

Integrating R Methods into SPSS*

Hongcheng Li

Department of Applied Mathematics, Shanghai Finance University, Shanghai
Email: lihc@shfc.edu.cn

Received: Nov. 10th, 2012; revised: Nov. 18th, 2012; accepted: Nov. 30th, 2012

Abstract: The paper introduces how to integrate the statistical methods and graphical methods of R into SPSS Version 17. It illustrates the procedures of calling the statistical methods of R, presenting the graphics of R, presenting the output of R in SPSS pivot tables. The way of calling the output of SPSS statistical procedures is introduced as well. The examples of the above procedures are given at the same time.

Keywords: SPSS; R; Linear Models; Xpath

SPSS 中集成 R 的统计方法*

李洪成

上海金融学院应用数学系, 上海
Email: lihc@shfc.edu.cn

收稿日期: 2012 年 11 月 10 日; 修回日期: 2012 年 11 月 18 日; 录用日期: 2012 年 11 月 30 日

摘要: 本文介绍了如何在 SPSS 版本 17 中集成 R 中的作图和统计方法。介绍了在 SPSS 中调用 R 的统计方法, 在 SPSS 中以数轴表的形式输出 R 的分析结果, 在 SPSS 中输出 R 的图形, 以及如何在 R 中调用 SPSS 统计分析的结果。并给出了各种方法的示例。

关键词: SPSS; R; 线性模型; Xpath

1. SPSS 和 R 简介

SPSS 是目前国内广泛应用的统计软件之一, SPSS 具有十分友好的图形用户界面, 它实现了传统的统计分析功能和方法, 新的版本中增加了神经网络和 EZ RFM 模块, 并且从版本 14 开始增加了编程扩展功能。SPSS 以其易用性、开放性和强大的统计分析功能而著名, SPSS 可以读入 SAS, Stata, Excel, 文本数据和数据库文件。而 R 则是有名的开源统计软件包, 它最早由新西兰奥克兰大学统计系的 Robert Gentleman 和 Ross Ihaka 两个人所开发, 现在已经成为免费软件基金会的 GNU 项目。R 软件和最早起源于贝尔实验室的 John Chambers 开发的 S 语言类似, 可以认为是

S 语言的一种实现。S 的代码基本上不用修改就可以在 R 下运行。R 包含所有的基本统计分析功能, 线性和非线性模型, 经典的统计检验, 时间序列分析, 聚类分析, 以及高级的作图技巧等。另外, R 可以运行时调用 C 或者 Fortran 的代码来完成计算量很大的工作。R 还有用户开发的各种特殊用途的插件包, 目前大约有 5300 多个插件包(2012 年 4 月)。SPSS 的 R 插件可以把 R 的功能和 SPSS 统计分析软件集成在一起, 充分利用 SPSS 软件的易用性和 R 的统计方法和强大的作图功能。国内的读者应用 SPSS 主要是通过图形用户界面, 对 SPSS 的编程了解较少。其实, SPSS 自发布伊始, 就是以编程方式出现的。随着其图形用户界面的发展, 大部分用户把编程忽略, 认为 SPSS 没有编程能力。其实, SPSS 除了其发行伊始就有的语

*资助信息: 上海高校示范性全英语教学课程项目资助。

法编程外，还可以集成 Visual Basic, Python 和 R 三种编程语言。本文对 SPSS 集成 R 的方法进行简介，以飨读者。

2. 在 SPSS 中调用 R 中的统计方法

通过 SPSS 的 R 插件，可以把 SPSS 当前数据集中的数据读入 R 中，从 R 中获取当前 SPSS 数据集中的数据字典，读取 SPSS 语法命令的输出结果，以枢轴表格式输出 R 的结果等。另外，也可以把 R 得到的分析结果写到 SPSS 数据集中，R 产生的图形输出也可以输出到 SPSS Viewer 中。

SPSS 的 R 插件提供了以下的函数类来和 R 进行交互：

`spssdata` 类，它提供了 SPSS 统计软件和 R 交换数据的方法。利用它可以把 SPSS 当前数据集中的数据读入到 R 的数据框(data frame)中，也可以把 R 数据框中的数据写到 SPSS 数据集中；

`spssdictionary` 类，提供了在 R 和 SPSS 集成环境中操作 SPSS 数据字典的方法；

`spssxmlworkspace` 类，提供了和 SPSS 内存中用于存贮 SPSS 统计分析输出的 XML 工作空间的交互方式；

`spsspivottable` 类，提供了把 R 的输出在 SPSS Viewer 中以枢轴表方式呈现的方法；

`spssRGRphics`，类提供了在 SPSS Viewer 中管理 R 输出的图形的显示方式的方法。

本节中以 SPSS 自带的 Cars.sav 数据为例。

2.1. SPSS 中调用 R 的方法

在 SPSS 命令语法程序中，把 R 的程序放在以下的程序块中：

```
BEGIN PROGRAM R.
```

```
R 语句
```

```
END PROGRAM
```

