

# IBM SPSS Statistics and open source: A powerful combination

Let's go →

# Executive summary

The purpose of this paper is to demonstrate the features and capabilities provided by the integration of IBM® SPSS® Statistics and open source programming languages like R and Python. Open source users get the access to superior 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 open source together makes the most of both worlds.



---

## Executive summary

---

Why integrate SPSS Statistics and open source?

---

Integrating open source with IBM SPSS Statistics

---

Extension bundles:  
Using R or Python programs created by others

---

Writing your own R or Python programs

---

Conclusion

---



# Why integrate SPSS Statistics and open source?

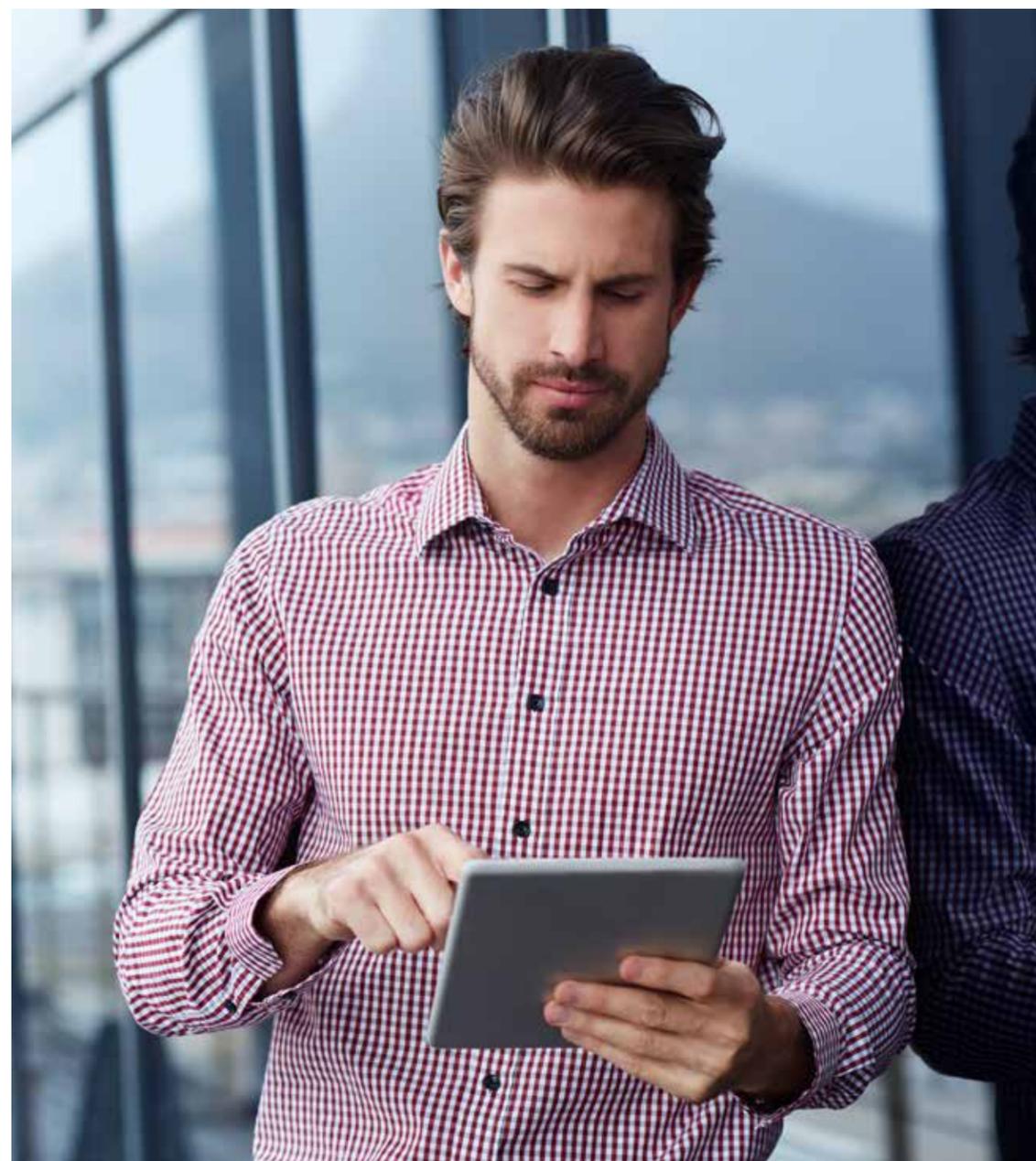
IBM SPSS Statistics is one of the world's leading statistical software solutions. It provides predictive models and advanced analytics to help solve business and research problems. For many businesses, research institutions and statisticians, it is the de facto standard for statistical analysis. Organizations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

- Understand data
- Analyze 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 analyze datasets, including survey data, corporate databases, data downloaded from

the web and much more. Advanced statistical procedures and visualization can provide a robust, user friendly and integrated platform for understanding your data and solving complex business and research problems.



