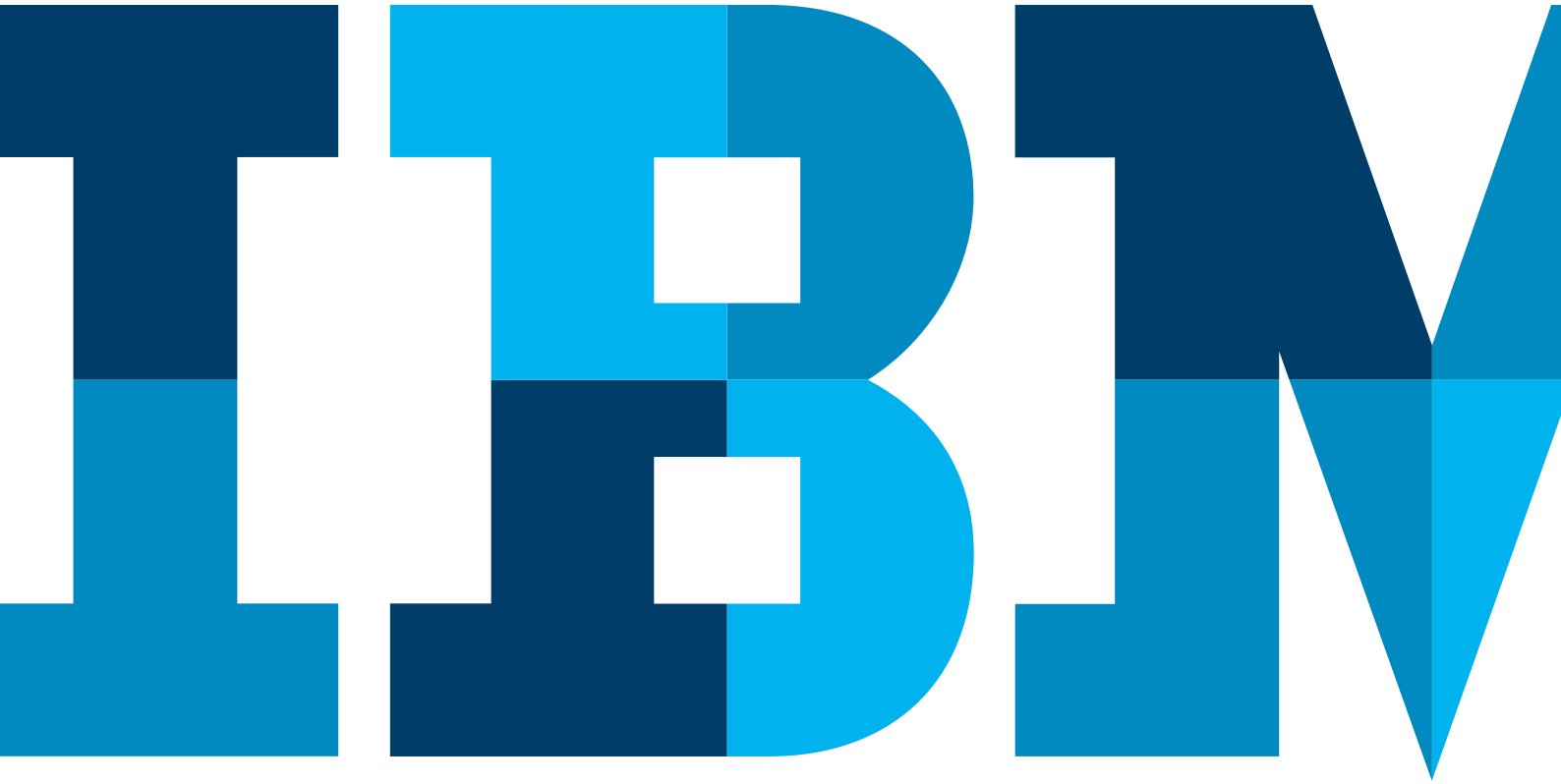
The power of IBM SPSS Statistics and R together



IBM

Contents

Executive summary

The purpose of this paper is to demonstrate the features and capabilities provided by the integration of IBM® SPSS Statistics and R. R users get the access to superb data management, ease of use and presentation quality output that is available from IBM SPSS Statistics. SPSS Statistics users get access to a rich, ever-expanding collection of statistical analysis and graphing libraries to help them gain deeper insights from their data. Using IBM SPSS Statistics and R together makes the most of both worlds.

Why integrate SPSS Statistics and R?