这样 R 的任何程序都可以在 SPSS 中调用并在 R 中执行，结果输出在 SPSS 的输出管理器中。

例如，在 SPSS 中调用 R 的线性模型来拟合变量 mpg 和 engine, horse, weight 间的关系。

```
BEGIN PROGRAM R.
```

```
  casedata<-spssdata.GetDataFromSPSS()
```

```
  model<-lm(mpg~engine+horse+weight,data=casedata)
  print(model)
```

```
END PROGRAM
```

在 SPSS 的 Viewer 中输出以下结果：

```
lm(formula = mpg ~ engine + horse + weight, data =
casedata)
```

```
Coefficients:
```

```
(Intercept)    engine      horse      weight
44.015491    -0.005527   -0.055633   -0.004617
```

2.2. SPSS 中以枢轴表输出 R 的结果

SPSS 的枢轴表是一种相当方便和直观的观察输出的方式，特别是对于复杂的数据结构和输出结果。例如三维的数据结构，有地区、性别、工资三个维度，以枢轴表方式输出不同地区和性别的平均工资和标准差，枢轴表可以方便的观察不同维度下的统计输出结果。下面的代码在 SPSS 中调用 R 的线性模型方法来拟合变量 mpg 和 engine, horse, weight 间的关系，并在 SPSS 中用枢轴表的方式输出线性回归的参数估计、标准误差、参数的 t 统计量以及相应的 p 值。

```
BEGIN PROGRAM R.
```

```
  casedata<-spssdata.GetDataFromSPSS()
  model<-lm(mpg~engine+horse+weight,data=casedata)
  sum<-summary(model)
  spsspivottable.Display(sum$coefficients,
  title="Regression Coefficients",
  rowdim="Predictors",hiderowdimtitle=FALSE,
  coldim="Coefficients and
  Stats",hidecoldimtitle=FALSE)
```

```
END PROGRAM
```

回归系数及其检验以枢轴表的形式输出(表 1)：

Table 1. Regression coefficients
表 1. 回归系数及检验

| Predictors | Coefficients and Stats | | | |
|-------------|------------------------|------------|---------|----------|
| | Estimate | Std. Error | t value | Pr(> t) |
| (Intercept) | 44.015 | 1.272 | 34.597 | 0 |
| Engine | -0.006 | 0.007 | -0.786 | 0.432 |
| Horse | -0.056 | 0.013 | -4.153 | 0 |
| Weight | -0.005 | 0.001 | -6.186 | 0 |

2.3. SPSS 中输出 R 的图形结果

用以下代码做出对拟合出的线性模型的残差的分析诊断图形(图 1):

BEGIN PROGRAM R.

```

casedata<-spssdata.GetDataFromSPSS()
model<-lm(mpg~engine+horse+weight,data=casedata)
par(mfrow=c(2,2))
plot(model)
    
```

END PROGRAM

从输出结果可知, 它给出了拟合值和残差的散点图(图 1 左上)、残差的正态 Q-Q 图、拟合值和标准残差平方根的散点图以及供进行影响点分析的杠杆值图。从图 1 右上的正态 Q-Q 图可知, 残差不是正态分布的。同时, 残差也不是同方差的, 随着拟合值的增大, 残差的波动变大。

2.4. 在 R 软件中调用 SPSS 的输出结果

在 R 中可以获得 SPSS 输出的各种结果。例如, 在 R 中调用 SPSS 方差分析的结果中的组内、组间自

由度和总的自由度, 以及方差分析的 p 值。代码如下:

OMS SELECT TABLES

```

/IF SUBTYPES=['ANOVA']
/DESTINATION FORMAT=OXML
XMLWORKSPACE='anova_table'
/TAG='anova_out'.
    
```

ONEWAY mpg BY origin

```

/MISSING ANALYSIS.
    
```

OMSEND TAG='anova_out'.