---

Executive summary

---

**Why integrate SPSS Statistics and open source?**

1 • 2 • 3 • 4

---

Integrating open source with IBM SPSS Statistics

---

Extension bundles:  
Using R or Python programs created by others

---

Writing your own R or Python programs

---

Conclusion

---



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 customizable 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 analyzing 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.

	manufact	model	mpg	engine_size	horsepower	curb_weight
1	Acura	Integra	28	1.8	140	2.639
2	Acura	TL	25	3.2	225	3.517
3	Acura	CL	26	3.2	225	3.470
4	Acura	RL	22	3.5	210	3.850
5	Audi	A4	27	1.8	150	2.998
6	Audi	A6	22	2.8	200	3.561
7	Audi	A8	21	4.2	310	3.902
8	BMW	323i	26	2.5	170	3.179
9	BMW	328i	24	2.8	193	3.197
10	BMW	528i	25	2.8	193	3.472

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

Executive summary

**Why integrate SPSS Statistics and open source?**

1 • 2 • 3 • 4

Integrating open source with IBM SPSS Statistics

Extension bundles:  
Using R or Python programs created by others

Writing your own R or Python programs

Conclusion



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 dialog box (Figure 2).

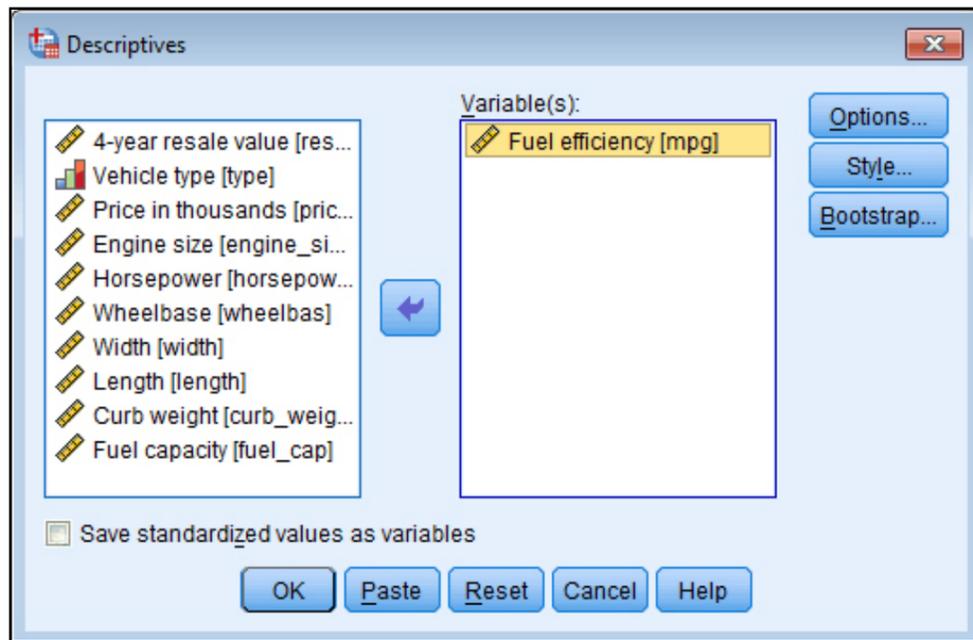


Figure 2: Descriptives dialog box.

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

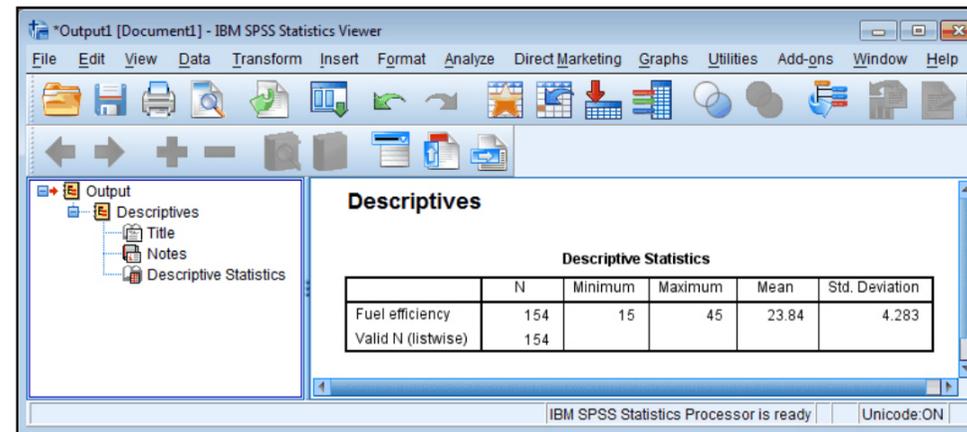


Figure 3: Output of descriptives dialog in SPSS Statistics Viewer.

R is an open source programming language and software environment for statistical computing and graphics ([www.r-project.org](http://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](https://www.gnu.org/licenses/gpl-3.0.html). 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.

Executive summary

Why integrate SPSS Statistics and open source?