IBM SPSS Statistics is one of the world's leading statistical software solutions. It provides predictive models and advanced analytics to help solve business and research problems. For many businesses, research institutions and statisticians, it is the de facto standard for statistical analysis. Organisations use SPSS Statistics to:

- Understand data
- Analyse trends
- Forecast and plan
- Validate assumptions
- Drive accurate conclusions.

SPSS Statistics has been continuously developed and tested since 1968. Over that period, many forms of statistical analysis have been embedded in the software. In addition, the algorithms that execute the equations have been tested by developers and users in academia, in laboratories and in virtually every type of business. As a result, users can be confident that the software has been thoroughly tested and its results found to be reliable.

The SPSS Statistics environment makes it easier for you to quickly access, manage and analyse datasets, including survey data, corporate databases, data downloaded from the web and much more. Advanced statistical procedures and visualisation can provide a robust, user friendly and integrated platform for understanding your data and solving complex business and research problems.

IBM SPSS Statistics can help you address ***all*** facets of the analytical process from data preparation and management to analysis and reporting. It provides tailored functionality and customisable interfaces for different skill levels and functional responsibilities. It also enables users to create high-resolution graphs and presentation-ready reports to easily communicate results.

For example, consider Robert, who is interested in analysing data on miles-per-gallon (mpg) for different types of cars. Figure 1 shows a segment of what the data looks like in the SPSS Statistics Data Editor.

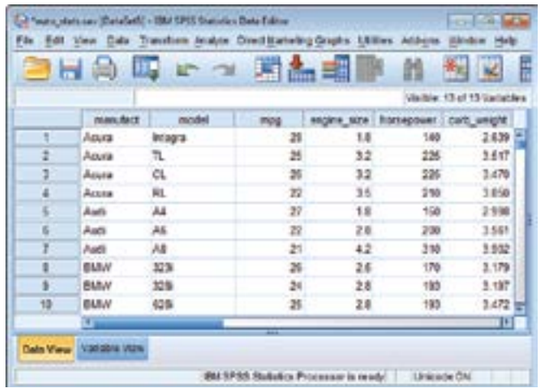


Figure 1: SPSS Statistics Data Editor displays mpg information for different makes and models.

For this analysis, Robert might first run the Descriptives procedure to get an idea of the distribution of the data for mpg. He does this from the Descriptives dialogue box (Figure 2).

Figure 2: Descriptives dialogue box.

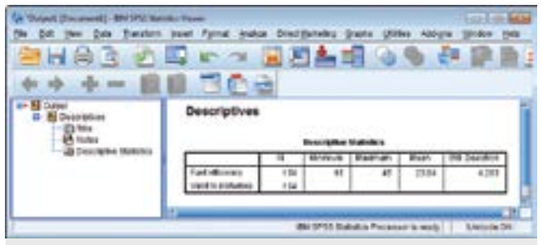


Figure 3: Output of descriptives dialogue in SPSS Statistics Viewer.

The output (in this case, tabular output) is shown in the SPSS Statistics Viewer (Figure 3).

R is an open source programming language and software environment for statistical computing and graphics (www.r-project.org). The R language has become very popular with statisticians and data miners for developing statistical software and is widely used for advanced data analysis. R provides a wide variety of advanced statistical and graphical techniques and is highly extensible. R is available as Free Software under the terms of the Free Software Foundation GNU General Public License. It runs on Windows and MacOS, a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux). R can be easily extended with packages.

Among the topics that R users commonly discuss are scalability and basic data and output management issues such as connecting to databases, improving output quality and sharing R algorithms with others not familiar with R. On the other hand, SPSS Statistics users might want to use some of the R functions that are not available within SPSS Statistics without having to learn R. After all, programming in R is not for everyone.

Therefore, integrating R with SPSS Statistics makes sense. The combined strength of both helps address the needs of the two user groups. SPSS Statistics is a convenient platform from which R users can handle large data sets and get high quality graphs and other forms of output. Some of the other benefits are the ease of use of SPSS Statistics and the ability to distribute integrated R packages to a wide range of users who are not familiar with R. This integration also provides SPSS Statistics users with easy access to nearly 4000 open source statistical functions.