BEGIN PROGRAM R.

```

handle <- "anova_table"
context <- "/outputTree"
p_path <- paste("//category[@varName='mpg']/category[@text='Sig.']/cell/@text")
df_path <- paste("//category[@varName='mpg']/category[@text='df']/cell/@text")
pvalue <- spssxmlworkspace.EvaluateXPath(handle, context,p_path)
dfvalue <- spssxmlworkspace.EvaluateXPath(handle,
    
```

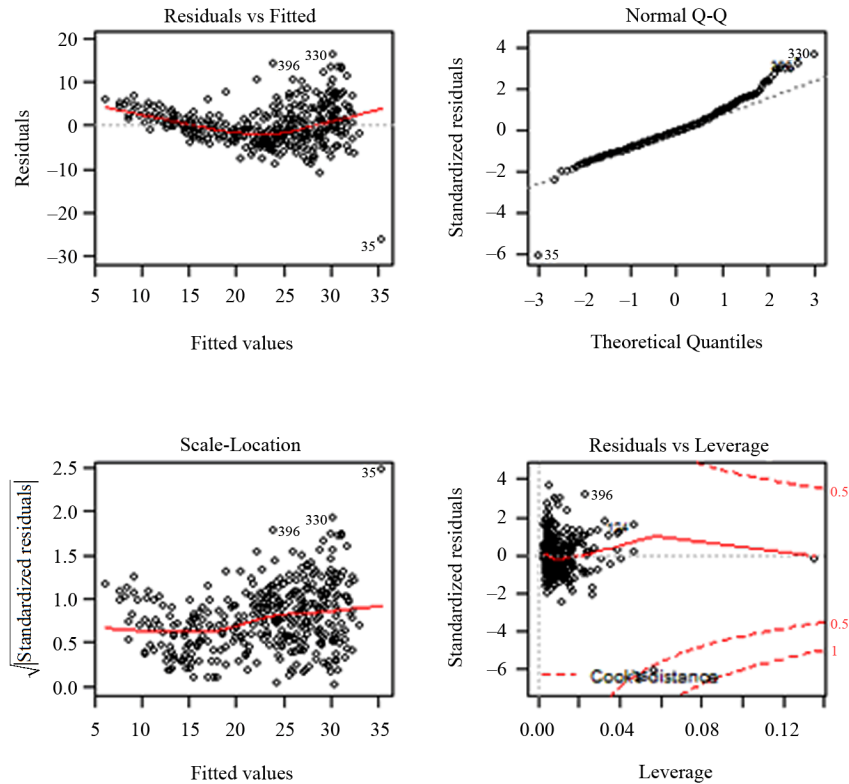


Figure 1. Diagnostic plots of residuals
图 1. 残差分析图

```
context,df_path)
cat("df=",dfvalue,"\n")
cat("p=",pvalue,"\n")
spssxmlworkspace.DeleteXmlWorkspaceObject(handle)
END PROGRAM.
```

输出结果为:

```
df:= 2 394 396
pvalue:= .000
```

3. 一个案例

在 SPSS 中目前还不能进行异方差的时间序列分析,而 R 中有许多异方差分析的插件包。本节应用 R 插件包 evir 中的数据集 BMW 进行异方差分析。数据集 BMW 记录了 BMW 公司从 1986 年 6 月到 1990 年 3 月股票的日对数收益,我们应该用数据来展示 SPSS 的 R 插件包的强大功能。

下面应用 SPSS 的 R 插件对 BMW 数据拟合一个 AR(1)/GARCH(1,1)模型,代码如下:

```
BEGIN PROGRAM R.
library(fGarch)
data(bmw,package="evir")
bmw.garch_norm = garchFit(~arma(1,0)+garch(1,1),
data=bmw,cond.dist="norm")
options(digits=3)
summary(bmw.garch_norm)
END PROGRAM.
```

得到 AR(1)/GARCH(1,1)模型的估计和检验结果如下:

Call:

```
garchFit(formula = ~arma(1, 0) + garch(1, 1), data =
bmw, cond.dist = "norm")
```

Mean and Variance Equation:

```
data ~ arma(1, 0) + garch(1, 1)
```

<environment: 0x02fd1880>

```
[data = bmw]
```

Conditional Distribution:

```
norm
```

Coefficient(s):

```
mu      ar1      omega      alpha1      beta1
4.0092e-04  9.8596e-02  8.9043e-06  1.0210e-01
```

```
8.5944e-01
```

Std. Errors:

based on Hessian

Error Analysis:

| Estimate | Std. Error | t value | Pr(> t) |
|----------|------------|-----------|-------------------|
| mu | 4.009e-04 | 1.579e-04 | 2.539 0.0111* |
| ar1 | 9.860e-02 | 1.431e-02 | 6.888 5.65e-12*** |
| omega | 8.904e-06 | 1.449e-06 | 6.145 7.97e-10*** |
| alpha1 | 1.021e-01 | 1.135e-02 | 8.994 < 2e-16*** |
| beta1 | 8.594e-01 | 1.581e-02 | 54.348 < 2e-16*** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log Likelihood:

```
17757 normalized: 2.89
```