1 • 2 • 3 • 4

Integrating open source with IBM SPSS Statistics

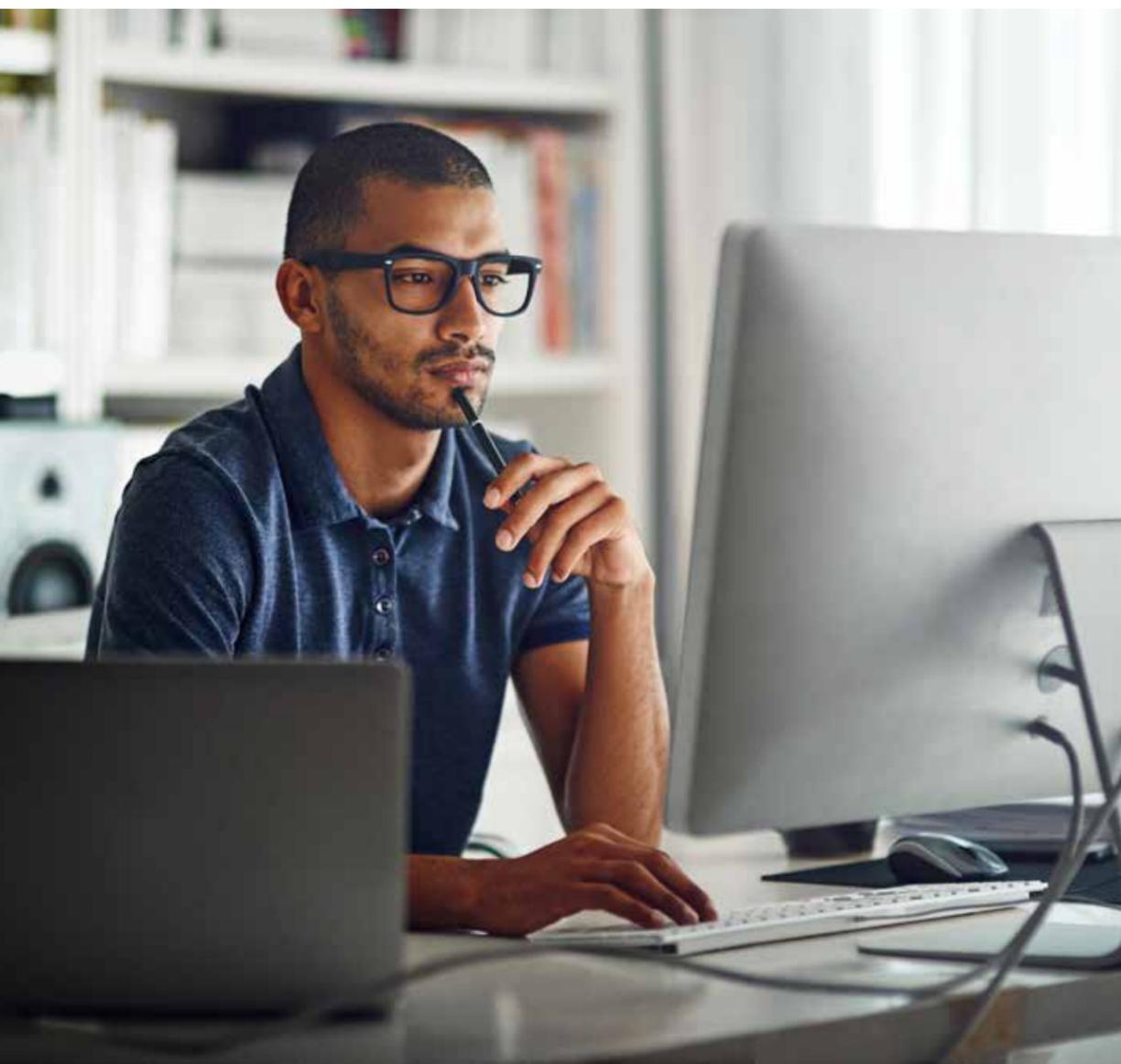
Extension bundles:  
Using R or Python programs created by others

Writing your own R or Python programs

Conclusion



On the other hand, Python is a general purpose and high level programming language (<https://www.python.org/>). You can use Python for developing desktop GUI applications, websites and web applications. Also, Python, as a high level programming language, allows you to focus on core functionality of the application by taking care of common programming tasks. The simple syntax rules of the programming language further makes it easier for you to keep the code base readable and application maintainable.



Among the topics that open source (R and Python) users commonly discuss are scalability, basic data and output management issues such as connecting to databases, improving output quality and sharing algorithms with others not familiar with programming. 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 programming. After all, programming in open source is not for everyone.

Therefore, integrating R or Python 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 or Python 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 or Python packages to a wide range of users who are not familiar with R or Python. This integration also provides SPSS Statistics users with easy access to nearly 4,000 open source statistical functions.

---

Executive summary

---

**Why integrate SPSS Statistics and open source?**

1 • 2 • 3 • 4

---

Integrating open source with IBM SPSS Statistics

---

Extension bundles:  
Using R or Python programs created by others

---

Writing your own R or Python programs

---

Conclusion

---



# Integrating open source with IBM SPSS Statistics

To use R or Python programming features with SPSS Statistics, you need the SPSS Statistics-Integration Plug-In for R and SPSS Statistics-Integration Plug-In for Python respectively.

The plug-in for R is available at no charge and is installed by default 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. 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.

The SPSS Statistics-Integration Plug-In for Python works similarly.



---

Executive summary

---

Why integrate SPSS Statistics and open source?

---

**Integrating open source with IBM SPSS Statistics**

---

Extension bundles:  
Using R or Python programs created by others

---

Writing your own R or Python programs

---

Conclusion

---



# Extension bundles: Using R or Python programs created by others

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

After an extension bundle is installed, its dialog 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 or Python program functions as if it were a native dialog box and a syntax command.

To understand an extension bundle better, consider again the example where the user is analyzing data on mpg. Suppose Robert is now interested in analyzing 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 ( $0 \leq q < 1$ ) for a linear model. In contrast, ordinary regression estimates only the conditional mean.

