

# 1. SAS语言与数据管理

## 1.1~1.2 SAS系统、SAS对文件管理

### SAS文件与逻辑库

- SAS逻辑库的文件为两级命名方式：`sas_libname.SAS_filename`
  - 例如：`sashelp.class` 表示sashelp逻辑库中的class数据集
- SAS逻辑库的创建：
  - 例如：`LIBNAME saslib 'd:\sas';`
  - 命名规则：
    - 不能超过8字符
    - 必须以字母或下划线开头
    - 其余字符必须是字母/数字/下划线
- 如果没有libname则默认为work逻辑库
- 浏览一个逻辑库中的所有文件

```
PROC CONTENTS DATA=libref._ALL_ NODS;  
RUN;
```

### 浏览SAS数据集

```
PROC CONTENTS DATA=SASDataSet;  
RUN;
```

### PDV

### SAS变量的长度

- 字符型变量
  - 包含任意值：字母、数字、特殊字符、空格
  - 如果连续赋值，则长度未第一个字符串长度，例如
- 数值型变量
  - 默认8字节浮点（只要未特殊声明，就按8字节处理）
  - 数值型的缺失值为小数点 `.`
- 日期型变量

```
x="ab";  
x="abcd";
```

此时x长度为2

- 如果字符串连接，则为长度之和
- 字符型的缺失值为空格

- SAS的日期为某日期到1960年1月1日的天数

## SAS变量的命名

- 要求：
  1. 1~32字符
  2. 必须以字母/下划线开头，后面的字符可以是字母/下划线/数字
  3. 可以大小写混合
  4. 不区分大小写

## PROC PRINT 浏览数据

- 语法：

```
PROC PRINT DATA=SASDataSet <NOOBS>;  
    VAR variables;  
RUN;
```

- 说明
  - NOOBS 去掉了观测数一行
  - VAR 语句可以选择显示的变量名称及顺序

## SAS的注释

```
/* comment */
```

```
* comment
```

## 1.3 生成SAS数据集

- 生成方法
  1. viewtable 直接输入
  2. DATA 步输入
  3. import wizard输入
  4. SAS/ACCESS

## FORMAT语句

- `FORMAT variable(s) format;`
- 在DATA步使用FORMAT，则格式将与变量永久关联

## SAS格式

- 标准形式： `<$> format<w>.<d>`
  - \$：若为字符型，则需要加\$
  - format：可以是SAS给定的格式名称，也可以是自定义的
  - w：格式的总的最大宽度，包括小数点和其他特殊字符
  - .：一定要有的元素
  - d：指定数值型变量的小数点位数
- 常用的一些格式：
  - \$w.：标准字符型数据

- `w.d` : 标准数值型数据
- `COMMAw.d` : 输入数据每三位一个逗号, 且有小数点
- `COMMAXw.d` : 输入数据每三位一个点作分割, 小数点为逗号
- `DOLLARw.d` : 输入的数据最左侧有一个美元, 且每三位有一个逗号, 且有小数点
- `EUROXw.d` : 输入的数据最左侧是欧元符号, 每三位用点分割, 小数点为逗号
- 常用的一些日期格式 (例1960年12月31日) :
  - `MMDDYY6.` : 123160
  - `MMDDYY8.` : 12/31/60
  - `MMDDYY10.` : 12/31/1960
  - `DDMMYY6.` : 311260
  - `DDMMYY8.` : 31/12/60
  - `DDMMYY10.` : 31/12/1960
  - `DATE7.` : 31DEC60
  - `DATE9.` : 31DEC1960
  - `WORDDATE.` : January 1, 1960
  - `WEEKDATE.` : Friday, January 1, 1960
  - `MONYY7.` : JAN1960
  - `YEAR4.` : 1960

## CARDS 直接输入变量

```
DATA SASDataSet;
  INPUT x y z;
  CARDS;
    1 2 3
    4 5 6
    7 8 9
  ;
RUN;
```

## SAS读取文件

```
DATA SASDataSet;
  INFILE 'c:\file.txt' <FIRSTOBS = 3>
                        <OBS=5>
                        <DLM=', '>
                        <MISSOVER>;

  INPUT x y z;
RUN;
```

- `FIRSTOBS=` 表示从第几行开始读入
- `OBS=` 表示读完第几行后就停止读入
- `DLM=', '` 表示分隔符为逗号
- `MISSOVER` 表示如果读入的数据有缺失值不要当作是一行输入完成直接跳转, 而是记成缺失值

## 1.4 SAS语言

# SAS 大小写

- SAS的语句通常不区分大小写，但字符型数据的值是区分的，例如“Beijing”与“BEIJING”SAS认为是不同的数值

## SAS的运算符

- 常用运算符：
  - + - \* \ \*\*
  - = EQ
  - ^= NE
  - > GT
  - < LT
  - >= GE
  - <= LE
  - IN
    - 例如 `city IN ('Beijing','Shanghai','NewYork')` 可用来判断city变量中是否含有上述三城市之一
  - & AND (与)
    - 例如 `(salary >= 1000) AND (salary<2000)`
    - 上表达式也可写成 `1000<= salary <2000`
  - | OR (或)
  - ^ NOT (非)
  - || 两个字符串连接
  - <> 两个数值取大
  - >< 两个数值取小

## SAS 常用内建函数

- 常用函数
  - Arithmetic: ABS, SQRT, DIM
  - Character: UPCASE, SUBSTR, TRIM
  - Date and Time: TODAY, DAY, MONTH, MDY
  - Mathematical: LOG, EXP, GAMMA
  - Noncentrality: CNONCT, FNONCT, TNONCT
  - Quantile: PROBIT, CINV, TINV, FINV
  - Probability and Density: PROBNORM, PROBT, POISSON,PDF,PDM
  - Random Number: RANUNI, RANNOR, RANEXP
  - Sample Statistic: SUM, MEAN, STD, VAR, RANGE
  - Special: PUT, INPUT, DIF, LAG
  - Trigonometric: SIN, TAN, ARCOS
  - Truncation: INT, CEIL, ROUND,FLOOR
  - Others: ZIPSTATE
- 统计函数
  - `value=CDF('分布',x,<参数表>)` 求分布函数x处概率

- `value=PDF('分布',x,<参数表>)` 求连续型函数x处密度
- `value=PMF('分布',x,<参数表>)` 求离散型x处分布概率
- `value=LOGPDF('分布',x,<参数表>)`
- `value=LOGPMF('分布',x,<参数表>)`
- 其中'分布'可选的选项有：
  - BERNOULI 伯努利
  - BETA
  - NORMAL 或 GAUSSIAN 正态
  - CAUCHY 柯西分布
  - CHICQUARED 卡方
  - GAMMA
  - F
  - EXPONENTIAL 指数分布
  - GEOMETRIC 几何分布
  - HYPERGEOMETRIC 超几何分布
  - LAPLACE 拉普拉斯分布
  - LOGISTIC 逻辑分布
  - LOGNORMAL 对数正态
  - POISSON 泊松分布
  - NEGBINOMIAL 负二项分布
  - PARETO 帕累托分布
  - T
  - UNIFORM 均匀分布
  - WALD 或 IGAUSS
  - WEIBULL 韦布尔
- 例如：
  - `PDF('NORMAL',1.96)`
- 分布函数
  - `PROBNORM(x)` 计算正态分布x处CDF值
  - `PROBCHI(x,df,nc)` 计算自由度为df, 非中心参数为nc的卡方分布在x处的CDF
  - `PROBGAM(x,a)` 计算形状参数为a的伽马分布在x的CDF
  - `POISSON(lambda,a)` 计算lambda的泊松分布的x处分布函数
  - `PROBNEGB(p,n,m)` 计算负二项分布概率分布
  - `PROBHYPER(nn,k,n,x,r)` 计算超几何分布, 其中有nn个产品, k个次品, 抽取n个样品, r是抽到不合格品率是合格品率的倍数
  - `PROBBNRM(x,y,r)` 计算标准二元正态的分布函数, r为相关系数
- 分位数函数
  - `CINV(p,df,nc)` 卡方分布分位数
  - `BETAINV(p,a,b)`
  - `FINV(p,ndf,ddf,nc)` F分布分位数
  - `TINV(p,df,nc)` T分布分位数
  - `PROBIT(p)` 计算标准正态分布分位数
  - `GAMINV(p,a)` 计算gamma分布的分位数
- 样本统计函数
  - `MEAN()`
  - `MAX()`

