SAS软件应用实习

实验要求：

1. 双变量Logistic回归分析；

2. 多元线性回归与相关分析;

3. 初步了解非参数检验。

第一部分 双变量的曲线回归

对于比较标准的非线性回归模型，可以通过对自变量的转换，将非线性模型转换为线性模型。而对于其他的一些非线性模型，在SAS系统内也可以通过NLIN过程来实现。

表 非线性模型到线性模型的转换

|  |  |  |  |
| --- | --- | --- | --- |
| 模型类型 | 非线性表达式 | 转换 | 线性表达式 |
| 多项式模型 | Y=a+bx+cx2 | Z=x2 | Y=a+bx+cz |
| 对数模型 | Y=a+blnx | Z=lnx | Y=a+bz |
| 指数模型 | Y=aebx | V=lny | v=lna+bx |
| 幂模型 | Y=axb | V=lny; u=lnx | V=lna+bu |
| 倒数模型 | Y=a+b/x | Z=1/x | Y=a+bz |

表4-1-5 越冬棉红铃虫的化蛹进度

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
| z | 3.5 | 6.4 | 14.6 | 31.4 | 45.6 | 60.4 | 75.2 | 90.2 | 95.4 | 97.5 |

程序1：

Data li\_414;

Input x z;

Y=log((100-z)/z);

Cards;

数据块

;

Proc gplot;

Plot z\*x;

Run;

Proc reg data=li\_414;

Model y=x;

Run;

|  |  |
| --- | --- |
| 运行结果： | 拟线性化方程：  Logistic曲线方程：  Logistic曲线标准型：  其中：  data new; a=exp(4.12861); put a; run; |

程序2：

Proc nlin data=li\_414;

Parameters a=50 R=0.1;

Model z=100/(1+a\*exp(-Rx));

Run;

|  |  |
| --- | --- |
| 运行结果： | Logistic曲线方程： |

程序3

Proc nlin data=li\_414;

Parameters B=100 a=50 R=0.1;

Model z=B/(1+a\*exp(-Rx));

Run;

运行结果:

|  |  |
| --- | --- |
|  | 曲线方程为： |

|  |  |
| --- | --- |
| NLIN过程基本语法格式 | 选项说明 |
| **Proc nlin data=数据集[选项];**  **Parameters 参数名=数值;/\*设置参数初始值，对结果有很大影响！\*/**  **Model 因变量=表达式[选项];**  Bounds 表达式；/\*规定参数的合理范围，各约束间用逗号隔开\*/  Der. 参数名.参数名=表达式;/\*给出需要求解的非线性回归模型对参数的一阶或二阶偏导\*/  Id 变量列表;  Output out=数据集[选项];  By 变量列表;  **RUN;** | 选项包括：  data=数据集，指定分析数据集；  outset=数据集名，指定每步迭代结果保存数据集；  best=n, 只输出残差平方和最小的n组，否则输出全部组合的残差平方和；  method=gauss/marquardt/newton/gradient/dud,设置模型参数估计迭代的方法。高斯牛顿法为默认，其余可选麦夸特法，牛顿法，梯度法和试位法。  Maxiter=n: 设置最大迭代次数；Print:显示每步迭代结果；nonprint：禁止结果输出；noinpoint：禁止每部迭代结果输出； |