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

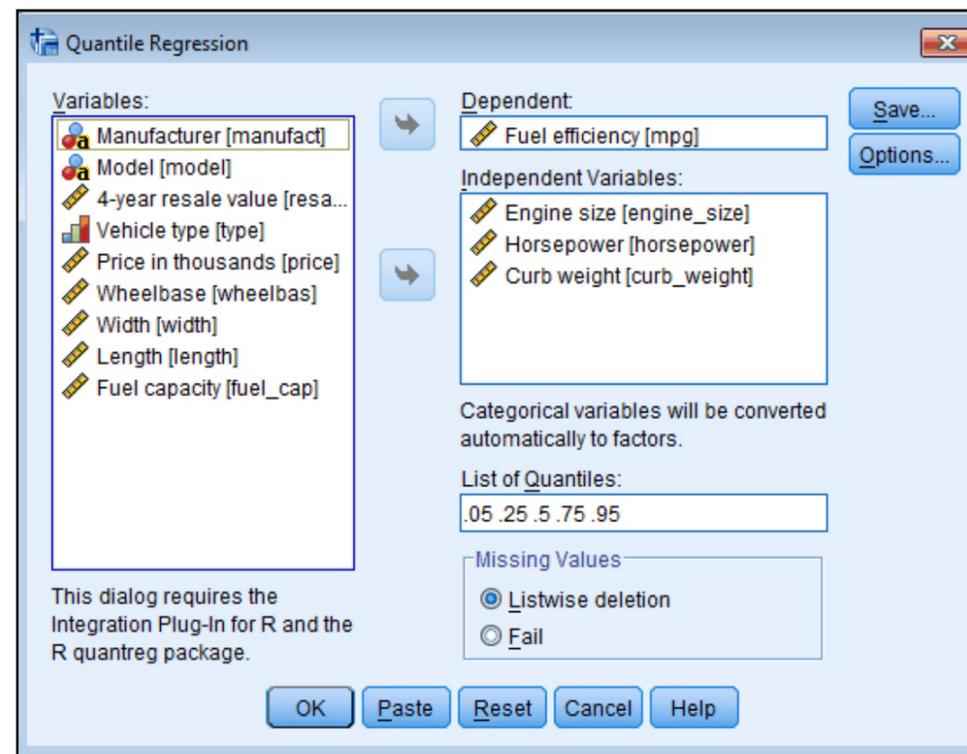


Figure 4: Quantile Regression dialog box.

Executive summary

Why integrate SPSS Statistics and open source?

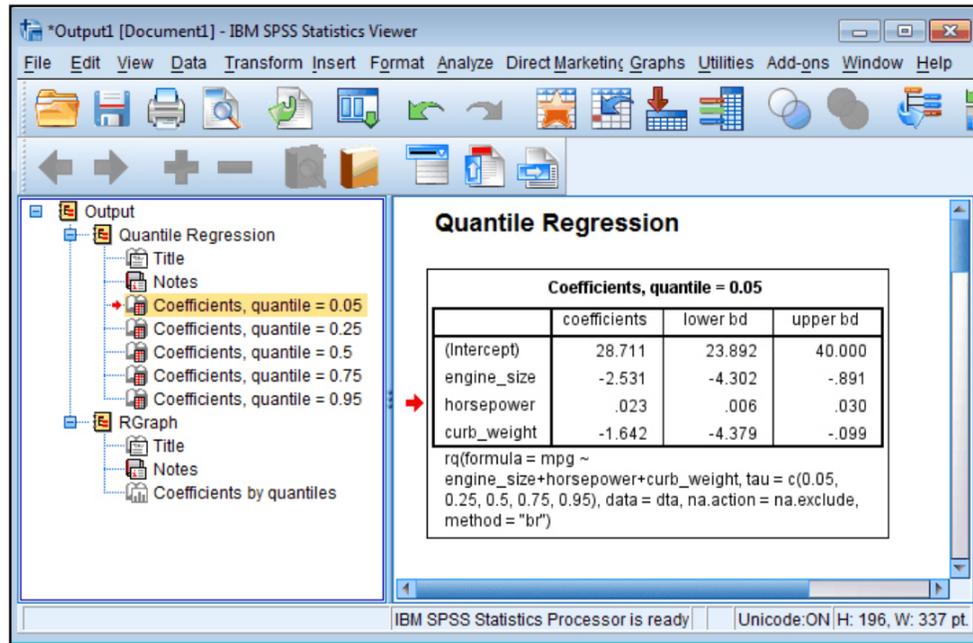
Integrating open source with IBM SPSS Statistics