- MIN()
- N() 非缺失值个数
- NMISS() 缺失值个数
- SUM()
- VAR()
- STD()
- STDERR() 标准误
- CV() 变异系数
  - 在概率论和统计学中，变异系数，又称“离散系数”（英文：coefficient of variation），是概率分布离散程度的一个归一化量度，其定义为标准差与平均值之比。变异系数（coefficient of variation）只在平均值不为零时有定义，而且一般适用于平均值大于零的情况。变异系数也被称为标准离差率或单位风险。
- RANGE() 极差
- CSS() 偏差平方和/矫正平方和:  $\sum_i (x_i - \bar{x})^2$
- USS() 未矫正平方和:  $\sum_i x_i^2$
- SKEWNESS() 偏度
- KURTOSIS() 峰度
- ORDINAL(k,x1,x2,...) 返回一系列x中第k小的数值
- 随机数函数
  - UNIFORM(seed) RANUNI(seed) 均匀分布
  - NORMAL(seed) RANNOR(seed) 正态分布
    - 若  $x \sim N(0, 1)$ ,  $y = sx + u$ ,  $z = e^y$ , 则  
 $y \sim N(u, s^2)$ ,  $z \sim \log N(e^{u+s/2}, e^{(2u+s)} * (e^s - 1))$
  - RANEXP(seed) 指数分布
  - RANGAM(seed,alpha)
  - RANTRI(seed,h)
  - RANCAU(seed)
  - RANBIN(seed)
  - RANPOI(seed,lambda)
  - RANTBL(seed,p1,p2,...)
- 其他函数
  - CEIL(x) 向上取整
  - FLOOR(x) 向下取整
  - INT(x) 截断取整
  - ROUND(x) 四舍五入取整
  - ABS(x) 求绝对值
  - MOD(x,y) 求x/y的余数
  - SQRT(x)

- DIGAMMA(x)
- GAMMA(x)
- LOG(x) ln
- LOG2(x)
- LOG10(x)
- ERF(x) 误差函数
- EXP(x) exp(x)
- SIN(x), COS(x), TAN(x), SINH(x), COSH(x), TANH(x)
- UPCASE(x) 转换为大写
- LOWCASE(x) 转换为小写
- SUBSTR(s,p,n) 从字符串s中第p个字符开始取n个字符子串
- LAGn(x) DIFn(x)

### LAG & DIF

	Obs	num	quality	lag0	lag	lag1	lag2	dif0	dif	dif1	dif2
	1	1	112	112	.	.	.	0	.	.	.
■	2	2	118	118	112	112	.	0	6	6	.
	3	3	132	132	118	118	112	0	14	14	20
	4	4	118	118	132	132	118	0	-14	-14	0
	5	5	126	126	118	118	132	0	8	8	-6
	6	6	141	141	126	126	118	0	15	15	23

- 日期函数
  - 日期的直接输入：记得引号和d '12JAN96'd
  - DATE() 取今天日期
  - TODAY() 取今天日期
  - DATETIME() 取今天日期+时间
  - TIME() 取今天时间
  - YEAR() 获取年
  - MONTH() 获取月
  - DAY() 获取日
  - WEEKDAY() 获取星期（周日=1，周六=7）
  - QTR(date) 由日期得到季度
  - MDY(month,day,year) 生成日期
  - HMS(hour,minute,second) 生成时间
  - DHMS(date,h,m,s) 生成日期+时间
  - DATEPART(datetime) 取日期
  - INTNX(interval,from,n) 计算从from开始，经过n个间隔后的日期，interval 可选：
    - YEAR, QTR, MONTH, WEEK, DAY

- `INTCK(interval, from, to)` 计算从from开始到to中间的间隔个数
  - 注意, 例如 `INTCK('YEAR', '31Dec1996'd, '1Jan1998'd)` 得到的结果=2

## 1.5 DATA 步入门

- data的循环范例

```
data a ;  
  put y= z= x= ;  
  input x y ;  
  z=x+y ;  
  put x= y= z= ;  
  cards ;  
  10 20  
  100 200  
  ;  
run;
```

日志:

y=. z=. x=.

x=10 y=20 z=30

y=. z=. x=.

x=100 y=200 z=300

y=. z=. x=.

- 从这个例子可以看出SAS数据步程序和普通程序的一个重大区别: SAS数据步如果有数据输入,比如用INPUT、SET、MERGE、UPDATE、MODIFY等语句读入数据,则数据步中隐含了一个循环,即数据步程序执行到最后一个语句后,会返回到数据步内的第一个可执行语句开始继续执行,直到读入数据语句 (INPUT、SET、MERGE、UPDATE、MODIFY等) 读入了数据结束标志为止才停止执行数据步,并把读入的各个观测写入在DATA语句中指定的数据集内。如果没有数据输入而只是直接计算,则数据步程序不需要此隐含循环
- 数据步因为有这样一个隐含循环,所以也提供了用来查询某一步是第几次循环的特殊变量`N`,它的值为数据步循环计数值。

## 1.6 创建SAS数据集

### 读入外部SAS数据集

- 将想要读入数据集的文件夹设置为逻辑库, 后直接调用即可

```
LIBNAME ...;
```



## 读入excel

```
PROC IMPORT OUT= WORK.DEMO
    DATAFILE= "C:\Users\\RepData2.xlsx"
    DBMS=EXCEL REPLACE;
    RANGE="Sheet1$";
    GETNAMES=YES;
    MIXED=NO;
    SCANTEXT=YES;
    USEDATE=YES;
    SCANTIME=YES;
RUN;
```

## 读入csv, txt等

```
DATA SAS-data-set ;
    变量属性设定语句;
    INFILE filename
    <FIRSTOBS=n1
    OBS=n2>;
    INPUT 语句;
    其它语句;
RUN;
```

例如

```
data da2;
    infile 'c:\ch02_15.txt';
    input x y z;
    mean=(x+y+z)/3;
run;
```

## INPUT的读取规则

- 见 Little SAS Book

## SAS 数据集的属性变量

- 可以通过 DATA ATTRIB 给各个变量的属性进行控制，也可以单独设置

```

data sales;
ATTRIB name LABEL="姓名" LENGTH=$6
        date LABEL="日期" FORMAT=yymmdd10.
        INFORMAT=mmddy8.
        amount LABEL="金额" FORMAT=10.2 ;
input name $ 1-6 date amount;
cards ;
王永成 10/15/98 17665
李宏志 1/3/1999 18178
贺佳 11/5/99 16254
;
proc print data=sales noobs label;
run;

