 | 3.41 | 0.0762 |
| **x1** | 1 | 938153.7036 | 938153.7036 | 22.89 | <.0001 |
| **x2** | 1 | 368954.7895 | 368954.7895 | 9.00 | 0.0059 |

两个协变量x1，x2和一个分组变量sex的完全随机化设计。P值和R2值表明模型总体显著。III型检验结果表明，身高、体重对体表面积都有非常显著性的影响（0.0001<0.05，0.0059<0.05），而男、女两性之间无显著性差别（0.0762>0.05）。

| **参数** | **估计值** |  | **标准误差** | **t 值** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | -1118.730592 | B | 497.2296650 | -2.25 | 0.0331 |
| **sex ’female’** | -136.828607 | B | 74.0867551 | -1.85 | 0.0762 |
| **sex ’male’** | 0.000000 | B | . | . | . |
| **x1** | 54.477217 |  | 11.3853803 | 4.78 | <.0001 |
| **x2** | 130.645108 |  | 43.5387744 | 3.00 | 0.0059 |

最小二乘均值

| **sex** | **y LSMEAN** | **标准误差** | **H0:LSMEAN=0** | **H0:LSMean1=LSMean2** | |
| --- | --- | --- | --- | --- | --- |
| **Pr > |t|** | **t 值** | **Pr > |t|** |
| **’female’** | 3876.62903 | 52.32694 | <.0001 | -1.85 | 0.0762 |
| **’male’** | 4013.45764 | 52.32694 | <.0001 |  |  |

选项tdiff输出修正后的均值，及其比较的t值和P值。