Integrating R with IBM SPSS Statistics

To use R programming features with SPSS Statistics, you need the SPSS Statistics-Integration Plug-In for R. This plug-in is available at no charge and is installed as part of SPSS Statistics-Essentials for R. This plug-in is necessary if you want to use extensions written either by you or by others. The SPSS Statistics-Integration Plug-In for R is part of a family of integration plug-ins that are available at no charge. This family includes plug-ins for Python, Java and .NET. R communicates with SPSS Statistics by means of APIs in the plug-in, and the integration requires writing R wrapper code. The SPSS Statistics-Integration Plug-in for R extends the SPSS Statistics command syntax language with the full capabilities of the R programming language. The plug-in also provides access to an R integrated development environment which makes it easy for users to develop, test and debug R programs for use with SPSS Statistics. It is available for Windows, Linux, Mac OS and SPSS Statistics Server.

After installing the SPSS Statistics-Integration Plug-In for R, you can choose to use R programs that have already been written or you can write your own.

Extension bundles: Using R programs created by others

SPSS Statistics and R integration enables you to take advantage of the R programs that others have written and packaged as extension bundles. The pre-coded algorithms obviate the need for intense R programming, especially if you are pressed for time or expertise in R programming is scarce. These R programs are deployed as extension bundles.

After an extension bundle is installed, its dialogue box is accessible from the SPSS Statistics menus, and the extension command can be run as if it were any built-in command. The R program functions as if it were a native dialogue box and a syntax command.

To understand an extension bundle better, consider again the example where the user is analysing data on mpg. Suppose Robert is now interested in analysing mpg as a function of engine size, horsepower and curb weight of the vehicle. However, he wants to go beyond standard linear regression and do the analysis with quantile regression. Quantile regression is provided as an extension bundle in the integration plug-in for R, and Robert can use it to understand the distribution of mpg as a function of the predictors. For the more technically minded, quantile regression estimates one or more conditional quantiles (0 <= q < 1) for a linear model. In contrast, ordinary regression estimates only the conditional mean.

The Quantile Regression dialogue box looks just like any native SPSS Statistics dialogue box (Figure 4).

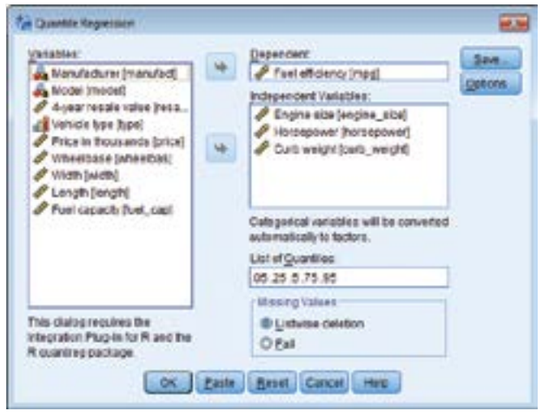


Figure 4: Quantile Regression dialog box.

Robert can simply specify the dependent and independent variables and the list of quantiles and then click OK to run the analysis. SPSS Statistics calls R and does the analysis using the R quantreg package, but Robert sees none of that. The results of the analysis from R are then presented as tabular and chart output in the SPSS Statistics Viewer. As in any typical regression analysis, results include tabular output of the regression coefficients. In this case, a separate table of regression coefficients is created for each specified quantile (Figure 5), where the table for the 0.05 quantile is displayed. The results shown in the table come from running the analysis in R, retrieving the results from R and displaying them in the SPSS Statistics Viewer.

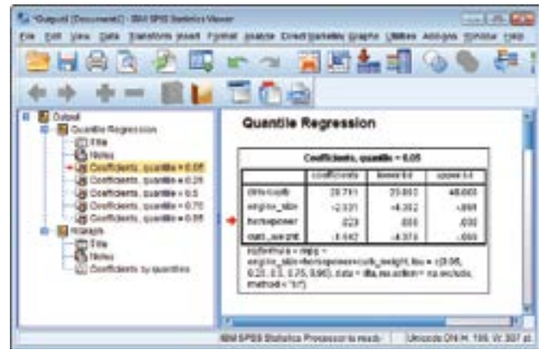


Figure 5: A table of regression coefficients.

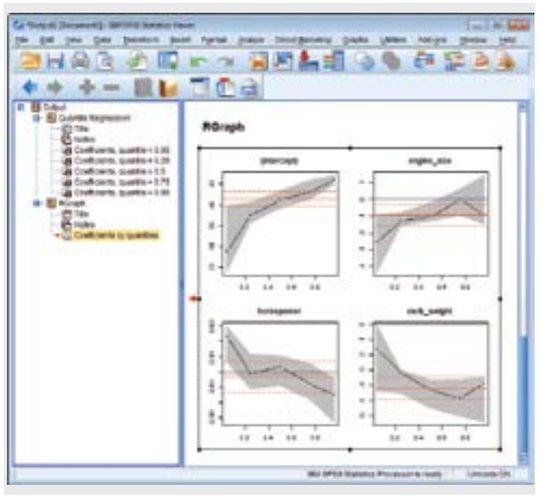


Figure 6: Chart coefficients for each of the predictors as a function of the specified quantiles.