```

- 可以设置的相关属性
  - LENGTH 变量名 <\$> 长度
  - INFORMAT 变量名 输入格式...
  - FORMAT 变量名 输出格式...
  - LABEL 变量名=字符串输入格式...
    - label的作用是给变量的名称的外观进行改变
    - format 的作用是给变量的值的外观进行改变
    - PROC PRINT 语句中若想要包括label需要单独说明（如上所示）

## 1.7 加工SAS数据集

### DATA SET

- SET 语句从一个SAS数据集中读入观测，并在DATA步中进一步处理

### 选择变量

- KEEP=
- DROP=

### 筛选数据

- IF... DELETE 语句

```

data a;
set old;
if sex='M' then delete;
run;

```

- WHERE 语句

- ```
LIBNAME libref 'SAS-data-library';
DATA output-SAS-data-set;
  SET input-SAS-data-set;
  WHERE where-expression;
  KEEP variable-list;
  LABEL variable = 'label'
        variable = 'label'
        variable = 'label';
  FORMAT variable(s) format;
RUN;
```

- where 配套的一些逻辑运算符

- BETWEEN ... AND
- IS NULL
- IS MISSING
- CONTAINS
  - where Job\_Title contains 'Rep';
  - 将查找是否含有字符串，位置不重要，区分大小写
- LIKE
  - 给定一个模式，看是否能模糊匹配
  - % 代替任意数量字符
    - where Name like '%N';
  - \_ 代替一个字符
    - where Name like 'T\_M%';

## 数据集的其他操作

- 累加求和

- 例：C++中的 `a=a+1` 在SAS中为： `a+1`

- RETAIN

- DATA步正常是包含循环的，RETAIN 语句让SAS在循环时保留该变量；
- 利用 RETAIN 即可完成 LAG 的操作

- ```
data d;
  input num score @@;
  datalines;
  1 12 2 13 3 14 9 11
  4 15 5 16 6 17 7 18
  ;
run;

data dd;
  format num score ret;
  ret=score;
  set d;
  retain score;
run;
```

```
proc print data=dd noobs;
run;
```

num	score	ret
1	12	.
2	13	12
3	14	13
9	11	14
4	15	11
5	16	15
6	17	16
7	18	17

- **PUT**
  - **PUT** 语句可以直接在日志窗口运行到某一行时就进行输出
  - 语法为 **PUT '辅助文字' 变量名;**
  - 不能输出数值常量或表达式
- **OUTPUT**
  - 作用是将当前的观测立刻输出到数据集中
  - 例如

```
data d;
  do k=1 to 50;
    OUTPUT;
  end;
run ;
proc print data=d noobs;
run;
```

若没有output, 则最终结果是最后一次循环的51  
 若有output, 则生成了从1到50的全部内容;

- 日历

```
data raw;
  format date date5.;
  do dd='01Jan2022'd to '31Dec2022'd;
    date=dd;
    output;
  end;
run;

data cal;
  set raw;
  array wday(7) Sun Mon Tue Wed Thu Fri Sat ;
  retain Sun Mon Tue Wed Thu Fri Sat;
  format Sun Mon Tue Wed Thu Fri Sat date7.;
  wday(weekday(date))=date;
  if weekday (date)=7 or day(date+1)=1 then do;
```

```

output; /*output的内容要压着七天才能output一次，这样才能保证retain的内容全部覆盖*/
call missing (of wday{*}); /*call missing(of ...)是将内容转换为缺失值*/
end;
run;

proc print data=cal;
run;

```

- SAS的另一个作用是输出到不同的数据集中

```

data classm classf;
set sasuser.class;
if sex='M' then output classm;
if sex='F' then output classf;
run;

```

- IF

- IF ... THEN DO;
 

```

      ...
      END;
      ELSE DO;
      ...
      END;
      
```

- SELECT

- SELECT(month);
 

```

      WHEN ('Feb', 'Mar', 'Apr') put '春天';
      WHEN ('May', 'Jun', 'Jul') put '夏天';
      OTHERWISE put '秋天或冬天';
      END;
      
```

```

SELECT;
  WHEN(age<=12) put '少年';
  WHEN(age<35) put '青年';
  OTHERWISE put '中老年';
END;

```

- DO

- DO i=... TO ... BY ... ;
 

```

      ...
      END;
      
```

```

/*先判断，再执行*/
DO WHILE()
  ...
END;

```

```
/*先执行，再判断*/
DO UNTIL(退出条件);
...
END;
```

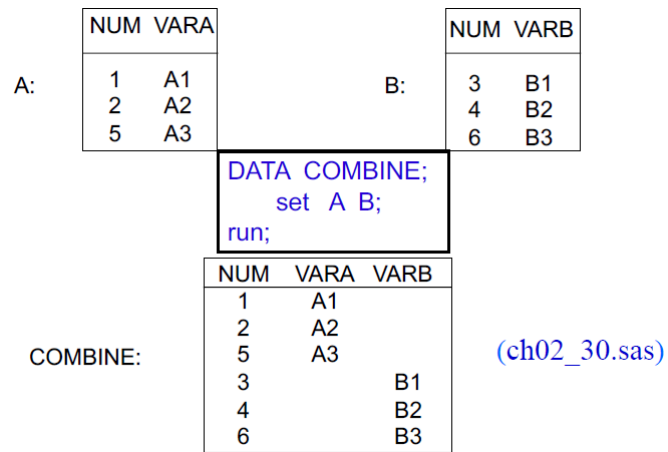
- 配合 LEAVE 可直接退出循环，相当于break;
- 配合 CONTINUE 可直接进入下一次循环;
- ARRAY数组
  - 声明方式：ARRAY 数组名(维数) 元素名列表 (初始列表)
  - ```
ARRAY test(3) math chinese english (0,0,0);
```
  - 声明也可以为 ARRAY tests(\*) 维度用\*代替，表示将有后面元素列表决定
  - 也可以定义二维数组：ARRAY table(2,2) x11 x12 x21 x22;
    - 说明table(1,1)为x11，table(1,2)为x12，table(2,1)为x21，table(2,2)为x22。
  - 同理可以生成字符型的数组
    - ARRAY 数组名(维数说明) \$ 元素长度说明  
数组元素名列表(初始值表);
    - ARRAY names(3) \$ 10 child father mother;
  - 若不想给数组每个元素起名字，还可以用temporary：ARRAY 数组名(维数说明)  
\_TEMPORARY\_ (初始值表);
    - ARRAY x(3) \_TEMPORARY\_ (0, 0, 0);
    - 说明了一个有三个元素的临时数组x。其元素为x(1)，x(2)，x(3)，即使变量x1，x2，x3存在也与此数组无关。
    - 临时数组的特点是它只用于中间计算，最终不被写入数据集。
    - 临时数组与其它变量不同的是，它在数据步隐含循环（后面会解释此概念）中能自动保留上一步得到的值。
    - 临时数组当然也可以有多维数组，或字符型数组。
  - 数组变量过多时，若连续，还可以用-替代
    - ```
input comp1-comp10 prin1-prin6 ;
ARRAY y(*) comp1-comp10 prin1-prin6;
```

## 1.8 合并SAS数据集

### 串接

- 直接串接

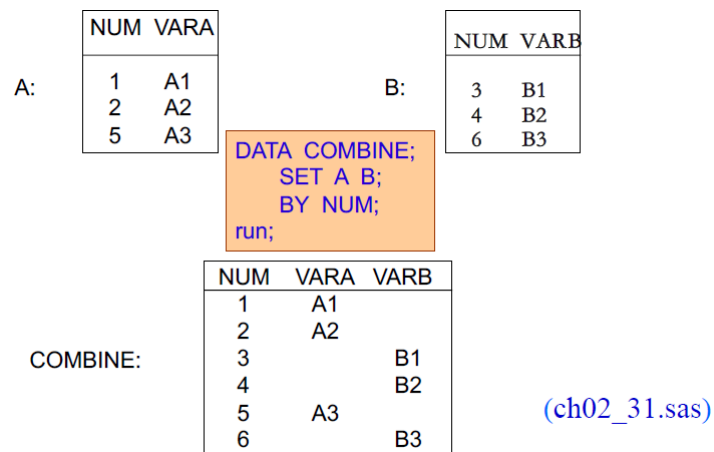
```
DATA SASDataSet;
  SET SASDataSet_1
    SASDataSet_2;
  ...
RUN;
```



- 插入式串接
  - 首先要进行排序, 再进行串接