Description:

```
Mon Nov 19 10:37:53 2012 by user: Administrator
```

Standardised Residuals Tests:

| | | Statistic | p-Value |
|-------------------|-----|-------------|---------|
| Jarque-Bera Test | R | Chi^2 11378 | 0 |
| Shapiro-Wilk Test | R | W NA | NA |
| Ljung-Box Test | R | Q(10) 15.2 | 0.126 |
| Ljung-Box Test | R | Q(15) 20.1 | 0.168 |
| Ljung-Box Test | R | Q(20) 30.5 | 0.0614 |
| Ljung-Box Test | R^2 | Q(10) 5.03 | 0.889 |
| Ljung-Box Test | R^2 | Q(15) 7.54 | 0.941 |
| Ljung-Box Test | R^2 | Q(20) 9.28 | 0.979 |
| LM Arch Test | R | TR^2 6.03 | 0.914 |

Information Criterion Statistics:

```
AIC BIC SIC HQIC
-5.78 -5.77 -5.78 -5.77
```

从输出结果可知,参数估计都是显著的。因此我们的模型为:

$$X_t = 4.009 \times 10^{-4} + 0.0986X_{t-1} + a_t$$

$$a_t = \sigma_t \epsilon_t$$

$$\sigma_t = \sqrt{8.904 \times 10^{-6} + 0.1021a_{t-1}^2 + 0.8594\sigma_{t-1}^2 \epsilon_t}$$

同时,这里 beta1 的估计值为 0.8594,说明 BMW 的对数收益有持续的波动聚集现象。这里的 AIC 和 BIC 相差不大,都较小,说明模型拟合较好。

我们可以进一步用该模型预测 1987 年 11 月 15 日

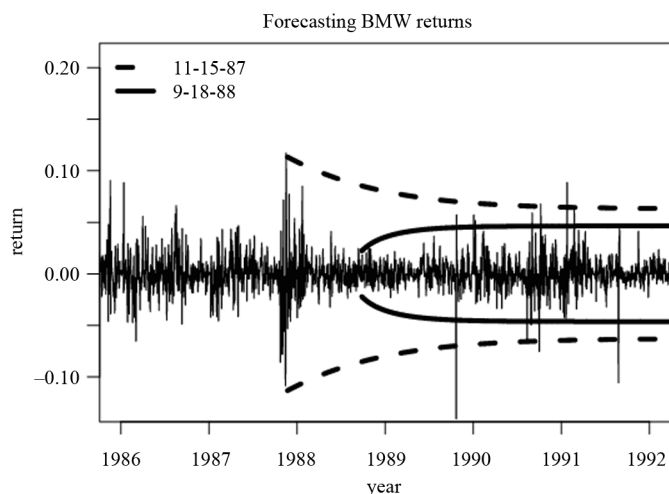


Figure 2. Prediction limits for forecasting BMW returns after Black Monday
图 2. BMW 黑色星期一之后的收益预测

(该日为 1987 年黑色星期一之后不久)和 1988 年 9 月 8 日以及它们之后的对数收益。用 SPSS 插件绘制 BMW 对数收益时序图和预测区间，并绘制它们的图形如图 2 所示。

从预测结果可知，在 1987 年 11 月 15 日由于黑色星期一的影响，收益波动很大，故此预测区间一开始较宽，随后随着波动变小，预测区间变窄。而在 1988 年 9 月 18 日的一段时期，收益的波动较小，因此初始的预测区间较小，随着波动的变大，预测区间变宽。

4. 结语

本文主要介绍了 SPSS 统计分析软件和自由软件 R 的集成的方法。通过 SPSS 的 R 插件，可以充分利

用目前两种统计分析软件的优点—利用 SPSS 的优秀图形用户界面和数据处理能力，利用 R 的强大的软件包和统计分析方法。同时本文介绍了 R 中调用 SPSS 程序的输出结果的方法，在第三部分给出了应用该插件拟合 BMW 对数收益的案例。

参考文献 (References)

- [1] SPSS Inc. SPSS statistics—R integration package. SPSS 17.0 Product CD-Rom, 2008.
- [2] R. Levesque and SPSS Inc. Programming and data management for SPSS statistics 17.0. SPSS 17.0 Product CD-Rom, 2008.
- [3] J. Verzani. Using R for introductory statistics. Boca Raton: Chapman & Hall/CRC, 2004.
- [4] D. Ruppert. Statistics and data analysis for financial engineering. Berlin: Springer, 2011.