The quantile regression procedure also produces chart output that shows the coefficients for each of the predictors as a function of the specified quantiles (Figure 6). The chart is actually generated by R but is automatically rendered in the SPSS Statistics Viewer. Again, all of that complexity is hidden from Robert, who just sees the chart output.

Where to find extension bundles

Extension bundles that implement a variety of R statistical algorithms are installed with the Essentials for R. You can find a listing of these extension bundles in the SPSS Statistics Help system under *Integration Plug-in for R Help > R Extension Commands for SPSS Statistics*.

Many more extension bundles that implement R statistical algorithms are available from the SPSS community on the IBM developerWorks site at: ibm.com/developerworks/spssdevcentral

Starting with SPSS Statistics 22, you can search for and download extension bundles, hosted on the SPSS community, from within SPSS Statistics. This feature is available from *Utilities > Extension Bundles > Download and Install Extension Bundles*. **Already installed bundles can be updated in the same way.**

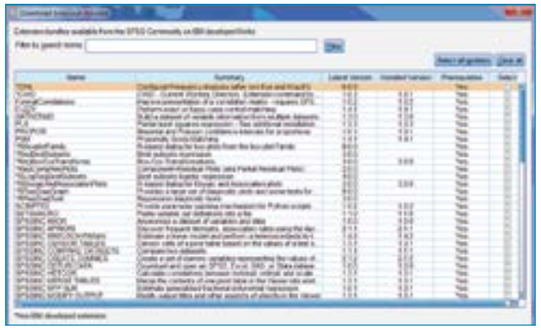


Figure 7: List of available R extension commands in SPSS Statistics.

Writing your own R programs

Using extension bundles is just one way of using R in SPSS Statistics. You can write your own R program and integrate it in SPSS Statistics at various levels by:

- Creating a custom dialogue that generates the syntax for an R extension command or explicit R code
- Creating an extension command implemented in R
- Running R code directly from within SPSS Statistics.

In these structures, you have access to both the R programming language and the functions specific to SPSS Statistics, provided in the R Integration Package for SPSS Statistics. You can also write R functions that use SPSS Statistics functionality from within R, but return results to R.

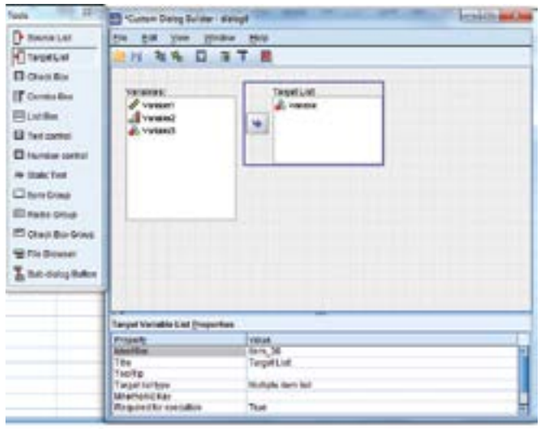


Figure 8: Example of the layout of a custom dialogue.

These functions enable you to:

- Read case data from the active dataset into R
- Get information about data in the active dataset
- Get output results from syntax commands
- Write results (back to a new dataset, to pivot table and to graphics) from R to be displayed in SPSS Statistics.

Creating a custom dialogue that generates the syntax for an R extension command or explicit R code

With the Custom Dialogue Builder, you can create a user interface that generates command syntax for an extension command implemented in R. You can then view the output (Figures 7-9) from running the dialogue in the SPSS Statistics Viewer. An R program can also be directly embedded in a custom dialogue.

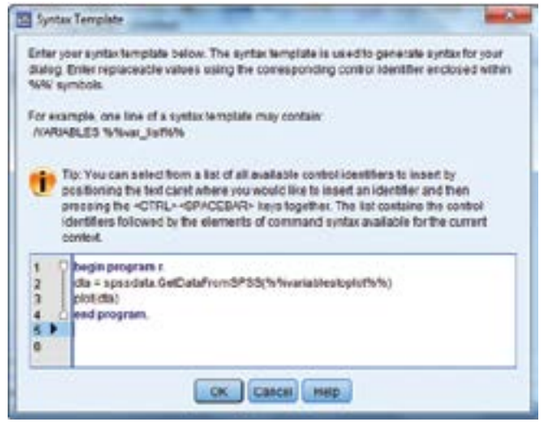


Figure 9: A syntax template for explicit R code.