```
PROC SORT DATA=SASDataset1
    OUT=Dataset1;
    BY variables;
RUN;
```

```
DATA Dataset;
    SET Dataset1 Dataset2;
    BY variables;
    ...
RUN;
```



## 并接

```
DATA SASDataSet;
    MERGE Dataset1 Dataset2;
    BY variables;
    ...
RUN;
```

NUM	VARA
1	A1
2	A2
3	A3

NUM	VARB
2	B1
2	B2
3	B3

```
DATA COMBINE;  
  merge A C;  
  by num;  
run;
```

NUM	VARA	VARB
1	A1	
2	A2	B1
2	A2	B2
3	A3	B3

(ch02\_34.sas)

in=

- 在连接过程中可以使用语句：

- `DATA Dataset;`

- `SET Dataset1(IN=var1) Dataset2(IN=var2);`

- `RUN;`

- 此时 `var1` `var2` 就成为了一个0-1的中间变量，标识某一行观测是否来自于第一个/第二个数据集，故可以用其来筛选数据：

- `DATA Dataset;`

- `MERGE Dataset1(IN=var1) Dataset2(IN=var2);`

- `BY NUM;`

- `IF var1 = 1;`

- `RUN;`



该程序则将保留第一个数据集中出现的观测

### Full dataset

Obs	num	vara	varb
1	1	A1	B1
2	2	A2	B2
3	3	A3	
4	4		B3

### if ia and ib

Obs	num	vara	varb
1	1	A1	B1
2	2	A2	B2

### if ia

Obs	num	vara	varb
1	1	A1	B1
2	2	A2	B2
3	3	A3	

## FIRST & LAST

- 在DATA步中, 若用 BY 进行分组, 则可以用 `FIRST.var` 和 `LAST.var` 取每个组第一个和最后一个元素

# 属性数据分析

## 1. 属性数据

- 所得到的信息是样本中个体的分类, 而不是定量变量的值。
- 属性数据类型
  - 名义变量
    - 名义变量的值之间无逻辑次序, 取值仅表示不同事物所属类别。
    - 可按任何次序排序编码
    - 例如性别,职业,地区,...都是名义变量.
  - 有序变量
    - 有序变量的值有明确的逻辑次序, 但各个值之间的距离并不清楚
    - 例如小中大
  - 区间变量
    - 区间变量是有大小顺序的连续数值变量, 且数值间的差值是有意义的。
    - 例如温度

- 但对两个数值的比率是没有意义的。例如由 $40/10=4$ ,而认为40度比10度热3倍的说法是不合适。
- 比率变量
  - 比率变量也是连续型的变量, 不仅数值之差有意义, 且要求有绝对的零点(具有明确的真正的零点)
  - 两数值的比率也是很重要。例如变量:饮料的体积,金子的重量等.

## 2. 列联表分析

### PROC FREQ 生成列联表

- 基本生成方法:
  - 当相关数据是具体到每一个个体详细生成时, 可直接用 PROC FREQ 进行生成

Obs	student	sex	major
1	1	男	是
2	2	男	非
3	3	女	是
4	4	男	非
5	5	女	是
6	6	女	是
7	7	男	非
8	8	男	非
9	9	男	是
10	10	女	是
11	11	男	非
12	12	女	是
13	13	男	是
14	14	男	是
15	15	男	非
16	16	女	是
17	17	男	是
18	18	男	非
19	19	女	非
20	20	男	是

```
PROC FREQ DATA=SASData;  
  TABLE var1*var2;  
RUN;
```

- 但若给定的数据本身就是已总结的数据时, 则需要加上 WEIGHT 选项

Obs	decision	defrace	numcell
1	是	白人	19
2	是	黑人	17
3	否	白人	141
4	否	黑人	149