请大家换用不同的方法进行参数估计并比较结果！

**第二部分 多元线性回归与相关**

已知1978~2006年全国GDP（y），第一产业x1、第二产业x2、工业生产总值x3、第三产业生产总值x4，请建立y对x1~x4的回归模型。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 年份 | y | X1 | X2 | X3 | X4 |
| 1978 | 16.8453 | 5.6092 | 7.4436 | 6.3735 | 3.7925 |
| 1979 | 21.383 | 6.6329 | 10.0585 | 8.7584 | 4.6916 |
| 1980 | 23.1716 | 6.5819 | 11.1424 | 9.6623 | 5.4473 |
| 1981 | 25.7289 | 7.0974 | 12.3181 | 10.6641 | 6.3134 |
| 1982 | 28.6247 | 7.7973 | 13.5101 | 11.288 | 7.3173 |
| 1983 | 32.3103 | 9.1952 | 14.6343 | 12.0864 | 8.4808 |
| 1984 | 36.4037 | 10.0688 | 16.1664 | 12.9822 | 10.1685 |
| 1985 | 45.0774 | 12.0844 | 19.7301 | 15.5838 | 13.2629 |
| 1986 | 51.4749 | 13.1395 | 22.252 | 17.4528 | 16.0834 |
| 1987 | 63.4135 | 17.4543 | 27.5363 | 21.4193 | 21.1229 |
| 1988 | 82.3484 | 19.4304 | 35.6294 | 28.8678 | 27.2886 |
| 1989 | 92.7143 | 21.2031 | 39.3832 | 33.0194 | 32.128 |
| 1990 | 101.4633 | 24.3772 | 40.5796 | 34.3378 | 36.5065 |
| 1991 | 117.4178 | 24.1941 | 48.4759 | 41.0114 | 44.7478 |
| 1992 | 147.5213 | 26.6151 | 62.6834 | 52.2893 | 58.2219 |
| 1993 | 188.8958 | 30.1613 | 82.3853 | 67.8921 | 76.3492 |
| 1994 | 253.0577 | 35.8777 | 111.324 | 91.4335 | 105.856 |
| 1995 | 320.4072 | 45.578 | 137.4362 | 102.6372 | 137.393 |
| 1996 | 397.5701 | 58.3757 | 167.9238 | 130.2389 | 171.2706 |
| 1997 | 475.8691 | 68.732 | 197.5005 | 157.0486 | 209.6366 |
| 1998 | 534.5969 | 70.7519 | 222.8439 | 174.1697 | 241.0011 |
| 1999 | 580.036 | 71.3285 | 238.4684 | 187.0766 | 270.2391 |
| 2000 | 656.4098 | 74.1104 | 268.3988 | 206.0297 | 313.9006 |
| 2001 | 728.0774 | 78.3636 | 297.0933 | 217.9077 | 352.6205 |
| 2002 | 812.8469 | 79.1826 | 328.0378 | 229.521 | 405.6265 |
| 2003 | 929.4858 | 83.2886 | 393.6734 | 268.2806 | 452.5238 |
| 2001 | 1133.8828 | 103.3327 | 504.571 | 341.5303 | 525.9791 |
| 2005 | 1519.90 | 112.59 | 655.27 | 469.28 | 752.04 |
| 2006 | 1790.66 | 123.25 | 774.66 | 584.41 | 892.75 |

来源（博文中对分析结果进行了解读）：http://blog.csdn.net/archielau/article/details/7972522

多元非线性回归方程通常转化为线性回归方程。转化时应首先选择适合的非线性回归形式，并将其线性化。对于实际问题，首先应对原始数据进行作图或通过观察，选择适当函数进行拟合。（以下程序运行正常）：

data ex;

input y x1-x4@@;

y1=log(y);z1=log(x1);z2=log(x2);z3=log(x3);z4=log(x4);/\*对数据做变化，取对数后再做回归分析\*/

cards;

[y x1-x4数据块]

;

proc reg;/\*reg调用回归模块\*/

model y1=z1 z2 z3 z4/cli;/\*表示以z1 z2 z3 z4为自变量，y1为应变量建立回归模型，/cli表示要求预测区间。加入/selection=stepwise时，表示逐步回归 \*/

proc corr; var z1-z4;/\*求相关系数矩阵\*/

run;

**第三部分 非参数检验**

非参数统计是统计分析的重要组成部分。可是与之很不相称的是它的理论发展远远不及参数检验完善，因而比较完善的可供使用的方法也不多**。在SAS中，非参数统计主要由UNIVARIATE过程、MEANS过程和NPAR1WAY过程来实现**，前两者可以进行配对设计差值的符号秩和检验（WILCOXON配对法）；后者是一个单因素的非参数方差分析过程，可进行成组设计的两样本（WILCOXON法）或多样本比较（KRUSKAL-WALLIS法）的秩和检验。下面将主要介绍NPAR1WAY过程。

由于在理论上还有争议，作为权威性的统计软件，SAS不提供非参检验两两比较的方法。据我所知，其余统计软件里也只有PEMS提供这一功能（医统·医百的配套软件，而非参两两比较是写入了该书的）。如果你需要这一结果，那么恐怕只有手算了。

语法格式

|  |  |
| --- | --- |
| PROC NPAR1WAY [DATA=<数据集名> [选项] ] ;  CLASS <处理因素变量名>;  EXACT <关键字>;  OUTPUT <OUT=数据集名> <选项>;  VAR <结果变量名>;  BY <变量名列>; | 必需，指定要分析的处理因素  要求程序在必要时计算确切概率  指定统计结果的输出数据集  指定要分析的应变量  统计按指定的变量分组进行 |