Creating an extension command implemented in R

An extension command is a custom SPSS Statistics command that is implemented in R, Python or Java. You can integrate an R program into SPSS Statistics by creating an extension command that implements the R program. Integrating an R algorithm into SPSS Statistics is particularly useful when a user needs an advanced statistical function but lacks the expertise or time to create such a program. In such a scenario, a methodology group, which creates R algorithms for much needed statistical functions, could write the code and distribute it as an extension command. The user can then use it just as if it were a built-in SPSS Statistics command.

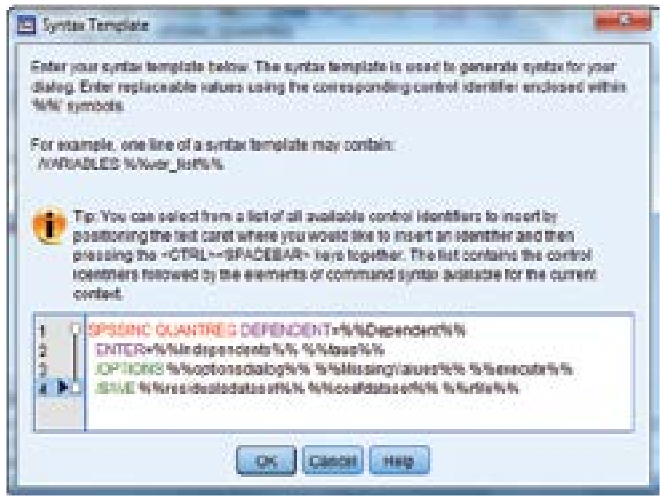


Figure 10: A syntax template for an extension command.

To create an extension command:

- Write the program as you would an R function
- Define the SPSS Statistics syntax for the extension command in an xml file that specifies the command name, the subcommands and the keywords (Figure 11)
- Declare the syntax in an “R Run Function” and call the function (Figure 12).

The user input is automatically validated and mapped to R variables and passed to the implementing function.

The R code calls APIs in the integration plug-in for R that interact with SPSS Statistics. Text in the R code that is intended for output, such as pivot table labels and error messages, can be enabled for translation.

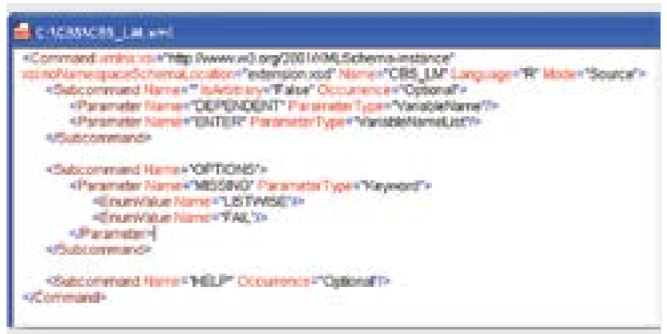


Figure 11: Syntax for the extension command.

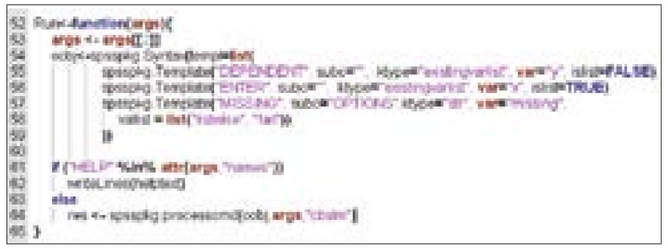


Figure 12: Declaring the syntax and calling the function.

Running R code directly from within SPSS Statistics

If you are familiar with R, you can run your own R code in SPSS Statistics. To run the R code, you enclose the code in a BEGIN PROGRAM R – END PROGRAM block of SPSS Statistics command syntax. When the syntax is submitted, the code inside the block is executed in R. The code is typically a combination of ordinary R code and calls to SPSS-specific R functions (provided with the Integration Plug-in for R) that enable R to interact with SPSS Statistics.

Figure 13 shows an example of an R program to run a regression.