```
PROC FREQ DATA=SASData;  
  TABLE var1*var2;  
  WEIGHT numcell;  
RUN;
```

- PROC FREQ的其他一些可选项

- PROC FREQ DATA=SASDataset  
ORDER=... /\*可选DATA, FORMATTED, FREQ, INTERNAL\*/  
NOPRINT; /\*该命令将取消结果输出\*/  
TABLES var1 \* (var2 var3) /\*通过添加括号可以在一个命令中生成两个列联表\*/  
/ NOCOL NOROW NOCUM NOFREQ NOPERCENT /\*注意TABLES的选项前有反斜线\*/  
MISSING /\*将显示缺失值，且计算时缺失值也包括在内\*/  
MISSPRINT /\*将显示缺失值，但计算时并不考虑在内\*/  
LIST /\*若有list则列联表将变成一维的列表\*/  
DEVIATION /\*显示偏差，偏差=实际频数-期望频数\*/  
OUT=...;  
WEIGHT ... ;  
BY ... ;  
RUN

- TABLES var1\*(var2 var3) 表示生成var1\*va2 var1\*var3 两份独立的列联表
- TABLES var1\*var2\*var3 表示按照var1分层，当var1取不同值时分别生成多个var2\*var3的表

### 3. 无关联性检验

#### Pearson卡方检验

- 计算公式：

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O：观测频数
- E：期望频数
- SAS语句

- PROC FREQ DATA=SASDataset;  
TABLES var1\*var2 / NOROW NOCOL NOPERCENT  
EXPECTED CHISQ CELLCHI2;  
WEIGHT NUMCELL;  
RUN;

- EXPECTED：生成期望频数

expected  
FREQ 过程

频数 期望值	表 - decision * defrace			
	defrace			
	decision	黑人	白人	合计
否		149	141	290
		147.67	142.33	
是		17	19	36
		18.331	17.669	
合计		166	160	326

- CHISQ：进行 $\chi^2$ 检验

chisq				
FREQ 过程				
频数	表 - decision * defrace			
	decision	defrace		
		黑人	白人	合计
否		149	141	290
是		17	19	36
合计		166	160	326

表 "defrace-decision" 的统计量				
统计量	自由度	值	概率	
卡方	1	0.2214	0.6379	
似然比卡方检验	1	0.2215	0.6379	
连续调整卡方	1	0.0863	0.7689	
Mantel-Haenszel 卡方	1	0.2208	0.6385	
Phi 系数		0.0261		
列联系数		0.0261		
Cramer V		0.0261		

Fisher 精确检验	
单元格 (1,1) 频数 (F)	149
左侧 Pr <= F	0.7412
右侧 Pr >= F	0.3843
表概率 (P)	0.1255
双侧 Pr <= P	0.7246

样本大小 = 326

o cellchi2: 输出 $\chi^2$ 贡献

cellchi2				
FREQ 过程				
频数 单元格卡方	表 - decision * defrace			
	decision	defrace		
		黑人	白人	合计
否		149 0.012	141 0.0125	290
是		17 0.0967	19 0.1003	36
合计		166	160	326

• 对卡方检验的说明:

o 卡方 (Pearson chisq)

- 广泛应用于**分类变量** (categorical data) 的**独立性检验**中, 也可用于**分类变量的比较检验**中。这两种检验都需要用到R×C列联表 (R×C contingency table), 其中R表示行 (Row), C表示列 (Column)。本文只讨论**行列变量都是无序变量的情形**, 最简单的情形是行与列都是**二分类无序变量**, 这样的数据也称为**四格表资料**。
- 一般来说, 利用Pearson's chi squared test对R×C列联表进行检验的理论上的要求:
  - 样本来自简单随机抽样。
  - 各个格子是相互独立的。
  - 样本量应尽可能大。总观察数应不小于40, 且每个格子的频数应大于等于5 (否则应考虑其他的检验方法, 后面会讲到)。
  - 依据样本数据计算出的理论频数应不小于5 (关于什么是理论频数, 马上就会讲到)。
- 卡方的p值只能说明是否存在关联性, 其大小不能说明关联程度的强弱
- 卡方检验的p值受到样本容量大小影响
- 若超过1/5的单元格的期望频数<5, 此时存在明显误差, 要改用精确p值

o 似然比卡方

- 与Pearson 卡方计算的结果相近

o 连续调整卡方

- 该修正由英国统计学家Frank Yates提出, 修正的目的是在小样本情况下, 降低将离散型频数数据近似到连续性卡方统计量的过程中的误差。然而关于耶茨连续性修正有很多争

论，许多人反对修正的理由是它的结果过于保守，导致有些可能显著的检验变得不显著了。

- Mantel-Haenszel 卡方

- **Mantel-Haenszel卡方检验也称线性趋势检验 (Test for Linear Trend) 或定序检验 (Linear by Linear Test) [1]。**

进行Mantel-Haenszel卡方检验，需要满足以下两个假设。

**假设1：**其中一个变量是有序分类变量。

**假设2：**另一个变量是有序分类变量（或二分类变量）。

- 首先需要通过列联表了解数据，如下图。本例中可观察到列联表对角线附近的观测数最多。例如有1-2个症状的观测的疼痛等级为 I 或 II，有5-6个症状的观测的疼痛登记为 III 或 IV，提示两个变量间可能存在关联。是否存在线性关联则需要Mantel-Haenszel卡方检验判断。
- Mantel-Haenszel卡方检验只能说明存在线性关系，但不能给出这种线性相关的强度和方向。**要判断线性相关的强度和方向，需要查看Pearson相关系数。**

## 关联性检验

### 检验方法的选择原则

在孙振球教授的《医学统计学》p.114，介绍了四格表的卡方检验的三个选择原则：

1. 当 $n \geq 40$ 且所有的 $E \geq 5$  (也就是 $a$ 、 $b$ 、 $c$ 、 $d$ 对应的卡方分布的理论频数)，可以用卡方检验的基本公式，但当 $p \approx \alpha$ 时，改用Fisher精确检验。
2. 当 $n \geq 40$ 但有 $1 \leq E < 5$ 时，用卡方检验的校正公式，或改用Fisher精确检验。
3. 当 $n < 40$ ，或 $E < 1$ 时，用四格表的Fisher精确检验。

当然，现在统计软件已经功能非常完善了，我认为不管表格中的数据属于哪种情况，直接用Fisher精确检验总会相对准确一些。个人之见，仅供参考

### Fisher 卡方

- Fisher精确检验 (Fisher exact test) 是一个比较另类的检验，它没有统计量，更没有繁琐的统计量的表格，它算出来的就是p值，但是它在大量样本情况下手算几乎是不可能的，因为它涉及到阶乘运算。
- Fisher确切概率法能够让我们准确地认识到假设检验中p值的含义：在原假设成立的情况下，发生当前甚至更极端的情形的概率。
- 总的来说，Fisher's exact test是最普适的，但为什么它远没有Pearson卡方检验有名呢？究其原因，卡方检验流传甚广的原因主要是计算简便，很多情况下不需计算器徒手就能算出来，相比之下，Fisher精确检验在大量样本情况下，在没有计算机的时代几乎算不出来。理论上来说，Fisher精确检验得到的结果才是准确的，卡方检验是利用了大量样本下渐近卡方分布的性质，即使是近似服从卡方分布，得到的结果仍是近似值。当然，在两组间的差异足够大的情况下，利用卡方检验得到的p值与利用Fisher精确检验得到的p值差别很小。
- SAS 语法

```
○ PROC FREQ DATA=DataSet;
  TABLES var1*var2 / EXACT;
  EXACT PCHI;
RUN;
```

## Cochran-Mantel-Haenszel检验 (CMH分层卡方检验)

- PROC FREQ DATA=SASData ORDER=DATA;  
TABLES vargrp\*var1\*var2 / CMH;  
WEIGHT n;  
RUN;

- CMH 即为分层卡方检验

- Cochran-Mantel-Haenszel, 简称CMH检验, 是研究两个我们关注的分类变量之间关联性的一种检验方法。但有时数据除了我们研究的变量外, 还混杂或隐含了其它的变量, 如果将这些变量纳入分析中, 则有可能得出完全不同的结论, 著名的Simpson悖论就是这个问题的典型案例。
- 换句话说, 在2 x 2 表格数据的基础上, 引入了第三个分类变量, 称之为混杂变量。混杂变量的引入使得该检验可以用于分析分层样本, 作为生物统计学领域的一种常用技术, 该检验常用于疾病对照研究。
- 对于这种分层的列联表, 通常可以各层单独做卡方检验。但除此之外, 我们还想知道在数据分层条件下, 总体的状态如何, 此时分层的作用就像是试验设计中的区组化, 虽然分层可能对卡方检验结果有影响, 但我们并不关注它, 而是考虑排除其影响后卡方检验的显著性。这种方法就是Cochran-Mantel-Haenszel检验, 简称CMH检验或MHC检验。
- 例如, 研究不同性别和候选人投票结果之间的关联:

- | 性别 | 候选人 A | 候选人 B |
|----|-------|-------|
| 女  | 942   | 737   |
| 男  | 737   | 699   |

- 这里有两个二分类变量, 第一个是投票者的性别, 第二个是候选人A和B。考虑到所有的投票者本身存在分层现象, 来自3个不同的州, 针对不同的州重新统计。
- 上述例子中, 投票者出现了分层现象, 来自3个不同的州。如果不考虑这个因素, 直接统计性别和候选人的频数分布, 采用卡方或者费舍尔精确检验来进行分析, 即使得到了阳性的结果, 也无法确定是不同性别之间真实存在投票的差异还是由于来自不同的州导致了这样的差异。
- 由于投票者的分层现象, 直接采用卡方或者费舍尔精确检验进行分析是不太合适的。在上述模型中, 投票者的分层就是一个典型的混杂变量, 对于这样的数据可以采用CMH检验进行分析。

- SAS输出结果分析

- 

Cochran-Mantel-Haenszel 统计量 (基于表评分)				
统计量	备择假设	自由度	值	概率
1	非零相关	1	0.2215	0.6379
2	行评分均值不同	1	0.2215	0.6379
3	常规关联	1	0.2215	0.6379

- 概率 (p值) >0.05, 接受原假设, 认为行列 (var1,var2) 之间不相关

- 

普通优比和相对风险				
统计量	方法	值	95% 置信限	
优比	Mantel-Haenszel	1.1355	0.6693	1.9266
	Logit	1.1341	0.6678	1.9260
相对风险 (第 1 列)	Mantel-Haenszel	1.1291	0.6808	1.8728
	Logit	1.1272	0.6794	1.8704
相对风险 (第 2 列)	Mantel-Haenszel	0.9945	0.9725	1.0170
	Logit	0.9945	0.9729	1.0166

优比齐性的 Breslow-Day 检验	
卡方	0.1173
自由度	2
Pr > 卡方	0.9430

- 优势比的值接近1，表明行列相关性弱

## McNemar 检验

