**15. 绘图Ⅲ—散点图,时间序列图,拟合回归图**

**三、散点图**

以两个定量变量的数值在二维坐标系中描点，生成的图形就是散点图。散点图能直观地显示出因变量随自变量变化而变化的大致趋势，据此可以选择合适的函数对数据点进行拟合。

基本语法：

PROC SGPLOT;

SCATTER X=*自变量* Y=*因变量* / *options*;

可选项：

（1）DATALABEL = variable-name

指定该变量的值作为各个数据点的标签，若不指定变量名，则使用Y变量的值；

（2）GROUP = variablename

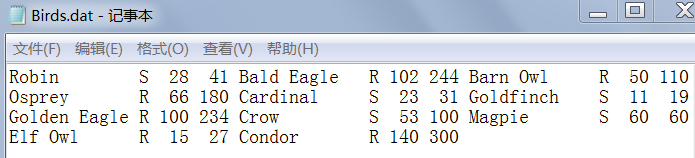
指定分组变量，按其值对数据分组分别处理；

（3）NOMISSINGGROUP

指定不包含分组变量缺失值的观测；

（4）TRANSPARENCY = n——指定透明度。

**例1** 关于鸟类的数据（C:\MyRawData\Birds.dat），变量包括名称、种类（S-鸣禽，R-猛禽）、身长（cm）、翼展（cm），注意每行有多组数据：



读入数据，绘制散点图，要求按类型分组。

**代码**：

**data** wings;

infile 'c:\MyRawData\Birds.dat';

input Name $12. Type $ Length Wingspan @@;

**run**;

\* Plot Wingspan by Length;

**proc** **format**;

value $birdtype

'S' = 'Songbirds'

'R' = 'Raptors';

**run**;

**proc** **sgplot** data = wings;

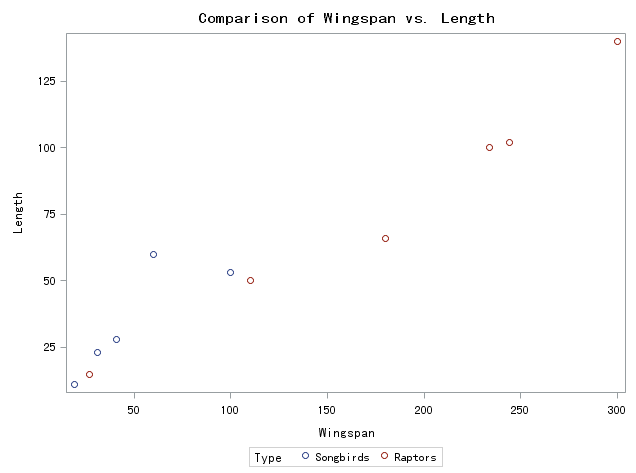
scatter X = Wingspan Y = Length / GROUP = Type;

format Type $birdtype.;

title 'Comparison of Wingspan vs. Length';

**run**;

**运行结果**：



**四、时间序列图**

时间序列图，是以时间为X轴,变量为Y轴的一种图（数据点之间依次用线段连接），观察变量是否随时间变化而呈现某种趋势。

基本语法：

PROC SGPLOT;

SERIES X=*时间变量* Y=*因变量* / *options*;

注：X变量必须事先排好序。

可选项：

（1）CURVELABEL = 'textstring'

给曲线添加标签，若不指定标签文本，则使用Y变量；

（2）DATALABEL = variable-name

指定该变量的值作为各个数据点的标签，若不指定变量名，则使用Y变量的值；

（3）MARKERS

给每个数据点加上标签；

（4）GROUP = variablename

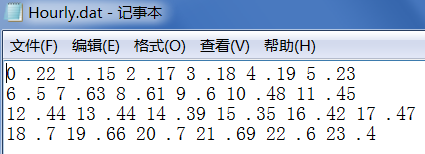
指定分组变量，按其值对数据分组分别处理；

（5）NOMISSINGGROUP

指定不包含分组变量缺失值的观测；

（6）TRANSPARENCY = n——指定透明度。

**例2** 某作家记录了某天的用电量数据（C:\MyRawData\Hourly.dat），变量包括时间（24个时刻）、用电量（千瓦时）：



读取数据，绘制用电量随时间变化的曲线图，画一条用电量=0.5的基准线。

**代码**：

**data** electricity;

infile 'c:\MyRawData\Hourly.dat';

input Time kWh @@;