【过程选项】NPAR1WAY过程常用的选项有：

MISSING 将缺失值也用于统计分析

ANOVA 同时进行方差分析

MEDIAN 要求进行中位数检验

NOPRINT 禁止统计结果在OUTPUT视窗内输出

SAVAGE 要求对样本进行SAVAGE得分分析

WILCOXON 要求进行WILCOXON秩和检验

NPAR1WAY过程不能处理按频数输入的资料。这意味着如果你的数据是以频数方式输入的，那么除非你将资料想办法转换成按例记录的资料，否则SAS无法处理。

“NPAR1WAY”可以这样来记：“NPAR”即“非参”的英文缩写，“WAY”是维数，更明确的说是因素的意思，而“1WAY”就代表一个因素，合起来“NPAR1WAY”说的是“单因素的非参数检验”。

我们常用的秩和检验就是WILCOXON秩和检验，对于其它方法，有兴趣的读者可参阅有关统计书籍。在省略所有选项的情况下，SAS系统默认输出所有的统计结果，这恰恰说明了非参数检验方法的不完善。如果你无法判断用那个结果，那么只看Wilcoxon秩和检验的分析结果就够了。输出结果中Z检验就是我们非常熟悉的u检验，只不过是国内外的叫法不同罢了。

两个独立样本的非参数检验（秩和测验）

两独立样本的非参数检验是在对总体分布不了解的情况下，通过分析样本数据，推荐样本来自的两个独立总体的分布是否存在显著差异，一般来说是推断两个独立总体的均值或中位数是否存在显著差异。关于样本是否为独立的，主要看在一个总体中抽取样本对在另一个总体中抽取样本有无影响。如果没有影响，则可以认为这两个总体是独立的。

零假设H0：样本来自的两独立总体的分布没有显著差异

检验方法有多种：

（1）两独立样本的Wilcoxon秩和检验，也成为Mann-Witeney U检验。该检验主要是通过对平均秩的研究来实现推断的。基本思路是：首先，将两组样本数据（X1,X2,…,Xm）和（y1,y2,…,yn）混合并按升序排序（m和n分别位两组样本的样本容量），求出每个数据各自的秩Ri；然后，分别对（X1,X2,…,Xm）和（y1,y2,…,yn）的秩求平均，得到两个平均秩Wx/m和Wy/n，如果这两个平均秩相差甚远，则倾向于拒绝零假设。

（2）两独立样本的K-S检验（Kolmogorov-Smirnov Z），该检验首先将两组样本混合并按升序排序，然后，分别计算两组样本秩的累计频数和每个点上得到累计频率；最后，将两个累计频率相减，得到差值序列数据。

（3）Wlad-Wolfwitz runs，该检验将两组样本混合并按升序排序，在数据排序的同时，两组样本的每个观察值对应的样本组标志值序列也随之重新排列；然后，对这个标志值序列求游程。如果样本所属的两总体的分布形态存在较大差距，那么计算出的游程数会相对比较小。如果游程数比较大，则应该是由于两样本数据充分混合的结果，那么它们的分布应该不存在显著差异。

【例】下表为来自两个样本的测量数据，经检验知两样本方差不齐，试做非参数比较量数据的差别。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A组 | 7 | 14 | 22 | 36 | 40 | 48 | 63 | 98 |
| B组 | 3 | 5 | 6 | 10 | 17 | 18 | 20 | 39 |

SAS程序如下：

Data sasuser;

Do g=1 to 2;

Input x@@; output;

End;

Datalines;

7 3 14 5 22 6 36 10 40 17 48 18 63 20 98 39

;

Proc npar1way wilcoxon;

Class g;

Var x;

Run;

SAS给出两组数据的基本信息（样本量，秩和等）；给出在零假设下各组统计量的期望值及标准差，最后还给出近似Z检验以及近似t检验所得的统计量和所对应的单、双侧概率值。另外，在默认状态下，SAS同时给出Kruskal-Wallis检验结果。

所不同的是，在两样本容量相同时，SAS以秩和较大者作为对象统计量进行概率值的计算，而非医学统计学教材中所说的以较小秩和为对象统计量。在两样本容量相同时，SAS以样本量较小组的秩和为对象统计量，这一点与教材相同。

实例：

**data** soybean;

group=**1**;

if \_n\_>**7** then group=**2**;

input value@@;

cards;