- ```
PROC FREQ DATA=SASData ORDER=DATA;
  TABLES var1*var2 / AGREE;
  WEIGHT;
RUN;
```

- AGREE 语句为McNemar检验

- 由美国心理学家、统计学家Quinn Michael McNemar提出，是对一对名义数据（paired nominal data）进行检验的方法，应用于行与列变量都是只有两个对立面的无序分类变量的2×2列联表（2×2 contingency table），以确定行与列的边际频数是否相等，即是否具有边际同质性（marginal homogeneity）。
- 这样说起来很拗口，其实就是**检验行列是否相关**，不妨举个例子看看：对于某种试验，有两种检验方法（Test1, Test2），均只有阳性和阴性（positive, negative）两种结果，这样检验方法和检验结果就是一个配对。现有n个样本，对它们进行这两种检验，样本检测结果只可能有四种结果：(positive, positive)、(positive, negative)、(negative, positive)、(negative, negative)
- 现我们的目的是检验这两种检验方法的结果是否有差别。可以看出，对角线上的元素a和d是表示两种检验结果相同（同为positive、同为negative），而c和b代表两种检验结果不同。
- 假设我们想知道禁烟广告对人们吸烟态度影响有多大，先调查100个测试者对吸烟的态度（支持或反对），然后给他们播放禁烟广告，再次询问他们对吸烟的态度。因为同一个测试者在两种不同条件下（看广告前和广告后）对同一个问题做出的两次回答，所以这个实验叫做配对实验设计。由于数据之间并不独立，不能采取卡方检验，这里只能采用配对样本的McNemar检验。
- 配对列联表也要求样本保持不变，如可以是部件加工前和加工后的比较，也可以是两种不同的评价方法的对比。表格可以进一步写成这样：

- |    |   | B            |              | 合计                  |
|----|---|--------------|--------------|---------------------|
|    |   | +            | -            |                     |
| A  | + | a            | b            | $n_{11}=a+b$        |
|    | - | c            | d            | $n_{12}=c+d$        |
| 合计 |   | $n_{c1}=a+c$ | $n_{c2}=b+d$ | $n= n_{11}+ n_{12}$ |

- 针对配对的四格表，有两种分析方法可以选择，即McNemar检验和Kappa检验。前者关注的是差异，后者关注的是一致性。
- a和d代表结果的一致性，b和c代表结果产生的变化。在McNemar检验中，原假设是对样本所施加的处理没有显著效应，也就是发生不同方向变化的可能性是一样的，有多少“- +”，就应该有多少“+ -”，即b=c，如果两者差异很大，则说明两种不同的处理有显著的差异，或一种处理的前后状态存在显著差异。
- McNemar检验与a和d两个格子的值无关，当这两个值很大时，即使检验结果显著，其实际意义也不是很大。因此我们需要考虑一致性的问题，这就需要Kappa检验。

## Kappa 检验

- SAS中的代码与McNemar相同：

- ```
PROC FREQ DATA=SASData ORDER=DATA;
  TABLES var1*var2 / AGREE;
  WEIGHT;
RUN;
```

- Kappa取值从-1~+1。-1代表完全不一致( $a=d=0$ 且 $b=c$ )；+1代表完全一致( $b=c=0$ )；0表示结果纯粹是瞎蒙的；负值代表结果比瞎蒙还差(当然也没有什么实际意义，实际上出现得很少)；正值越接近1代表一致性越好。通常0.75以上表示一致性较满意，0.4以下一致性不好。但是对于测量系统来说，需要在0.9以上才能说是好的测量系统。
- 例2：某工厂针对注塑产品表面质量一般采用人工和设备两种方式进行检验。为了了解两种检验方式的一致性，随机选择35件样品，采用两种方式分别进行检验，结果如下表。

○

		设备		合计
		合格	不合格	
人工	合格	25	3	28
	不合格	5	2	7
合计		30	5	35

- 根据上面的公式计算出Kappa值为0.2，说明两种检验的结果一致性很差。
- 可能有人会问，这个分析并没有告诉我们哪一种更好。为了确认哪一种方法更好，可以加入标准这个因素，即由专家对样品进行仔细鉴别，确定标准的结果，然后再将两种检验方法的结果分别与此对比。其中的一张表是这样的：

		标准		合计
		合格	不合格	
人工	合格	28	0	28
	不合格	1	6	7
合计		29	6	35

根据此表计算出的Kappa值是0.906，说明人工检验的准确率是很高的。

- Kappa值也可以分析多于四格表的列联表,如下例:
- 例3：某个考试共有80道单选题，每题的有A、B、C、D四个答案，为考察某个考生的成绩是不是随便猜的，可以用Kappa分析做一个较确切的判断。数据表如下：

○

		标准答案				合计
		A	B	C	D	
试卷答案	A	19	1	1	0	21
	B	0	18	1	2	21
	C	1	0	18	1	20
	D	0	1	0	17	18
合计		20	20	20	20	80

- 完全一致的有 $19+18+18+17=72$ ，计算得 $P_0=72/80=0.9$ 。

$$P_e = (21 \times 20 + 21 \times 20 + 20 \times 20 + 18 \times 20) / 80^2 = 0.25。$$

由此计算出 $Kappa = (0.9 - 0.25) / (1 - 0.25) = 0.867$ 。这个值比较大，说明学生的答案不是瞎蒙的，是真的学会了。

## SAS中的Measures语句

- ```
PROC FREQ DATA=DataSet;
    TABLES var1*var2 / MEASURES;
RUN;
```

- 通过 MEASURES 语句可以给出一系列统计量列表，这些统计量都是用来测定行列之间是否相关的，即是否当行变量增加时列变量也增加。
- 对于定序变量的检验：gamma, Kendall's tau-b, Stuart's tau-c, Somers'D
  - 对有序变量,列联表中变量各测量水平必须按顺序排列,否则有序关联性的度量是不准确的.
- 对于名义变量的检验：lambda asymmetric, lambda symmetric, uncertainty coefficients



## 优势比与相对风险

### riskdiff检验

- ```
PROC FREQ DATA=SASData;  
  TABLES var1*var2 / RISKDIFF;  
  WEIGHT n;  
RUN;
```

## 回归分析

### 1. 相关分析

- 相关系数
  - Pearson 相关系数
  - Spearman 相关系数
  - kendall 相关系数
  - ```
PROC CORR DATA=Dataset ... ; /*省略号处可选: KENDALL PEARSON COV*/  
  VAR ...;  
  WITH ...; /*与VAR语句搭配使用,可生成相关系数矩阵*/  
  PARTIAL variables ; /*用来计算偏相关系数,即认为该变量为常数后的相关系数*/  
  BY variables; /*BY语句表示将数据按照该变量进行排序*/  
RUN;
```
  - with 的说明:
    - 当输入 VAR x1 x2; WITH y1 y2 y3 得到的结果为:  
 
$$\begin{bmatrix} r(Y1,X1) & r(Y1,X2) \\ r(Y2,X1) & r(Y2,X2) \\ r(Y3,X1) & r(Y3,X2) \end{bmatrix}$$
    - 关于相关系数的说明:
      - 相关系数很强并不表示变量间一定有因果关系
      - 相关系数是说明线性联系程度的,相关系数接近于0的变量间可能存在非线性联系（可能是曲线关系）；
      - 有时个别极端数据可能影响相关系数
- 散点图
  - ```
PROC PLOT DATA=Dataset;  
  PLOT var1*var2='*';  
RUN;
```
  -

### 2. 一元线性回归

- SAS语句