**Extension bundles:  
Using R or Python programs  
created by others**

1 • 2 • 3

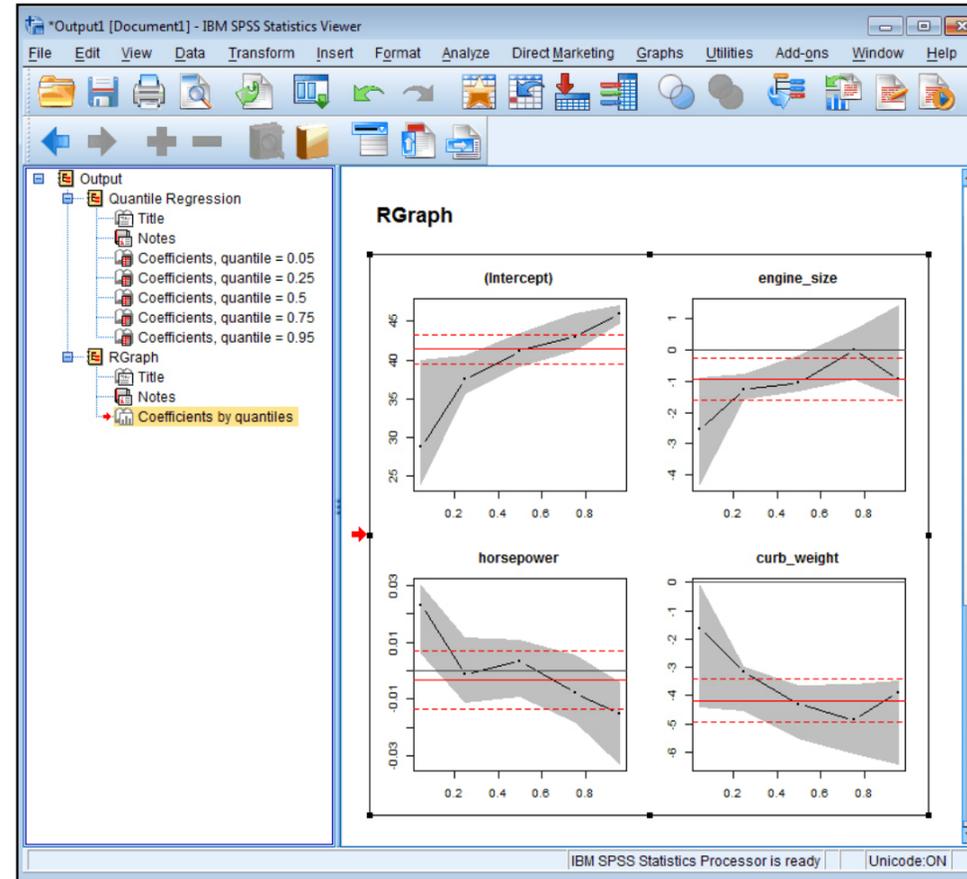
Writing your own R or Python programs

Conclusion



**Figure 5:** A table of regression coefficients.

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 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.

Executive summary

Why integrate SPSS Statistics and open source?

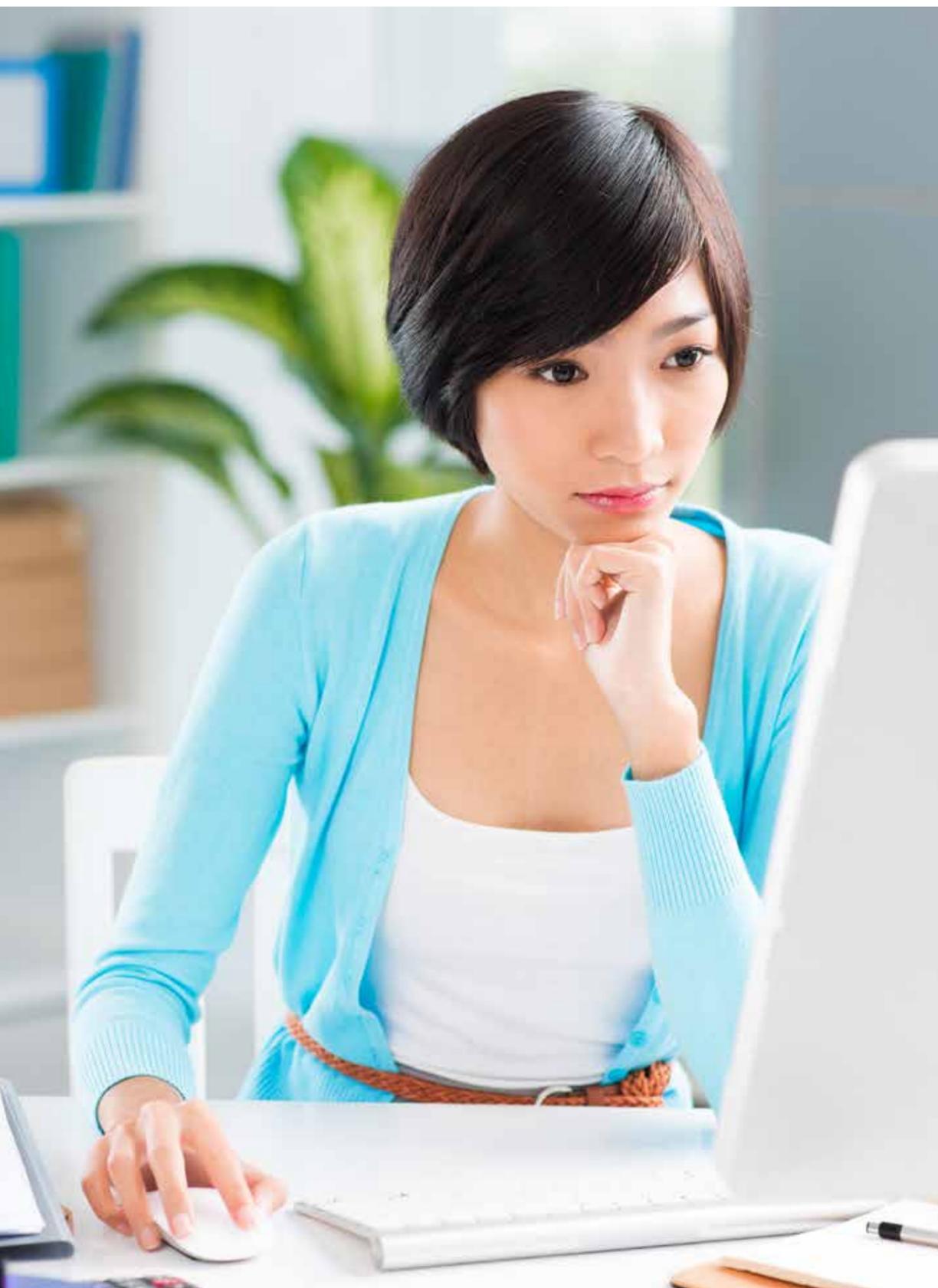
Integrating open source with IBM SPSS Statistics

Extension bundles:  
Using R or Python programs created by others

1 • 2 • 3

Writing your own R or Python programs

Conclusion



---

## Where to find extension bundles

Extension bundles that implement a variety of R or Python statistical algorithms are installed with Essentials for R or Essentials for Python.