**run**;

\* Plot temperatures by time;

**proc** **sgplot** data = electricity;

series X = Time Y = kWh / MARKERS;

REFLINE **0.5** / LABEL = ('0.5 kh') TRANSPARENCY = **0.5**;

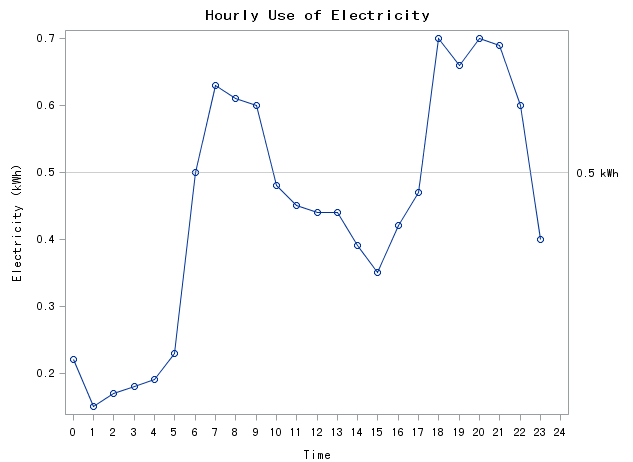
xaxis values = (**0** TO **24** BY **1**);

yaxis label = 'Electricity (kWh)';

title 'Hourly Use of Electricity';

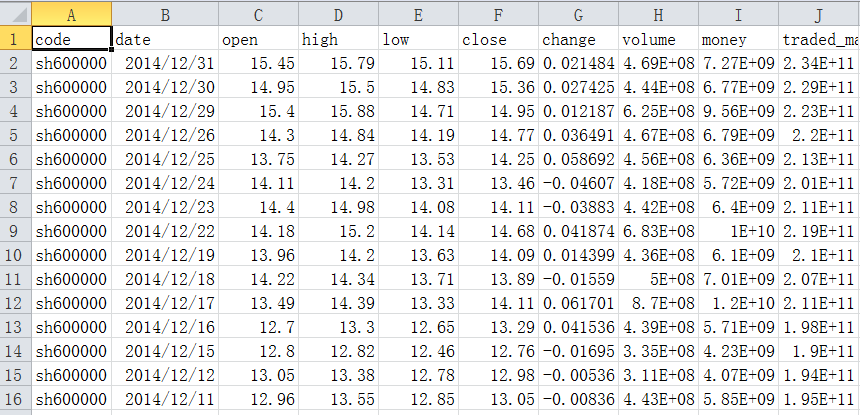
**run**;

**运行结果**：



**例3** 2013-2014年股票代码为sh600000的股票数据如下：（部分）

（C:\MyRawData\sh600000.csv）



读入数据，日期只选择2014年，只保留变量date, open（开盘价）, high（最高价）, low（最低价）, close（收盘价）。按日期从小到大排序，并绘制该股票2014年的开盘价和收盘价的变化趋势图。

代码：

**proc** **import** datafile = 'C:\MyRawData\sh600000.csv' DBMS=CSV OUT = sh600000 REPLACE;

getnames = YES;

datarow = **2**;

**run**;

**data** sh600000\_2014;

set sh600000;

keep date open high low close;

where **'1Jan2014'D** <= date <= **'31Dec2014'D**;

**proc** **sort** data = sh600000\_2014;

by date;

**run**;

**proc** **print** data = sh600000\_2014;

title 'sh600000股票2014年的数据';

**run**;

**proc** **sgplot** data = sh600000\_2014;

series X = date Y = open / name = 'kaipan' legendlabel='开盘价';

series X = date Y = close / name = 'shoupan' legendlabel='收盘价';

title '股票(sh600000)2014年开盘价、收盘价的变化趋势';

keylegend 'kaipan' 'shoupan' / LOCATION = INSIDE POSITION = TOP;

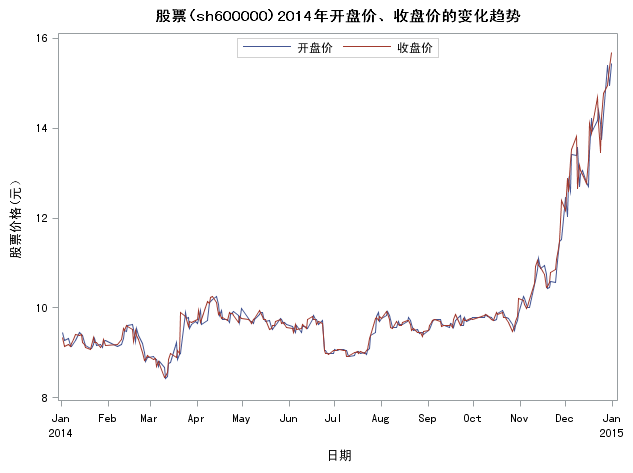
xaxis label = '日期';

yaxis label = '股票价格(元）';