```

o PROC REG DATA= Dataset
    CORR /*可以指定生成相关系数*/
    ALPHA= /*可以指定置信度*/;
MODEL 因变量 = 自变量1 自变量2 ... / <options1>;
OUTPUT OUT=Dataset2
    <options2> = names;
ID 变量;
PLOT 因变量*自变量 / <options3>;
RUN;

```

■ MODEL <options1> 可选的命令有(注意要添加 /):

- RIDGE= 岭回归
- AIC BIC MSE RMSE SSE CP 控制输出模型的各种统计指标
- P: 计算模型每个  $x$  对应的  $\hat{y}_i$
- CLI: 计算当前模型的  $100(1 - \alpha)\%$  置信限预测值
- CLM: 计算当前模型的  $100(1 - \alpha)\%$  置信限模型均值
- NOINT: 表示截距项不纳入模型中
- R: 残差分析

■ ID 的作用:

- ID x;

输出统计量										
观测	x	因变量	预测值	标准误差均值预测	95% 置信限均值		95% 置信限预测		残差	
1	352	166	165.6636	5.6526	152.8765	178.4508	137.3315	193.9958	0.3364	
2	373	153	172.2149	5.2060	160.4380	183.9918	144.3242	200.1056	-19.2149	
3	411	177	184.0695	4.4699	173.9579	194.1810	156.8401	211.2989	-7.0695	
4	441	201	193.4283	3.9824	184.4195	202.4372	166.5889	220.2678	7.5717	
5	462	216	199.9796	3.7101	191.5867	208.3724	173.3406	226.6186	16.0204	
6	490	208	208.7145	3.4616	200.8838	216.5453	182.2473	235.1818	-0.7145	
7	529	227	220.8811	3.3803	213.2344	228.5278	194.4677	247.2945	6.1189	
8	577	238	235.8553	3.7174	227.4461	244.2646	209.2111	262.4995	7.1447	

- 无 ID x;

输出统计量							
观测	因变量	预测值	标准误差均值预测	95% 置信限均值		95% 置信限预测	
1	166	165.6636	5.6526	152.8765	178.4508	137.3315	193.9958
2	153	172.2149	5.2060	160.4380	183.9918	144.3242	200.1056
3	177	184.0695	4.4699	173.9579	194.1810	156.8401	211.2989
4	201	193.4283	3.9824	184.4195	202.4372	166.5889	220.2678
5	216	199.9796	3.7101	191.5867	208.3724	173.3406	226.6186
6	208	208.7145	3.4616	200.8838	216.5453	182.2473	235.1818
7	227	220.8811	3.3803	213.2344	228.5278	194.4677	247.2945
8	238	235.8553	3.7174	227.4461	244.2646	209.2111	262.4995
9	268	255.8209	4.7213	245.1405	266.5014	228.3752	283.2667

■ OUTPUT <options2>=names

- 左侧 options2 为想要输出的统计量,可选的选择有
  - COOKD=: 计算Cook'D 距离
  - H=: 计算帽子矩阵  $X(X'X)^{-1}X'$
  - LCL/L95: 给定/95%置信度下的模型预测值下限
  - UCL/U95: 给定/95%置信度下的模型预测值上限
  - LCLM: 均值的给定置信度下限
  - UCLM: 均值的给定置信度上限
  - PREDICTED: 输出预测值
  - RESIDUAL: 输出残差项
- 右侧 names 为给统计量起的名字

- PLOT options
  - conf95: 输出均值置信线
  - pred95: 输出预测置信线
  - overlay: 多图叠加在一幅图
  - AIC CP MSE SSE: 图形左边显示相应统计量

### 3. 多元线性回归

- F检验
- t检验

```
PROC REG DATA=Dataset;
    MODEL y = x1 x2 ... xn;
RUN;
```

### 4. 变量选择

- 选择准则
  - 均方误MSE 最小
  - 预测均方误差 最小
  - Cp准则 最小
  - AIC或BIC准则 最小
  - Adj R2 最大
  - Schwarz's Bayesian Criterion
- SAS变量筛选方法

```
PROC REG DATA=Dataset;
    MODEL y = x1 ... xn
    / P CLI CLM R
    SELECTION = <options1> SLENTY= ...
    SLSTAY=...
    BEST = n
    AIC SBC RMSE
    INCLUDE = ... ;
RUN;
```

- SELECTION= 可选选项

- NONE: 全模型
- FORWARD 向前增选法
  - 从无变量开始,逐个加入变量计算F统计量的p值与 SLENTY (若未声明则为0.5)比较. 如果没有F统计量大于 SLENTY 则结束选择, 否则则将向模型中不断添加F统计量最大的变量. 如此往复,直到不再有新的待加入统计量能带来显著的F.
- BACKWARD: 向后剔除法
  - 从全模型开始,逐个剔除变量,每次都要计算模型的F统计量,直到剩余模型再进行删除会导致模型不再显著, 或者说直到剩余模型的F统计量为 SLSTAY 显著水平. 每步的筛选都将删除对模型的F统计量最小的变量.
- STEPWISE: 逐步筛选法
  - 逐步回归的基本思想是将变量逐个引入模型, 每引入一个解释变量后都要进行F检验, 并对已经选入的解释变量逐个进行t检验, 当原来引入的解释变量由于后面解释

变量的引入变得不再显著时，则将其删除。以确保每次引入新的变量之前回归方程中只包含显著性变量。

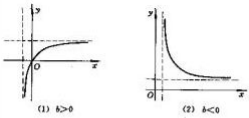
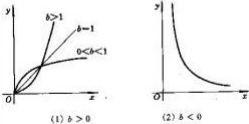
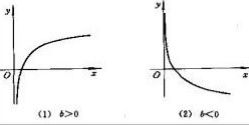
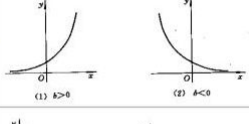
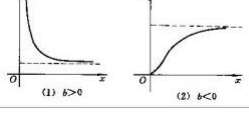
- 逐步回归法可以认为是向前引入法与向后剔除法的综合。**逐步回归法克服了向前引入法与向后剔除法的缺点，吸收两种方法的优点。**逐步回归法是以向前引入为主，变量可进可出的变量选取方法。它的基本思想是，当被选入的变量在新变量引入后变得不重要时，可以将其剔除，而被剔除的变量当它在新变量引入后变得重要时，又可以重新选入方程。
- 逐步回归分析是多元回归分析中的一种方法。回归分析是用于研究多个变量之间相互依赖的关系，而逐步回归分析往往用于建立最优或合适的回归模型，从而更加深入地研究变量之间的依赖关系。

- MAXR
- MINR
- RSQUARE
- ADJRSQ
- CP

○ BEST=n

- 配合 RSQUARE ADJRSQ CP 模型选择使用
- best表示在筛选过程中,对于每个变量个数只输出效果最好的n个子集和相应的统计量

## 5. 非线性回归

曲线方程	变换公式	变换后的线性方程	曲线图形
$\frac{1}{y} = a + \frac{b}{x}$	$X = \frac{1}{x}$ $Y = \frac{1}{y}$	$Y = a + bX$	
$y = ax^b$	$X = \ln x$ $Y = \ln y$	$Y = a' + bX (a' = \ln a)$	
$y = a + b \ln x$	$X = \ln x$ $Y = y$	$Y = a + bX$	
$y = ae^{bx}$	$X = x$ $Y = \ln y$	$Y = a' + bX (a' = \ln a)$	
$y = ae^{\frac{b}{x}}$	$X = \frac{1}{x}$ $Y = \ln y$	$Y = a' + bX (a' = \ln a)$	

### • 多项式回归

- 若进入回归模型的变量有一定的优先次序（如对多项式,线性项先于二次项,二次项先于三次项等）,应该用I型平方和(SS1)及相应的F统计量
- 若平等地考虑各个变量是否进入回归模型，则可用II型平方和(SS2)及其相应的F统计量
- 例题
  - 利用proc reg进行回归