Many more extension bundles that implement R or Python statistical algorithms are available from the SPSS community on the IBM developerWorks® site at: [developer.ibm.com/predictiveanalytics/downloads/](https://developer.ibm.com/predictiveanalytics/downloads/)

You can also search for and download extension bundles, hosted on the SPSS community, from within SPSS Statistics. This feature is available from Extensions > Extension Bundles. Already installed bundles can be updated in the same way.

---

Executive summary

---

Why integrate SPSS Statistics and open source?

---

Integrating open source with IBM SPSS Statistics

---

**Extension bundles:  
Using R or Python programs  
created by others**

1 • 2 • 3

---

Writing your own R or Python programs

---

Conclusion

---



# Writing your own R or Python programs

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

- Creating a custom dialog that generates the syntax for an R or Python extension command or explicit R or Python code
- Creating an extension command
- Running R or Python code directly from within SPSS Statistics

These functions enable you to:

- Read case data from the active dataset into R or Python
- 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 or Python to be displayed in SPSS Statistics



---

Executive summary

---

Why integrate SPSS Statistics and open source?

---

Integrating open source with IBM SPSS Statistics

---

Extension bundles:  
Using R or Python programs created by others

---

**Writing your own R or Python programs**

**1 • 2 • 3 • 4 • 5**

---

Conclusion

---



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

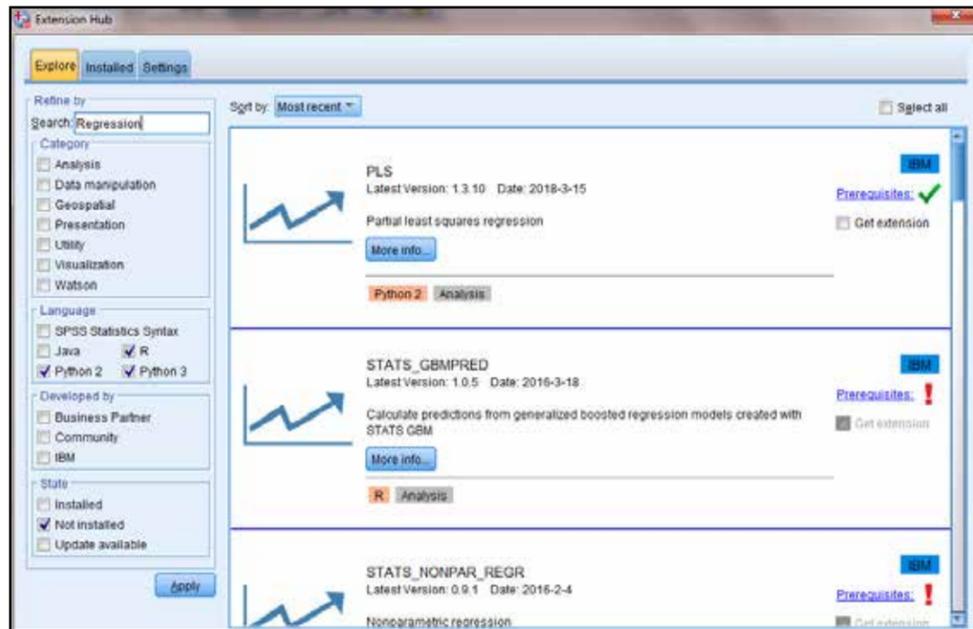


Figure 7: List of available R extension commands in SPSS Statistics.

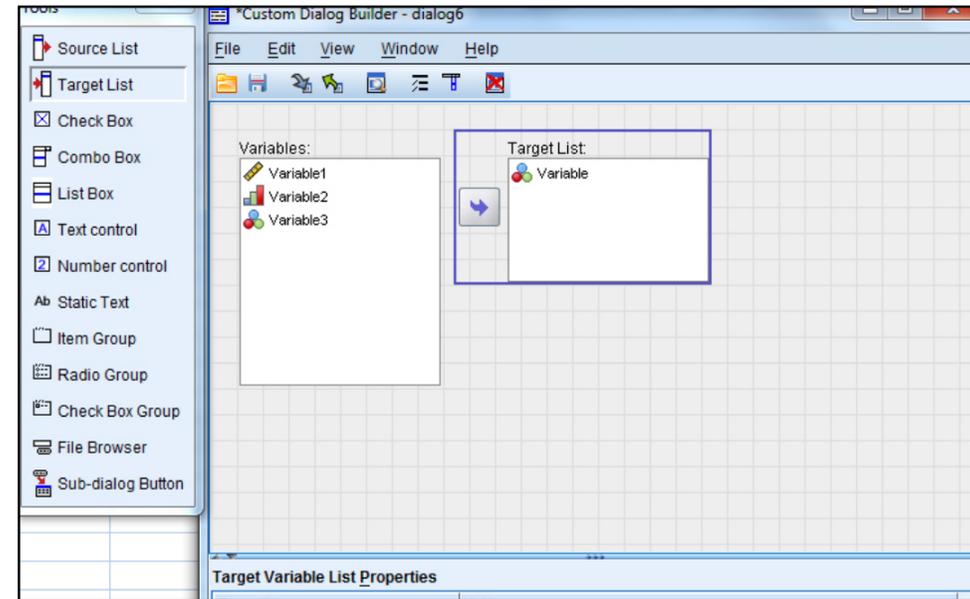


Figure 8: Example of the layout of a custom dialog.