The functions that facilitate the process of using R programming features with SPSS command syntax include:

- spssdata.GetDataFromSPSS (Gets data from the active dataset)
- spssdictionary.GetDictionaryFromSPSS (Gets variable dictionary information from the active dataset)
- spsspivottable.Display (Renders tabular output from R as a pivot table that can be displayed in the IBM SPSS Statistics Viewer or can be written to an external file with the SPSS Statistics Output Management System. Pivot tables produced with this function are just like pivot tables produced by native SPSS code).

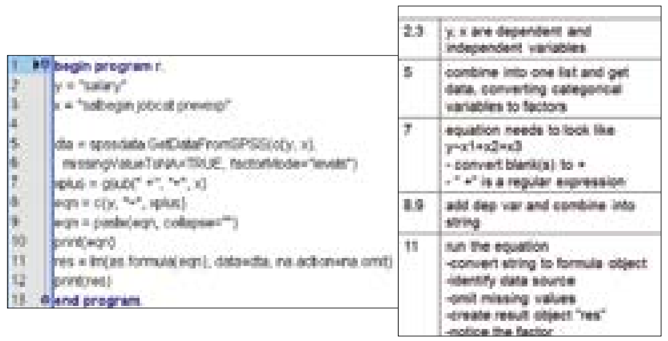


Figure 13: An R program for running a regression.

All artifacts can be easily packaged to create an extension bundle for easy distribution and installation. The extension bundle contains:

- A custom dialogue package (.spd) file that specifies the custom dialogue box
- An XML file that specifies the syntax of the extension command
- The Implementation code file(s) written in R.Other users who have installed Essentials for R can simply install the extension bundle from the SPSS Statistics menus. They can then use the dialogue and extension command in the same manner as a native dialogue or command. To enable collaboration and knowledge sharing, contributed R-based extensions can be hosted on the SPSS Community website if appropriate.

Conclusion

Both SPSS Statistics and R can independently boast strengths that have been tested over time and are strongly accepted in the statistical community. Moreover, these strengths complement each other to create an even more powerful set of statistical functions and features that benefit the statistical community as a whole.

R users can access superior data management capabilities, which enables them to handle much larger data sets. Also, the Output Management System from SPSS Statistics provides R users with a richer set of graphical and pivot table output options, which can lead to a better user experience. Finally, SPSS Statistics acts as an ideal deployment vehicle to distribute R packages to a wide range of users.

SPSS Statistics users gain access to many more statistical functions, which enables them to carry out complicated analysis without the hassles of learning a complex programming language such as R. The advantages of using R and SPSS Statistics together are many and worth considering.

Resources

The following resources are available to help users who want to use R in SPSS Statistics:

- **Instructions for getting Essentials for R are included in the SPSS Statistics Help system under** *Integration Plug-in for R Help > How to get the IBM SPSS Statistics—Integration Plug-in for R*
- **Tutorials are available from** *Help > Working with R*
- **Complete documentation for the Integration Plug-in for R is available in the SPSS Statistics Help system under "Integration Plug-in for R Help"**
- **Questions about using R in SPSS Statistics can be posted to the forum on R Programmability that is hosted in the SPSS community on developerWorks (ibm.com/developerworks/spssdevcentral)**
- **Detailed information about creating extension commands can be found in the article "Writing IBM SPSS Statistics Extension Commands," and the chapter on Extension Commands in *Programming and Data Management for IBM SPSS Statistics*, both of which are available from the SPSS community**
- **A list of extension bundles that are available for download from the SPSS community is provided in the article "Extension Bundles from IBM SPSS," which can be found in the SPSS community. The list includes extension bundles that implement Python extension commands along with those that implement R extension commands.**

About Business Analytics

IBM Business Analytics software delivers data-driven insights that help organisations work smarter and outperform their peers. This comprehensive portfolio includes solutions for business intelligence, predictive analytics and decision management, performance management and risk management.

Business Analytics solutions enable companies to identify and visualise trends and patterns in such areas as customer analytics that can have a profound effect on business performance. They can compare scenarios; anticipate potential threats and opportunities; better plan, budget and forecast resources; balance risks against expected returns and work to meet regulatory requirements. By making analytics widely available, organisations can align tactical and strategic decision making to achieve business goals.
For more information, see **ibm.com**/business-analytics.

Request a call

To request a call or to ask a question, go to **ibm.com**/business-analytics/contactus. An IBM representative will respond to your inquiry within two business days.

IBM United Kingdom Limited
PO Box 41, North Harbour
Portsmouth, Hampshire PO6 3AU
United Kingdom

IBM Ireland Limited
Oldbrook House
24-32 Pembroke Road
Dublin 4

IBM Ireland registered in Ireland under company number 16226.

YTW03350-GBEN-01