- ```

title ' reg52B.sas--试验温度数据';
data reg52;
input t tc @@;
tt=t*t; ttt=tt*t;
cards;
5 99.2 10 99.7 15 99.9 20 100.2 25 100.3
30 100.4 35 100.4 40 100.3 45 100 50 99.8 55 99.4
;
proc reg data=reg52;
Lin: model tc = t / ss1 ss2;
Quad: model tc = t tt / ss1 ss2;
Cubic: model tc = t tt ttt / ss1 ss2;
run;

```

- 利用proc glm 进行回归

- ```

title 'sas0710.sas';

data reg52;
input t TC @@;
cards;
5 99.2 10 99.7 15 99.9 20 100.2 25 100.3
30 100.4 35 100.4 40 100.3 45 100 50 99.8 55 99.4
;

proc glm data=reg52;
model tc=t*t t*t*t / ss1 ss2;
title2 'Fittting polynomial Models with PROC GLM';
run;
quit;

```

- 多项式回归的说明:

- 若用 PROC GLM 进行回归,则不需要提前在数据集中生成高次的变量,而是在 MODEL 中直接计算即可
  - PROC GLM 无法提供回归诊断, 但直观的给出了SS1和SS2的p值,因此可以用GLM进行判断合适的回归阶次,若需要回归诊断再用 PROC REG

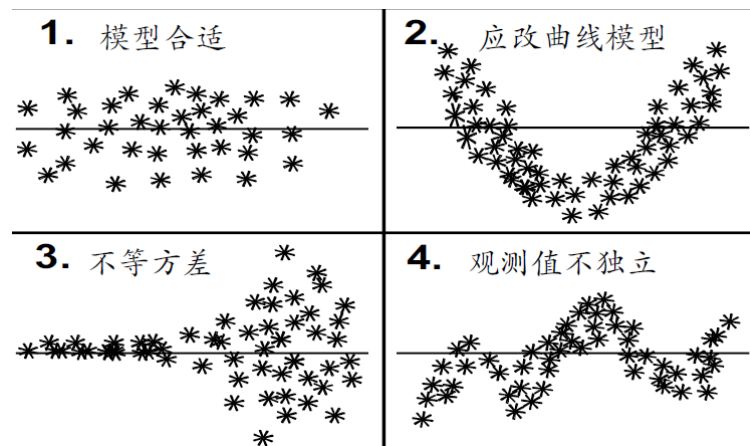
## 6. 回归诊断

- 必要性
  - 回归分析中讨论的估计和检验问题并不能用于验证回归模型的各项假定.另数据中的异常点可能使回归结果不稳定或不适用.
- 诊断的主要目标
  - 异常值(outliers)或强影响点的检查;
  - 误差项是否同方差, 不相关, 正态分布;
  - 自变量间是否存在多重共线性
- 特异值的类型
  1. 离群值: 因变量Y的特异值
  2. 高杠杆点: 自变量X的特异值
  3. 强影响点: 对回归分析造成很大影响的特异值

- 一个数据点可能是以上的一种或多种特异值. 辨别特异值时, 应当考虑上面这三种可能性
- 相对于离群值和高杠杆点, 强影响点对于数据分析的影响最大。

## 残差分析

- 



- 绘制残差-预测散点图

```
PROC REG DATA=...;
MODEL y = x1 ... xn / P R;
PLOT R.*P.;
RUN;
```

## 识别异常值点与强影响点

- ```
PROC REG DATA=Dataset;
MODEL y = x1 ... xn / R;
RUN;
```

- MODEL /R 的添加就会在模型输出中生成用于识别异常值的统计量

- 相关统计量或检验方法

- 学生化残差

- 若  $|\text{标准化残差} / \text{学生化残差}| > 3$ , 则观测点为异常点
- 若  $|\text{标准化残差} / \text{学生化残差}| > 2$ , 则观测点为可疑点

- Cook'D

- 对一个观测值其Cook D 统计量的值超过  $4/n$  时 ( $n$  为样本容量), 这个观测存在反常效应 (经验结论).
- 通常, 超过1时, 认为存在强影响点.

- DFFITS统计量

- DFFITS值反映去掉了某一个数据值之后, 新建立的模型对于其他点的拟合残差的大小变化情况。
- 一般来说, 当DFFITS大于 / 小于某个阈值的时候, 则可以认为这是一个强影响点。
- 然而, 在实际应用中, 对于阈值的设定是相对主观的, 不同的研究可能使用不同的阈值, 只要特异值的DFFITS明显不同于其他数据点, 就有可能被当作一个强影响点分析。

```
PROC REG DATA=Dataset;
MODEL y=x1 ... xn / INFLUENCE;
RUN;
```

- /INFLUENCE 生成DFFITS有关内容

- 偏杠杆图

- ```
PROC REG DATA=Dataset;  
    MODEL y=x1 ... xn / PARTIAL  
PARTIALDATA;  
RUN;
```

- `PARTIAL` 可以生成杠杆图, `PARTIALDATA` 可以生成偏杠杆数据

- 偏杠杆图是使有影响的观测可视化的方法, 偏杠杆图是两个回归的残差的散点图.
- 例如对变量`xr`的偏杠杆图: 纵轴是`Y`关于除`xr`以外所有`x`的回归的残差, 横轴是`xr`关于所有`x`的回归的残差.
- 有影响的观测通常分离与其它数据点或在某一轴上有极端数值.
- 偏杠杆图还可识别要加入哪些变量的高次项.

- 对于问题点的处理

1. 复验数据, 确认并无数据输入错误发生;
2. 若数据是有效的, 模型可能不合适. 拟合此数据可能需要使用高阶模型, 也可能数据是反常的;
3. 一般不剔除数据, 某些有影响的观测提供重要的信息. 若要剔除数据, 应给出必要的描述和说明.

## 共线性诊断

- 共线性问题

1. 自变量之间的共线性关系会隐藏变量的显著性
2. 会增加参数估计的方差
3. 产生不稳定模型

- 方差膨胀因子VIF

- ```
PROC REG DATA= ... ;  
    MODEL y=x1... / VIF;  
RUN;
```

- 一般认为 $VIF > 10$ 则出现强共线性

- 条件指数和方差比例

- ```
PROC REG DATA= ... ;  
    MODEL y=x1... / COLLIN 或 COLLINOINT;  
RUN;
```

- 

## 7. 案例