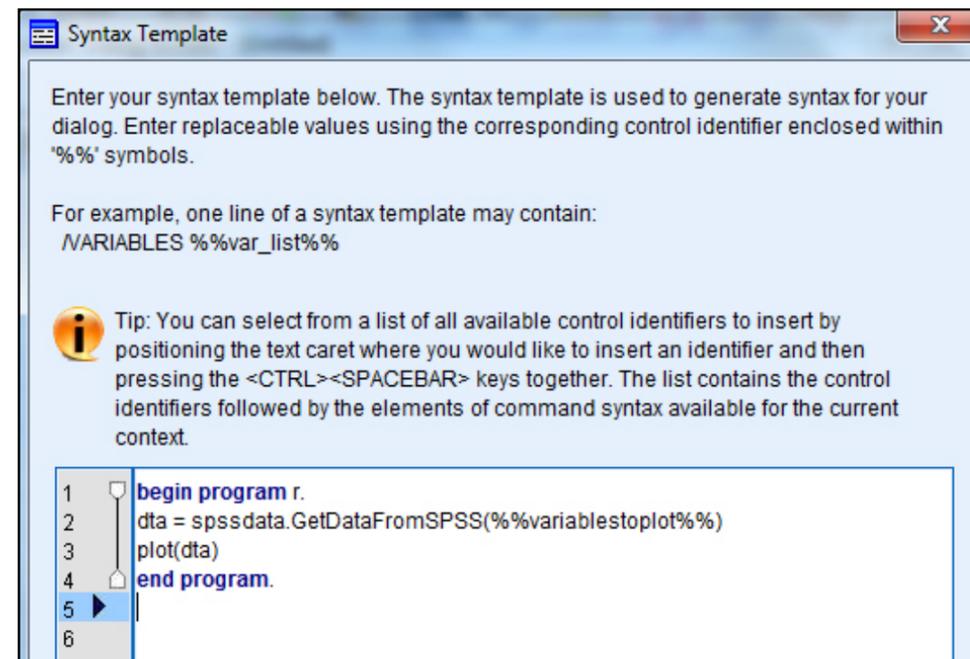


Figure 9: A syntax template for explicit R code.

Executive summary

Why integrate SPSS Statistics and open source?

Integrating open source with IBM SPSS Statistics

Extension bundles: Using R or Python programs created by others

Writing your own R or Python programs

1 • 2 • 3 • 4 • 5

Conclusion



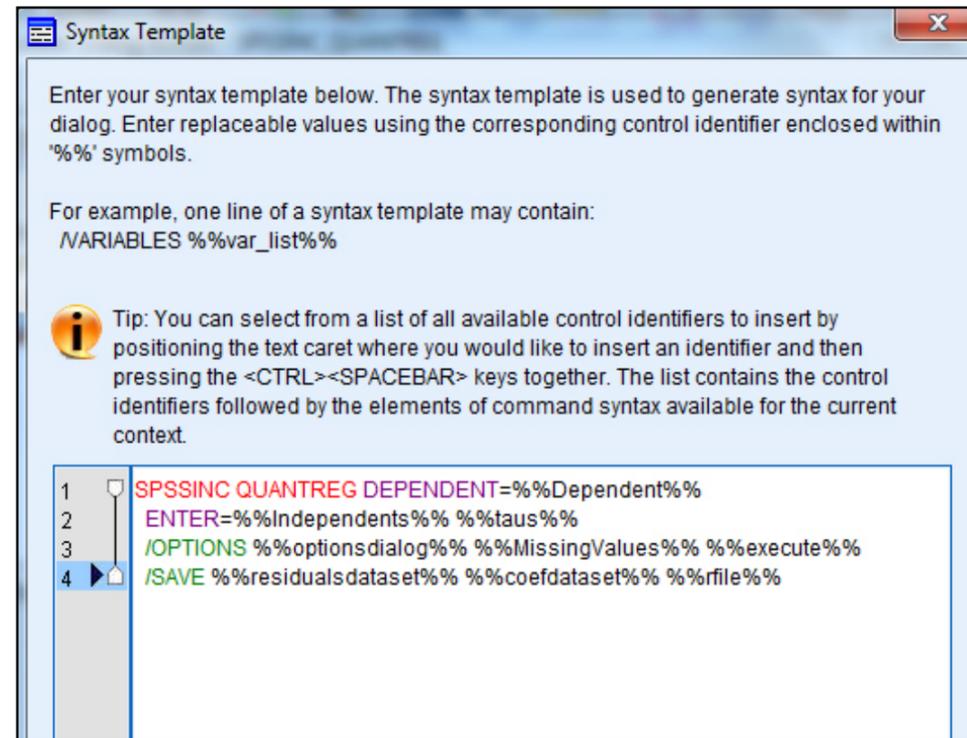
## Creating an extension command

An extension command is a custom SPSS Statistics command that is implemented in R, Python or Java. You can integrate an R or Python program into SPSS Statistics by creating an extension command that implements the R or Python program (Figure 10). Integrating an R or Python 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 or Python 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.

To create an extension command:

- Write the program as you would an R or Python 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” or “Python Run Function” and call the function (Figure 12)

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



**Figure 10:** A syntax template for an extension command.

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

Executive summary

Why integrate SPSS Statistics and open source?