0.82 **0.87** **0.97** **1.21** **1.64** **2.08** **2.13**

0.24 **0.24** **0.29** **0.33** **0.44** **0.58** **0.63**

0.72 **0.87** **1.01**

;

**proc** **gchart**;

vbar value / group=group;

**proc** **npar1way** data=soybean WILCOXON ;

class group;

var value;

**run**;

多个样本的非参数检验的两两比较

由于各个样本的总体呈偏态分布或方差不齐,无法使用方差分析中的两两比较，检验其总体分布是否相同，常用的非参数秩和检验方法是Kruskal-Wallis 法,在SAS 软件中实现的过程步有以下三种方法: NPAR1WAY 过程、FREQ过程以及RAN K 和ANOVA 两过程的结合，而样本间两两比较则可以通过RAN K 和ANOVA 两过程的结合,采用MEANS 语句来实现，即先对原始数据进行排序，求相应的秩次，然后对秩进行参数的多重比较。

调用FREQ 过程产生的第二个CMH统计量” row mean scores differ” (Kruskal-Wallis结果)、NPAR1WAY 过程产生的卡方统计量以及ANOVA 过程产生的R-Square 与υ总(总自由度)之积,即为Kruskal-Wallis 检验结果。此处ANOVA 过程步中的MEANS 规定采用LSD 法进行两两比较，也可使用其他方法。

data sample ;

do group = 1 to 3 ;

input x @@;

output ;

end ;

cards ;

9.8 0.6 0.4 10.2 1.2 1.9 10.6 2.0 2.2 13.0

2.4 2.5 14.0 3.1 2.8 14.8 4.1 3.1 15.6 5.0

3.7 15.6 5.9 3.9 21.6 7.4 4.6 24.0 13.6 7.0

;

proc freq ;

tables group\*x/ scores = rank cmh2 noprint;

run;

proc npar1way wilcoxon;

class group;

var x;

run;

proc rank data =sample out = a ;

var x;

ranks r;

proc anova;

class group ;

model r = group ;

means group/ lsd snk ;

run;

quit;

成组设计的等级资料或频数表资料多个样本及其两两比较

这种类型的资料为成组设计的等级资料或频数表资料,但频数表资料时各个样本的总体呈偏态分布或方差不齐,检验其总体分布是否相同，常用的非参数秩和检验方法也是Kruskal-Wallis 法,在SAS 软件中实现的过程步同上,程序略有差别。

data sample2;

do effect = 1 to 4 ;

do group = 1 to 3 ;

input f @@;

output ;

end ;

end ;

cards ;

10 24 48 17 41 65 19 33 36 4 7 8

;

proc freq;

tables group\*effect/ scores = rank cmh2 noprint;

weight f;

run;

data a;

set sample2;

do i = 1 to f ;

output;

end ;

run;

proc npar1way wilcoxon data=a;

var effect;

class group;

run;

proc rank data=a out=b;

var effect ;

ranks r;

run;

proc glm;

class group;

model r = group ;

means group/lsd snk tukey dunnett ;

run;

quit;

随机区组设计的定量资料多个样本及其两两比较

各个样本的总体呈偏态分布或方差不齐,检验其总体分布是否相同,常用的非参数秩和检验方法是Friedman 法,在SAS 软件中实现的过程步有以下两种方法: FREQ 过程以及RAN K和GLM 两过程的结合。各个样本两两比较可通过RAN K 和GLM 两过程的结合, 采用LSMEANS 语句来实现

调用FREQ 过程产生的第二个CMH统计量和GLM 过程产生的R2 与(υ总-υ区组) 之积,即

为Friedman 检验结果。LSMEANS 语句规定对调整后的样本均数进行两两比较。

data sample3;

do block = 1 to 7 ;

do treat = 1 to 4 ;

input x @@;

output;

end;

end ;

cards ;

63 188 138 54 90 238 220 144 54

300 83 92 45 140 213 100 54 175

150 36 72 300 163 90 64 207 185 87

;

proc freq ;

tables block\*treat\*x/ scores = rank cmh2 noprint;

run;

proc sort data=sample3 out=a;

by block ;

run;

proc rank data=a out=b;

var x;

by block;

ranks r ;

run;

proc glm data=b ;

class treat block ;

model r = treat block/ ss1 ;

lsmeans treat/ stderr pdiff ;

run;

quit;

作业：

Page 229：10，13，14