**run**;

运行结果：（部分）





程序说明：

（1）csv文件是逗号分隔的文本文档，不是Excel表格；

（2）“/name =”语句是给曲线起名，后面“keylegend”语句使用它。

**五、拟合（回归）曲线图**

散点图只能大致显示两个变量之间的关系，进一步还可以绘制一条拟合曲线（应用最小二乘法原理，让误差的平方和最小）。

这种拟合曲线大概只能为后续真正做数据拟合提供参考信息，要拟合数据是用REG、GLM、NLIN过程步。

用SGPLOT过程步可以绘制拟合曲线图。

基本语法：

PROC SGPLOT data = 数据集;

拟合方式名X=*自变量* Y=*因变量* / *options*;

可选拟合方式有3种：

REG——线性或曲线回归，；

LOESS——局部加权回归（loess）；

PBSPLINE ——惩罚B样条（B样条拟合，加入惩罚函数

避免边界附近“过拟合”）

注：REG可以带可选参数 “/degree = n”，指定以n次多项式拟合。例如，

reg x=height y=weight / degree=**3**;

可选项：

（1）ALPHA = p

设定置信限: [0,1]，默认为0.05（95%可信）

（2）CLI

给个体预测值加上预测限（只限REG和PBSPLINE）；

（3）CLM

给均值预测值加上置信限；

（4）CURVELABEL = 'textstring'

为拟合曲线添加标注（默认是用Y变量）；

（5）GROUP = variablename

指定分组变量，按其值对数据分组分别处理；

（6）NOLEGCLI、NOLEGCLM、NOLEGFIT

不显示CLI带、CLM带、拟合曲线的图例说明；

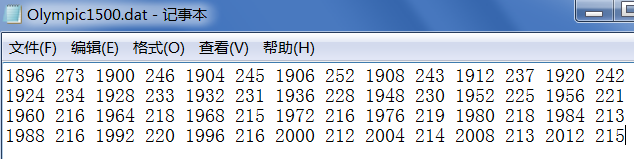
（7）NOMARKERS

不对数据点做标记；

（8）CLMTRANSPARENCY = n

指定置信限的透明度。

**例4** 奥运会男子1500米跑的数据（C:\MyRawData\Olympic1500.dat），变量包括年份、时间（秒），注意每行有多组数据：



读入数据，用REG和LOESS拟合方式绘制时间值的拟合曲线。

代码：

**data** Olympic1500;

infile 'C:\MyRawData\Olympic1500.dat';

input Year Men @@;

**run**;

**proc** **sgplot** data = Olympic1500;

LOESS X = Year Y = Men / NOMARKERS CLM NOLEGCLM;

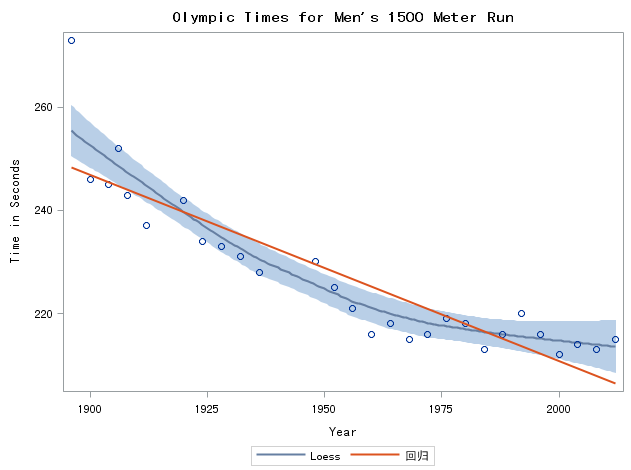
REG X = Year Y = Men;

label Men = 'Time in Seconds';

title "Olympic Times for Men's 1500 Meter Run";

**run**;

运行结果：



注：为了避免重复标记，用了NOMARKERS可选项，只让一种拟合曲线标记了散点。