Integrating open source with IBM SPSS Statistics

Extension bundles: Using R or Python programs created by others

Writing your own R or Python programs

1 • 2 • 3 • 4 • 5

Conclusion

```

C:\CBS\CBS_LM.xml
<Command xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="extension.xsd" Name="CBS_LM" Language="R" Mode="Source">
  <Subcommand Name="" IsArbitrary="False" Occurrence="Optional">
    <Parameter Name="DEPENDENT" ParameterType="VariableName"/>
    <Parameter Name="ENTER" ParameterType="VariableNameList"/>
  </Subcommand>

  <Subcommand Name="OPTIONS">
    <Parameter Name="MISSING" ParameterType="Keyword">
      <EnumValue Name="LISTWISE"/>
      <EnumValue Name="FAIL"/>
    </Parameter>
  </Subcommand>

  <Subcommand Name="HELP" Occurrence="Optional"/>
</Command>

```

**Figure 11:** Syntax for the extension command.

```

52 Run<-function(args){
53   args <- args[[2]]
54   oobj<-spsspkg.Syntax(temp=list(
55     spsspkg.Template("DEPENDENT", subc="", ktype="existingvarlist", var="y", islist=FALSE),
56     spsspkg.Template("ENTER", subc="", ktype="existingvarlist", var="x", islist=TRUE),
57     spsspkg.Template("MISSING", subc="OPTIONS", ktype="str", var="missing",
58       vallist = list("listwise", "fail")),
59   ))
60
61   if ("HELP" %in% attr(args,"names"))
62     writeLines(helptext)
63   else
64     res <- spsspkg.processcmd(oobj,args,"cbslm")
65 }

```

**Figure 12:** Declaring the syntax and calling the function.

## Running R or Python 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. Similarly, to run Python code, you enclose the code in a BEGIN PROGRAM PYTHON - END PROGRAM block of SPSS Statistics command syntax.

When the syntax is submitted, the code inside the block is executed in R or Python respectively. The code is typically a combination of ordinary R or Python code and calls to SPSS-specific R functions (provided with the Integration Plug-in for R) or Python functions (provided with the integration plug-in for Python) respectively that enable R or Python to interact with SPSS Statistics.

*Interaction between R or Python and SPSS Statistics enables your programmers to write code unique to your organization and execute it with common SPSS command syntax.*

Executive summary

Why integrate SPSS Statistics and open source?

Integrating open source with IBM SPSS Statistics

Extension bundles: Using R or Python programs created by others

Writing your own R or Python programs

1 • 2 • 3 • 4 • 5

Conclusion



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)

<pre>1 begin program r. 2 y = "salary" 3 x = "salbegin jobcat preexp" 4 5 dta = spssdata.GetDataFromSPSS(c(y, x), 6   missingValueToNA=TRUE, factorMode="levels") 7 xplus = gsub(" ", "+", x) 8 eqn = c(y, "=", xplus) 9 eqn = paste(eqn, collapse="") 10 print(eqn) 11 res = lm(as.formula(eqn), data=dta, na.action=na.omit) 12 print(res) 13 end program</pre>	<p>2,3 y, x are dependent and independent variables</p> <p>combine into one list and get data, converting categorical variables to factors</p> <p>equation needs to look like <math>y=x1+x2+x3</math> - convert blank(s) to + - " " is a regular expression</p> <p>add dep var and combine into string</p> <p>run the equation -convert string to formula object -identify data source -omit missing values -create result object "res" -notice the factor</p>
---	--

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 dialog package (.spd) file that specifies the custom dialog box
- An XML file that specifies the syntax of the extension command
- The Implementation code file(s) written in R or Python

Other users who have installed Essentials for R can simply install the extension bundle from the SPSS Statistics menus. They can then use the dialog and extension command in the same manner as a native dialog or command. To enable collaboration and knowledge sharing, contributed R- or Python-based extensions can be hosted on the SPSS Community website if appropriate.

*Extension bundles make it easy to distribute R or Python programs to users who don't have open source skills.*

Executive summary

Why integrate SPSS Statistics and open source?

Integrating open source with IBM SPSS Statistics

Extension bundles: Using R or Python programs created by others

Writing your own R or Python programs

1 • 2 • 3 • 4 • 5

Conclusion



# Conclusion

Both SPSS Statistics and open source (R or Python) 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 or Python 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 or Python 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 or Python. The advantages of using open source (R or Python) and SPSS Statistics together are many and worth considering.

## Try SPSS Statistics now

To start your 14-day free trial of SPSS Statistics, visit: [ibm.com/spssstatistics/free](https://ibm.com/spssstatistics/free)

## For more information

To learn more about using SPSS Statistics with open source (R or Python) programming languages and to access a rich set of resources, visit the following websites:

<https://developer.ibm.com/predictiveanalytics/>

<http://ibmpredictiveanalytics.github.io/>

---

Executive summary

---

Why integrate SPSS Statistics and open source?

---

Integrating open source with IBM SPSS Statistics

---

Extension bundles: Using R or Python programs created by others

---

Writing your own R or Python programs

---

**Conclusion**

---



© Copyright IBM Corporation 2018

IBM Corporation  
New Orchard Road  
Armonk, NY 10504

Produced in the United States of America  
December 2018

IBM, the IBM logo, ibm.com, developerWorks and SPSS are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at [ibm.com/legal/copytrade.shtml](http://ibm.com/legal/copytrade.shtml)

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

UNIX is a registered trademark of The Open Group in the United States and other countries.

It is the user’s responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

YTW03350-USEN-02

---

Executive summary

---

Why integrate SPSS Statistics and open source?

---

Integrating open source with IBM SPSS Statistics

---

Extension bundles:  
Using R or Python programs created by others

---

Writing your own R or Python programs

---

Conclusion

---