

Revolution R Enterprise 5.1 Beta 2 Overview

Revolution R Enterprise includes:

- Latest stable version of Open-Source R (currently 2.13.2; move to 2.14.2 mid-March)
- High performance math libraries
- RevoScaleR package: Scalable data management and analysis (same code scales from data frames to local, high performance .xdf files to distributed context on a cluster).
- R Productivity Environment (RPE) on Windows only: an Integrated Development Environment for R based on Visual Studio technology, including a visual debugger for R

Beta Plans

Beta 1 (mid-Feb.):

- “Big data” Generalized Linear Models (GLM) in the RevoScaleR package

Beta 2 (early March):

- Enhancements/bug fixes for GLM in RevoScaleR

Beta 3 and later (starting mid-March):

- Support for distributed computing with Platform Computing’s LSF workload manager for Red Hat Enterprise Linux in the RevoScaleR package. This will complement the currently supported distributed computing on Microsoft HPC clusters.
- Support for Azure Burst with Microsoft HPC clusters
- Enhancements/bug fixes for GLM in RevoScaleR
- Enhancements/bug fixes for LSF support in RevoScaleR
- R 2.14.2

Feature Overviews

- **Big Data Generalized Linear Models**
 - The new RevoScaleR function, **rxGlm**, provides a fast, scalable, distributable implementation of generalized linear models
 - Impressive speed-ups relative to glm on in-memory data frames.
 - Example: Tweedie family with 1 million observations and 78 estimated coefficients (categorical data) took 17 seconds with **rxGlm** compared with 377 seconds for **glm** on a quadcore laptop.
 - Scale your analysis to tens or hundreds of millions of observations without increasing memory requirements by storing your data in a native .xdf file,
 - Automatically distribute computations across nodes of a cluster (currently Microsoft HPC Server) for even faster analysis
 - For examples using **rxGlm**, see Chapter 9 of the *RevoScaleR User’s Guide*.

Changes from Beta 1:

- The ***rxPredict*** function now returns either an RxDfData data source or a data frame, for consistency with other RevoScaleR functions.
- The ***rxFactors*** function now can be used to easily convert all factors to have sorted levels; simply use `rxFactors(data, sortLevels=TRUE)`.
- The ***rxGlm*** function now supports offset arguments to be specified as part of the formula and with transformation functions (such as `log(x)`). ***rxPredict*** now properly handles the *offset* argument for ***rxGlm*** objects.
- A code snippet for ***rxGlm*** is now available in the R Productivity Environment.
- ODBC import has improved memory usage.
- ***rxLogit*** and ***rxGlm*** now report correct degrees of freedom if there are *fweights*.
- The *writeModelVars* argument for ***rxPredict*** is now supported when the input data is a data frame.
- A bug has been fixed in ***rxHistogram***, where under some circumstances the lowest value was being omitted.

Changes from Previous Versions:

- The ***rxLogit*** function now returns p-values based on the Z-statistic rather than on the t-statistic for consistency with ***rxGlm***.
- Distributed computing job objects have changed; jobs created with Revolution R Enterprise Version 5.0.x are not compatible with Revolution R Enterprise 5.1. If you have older jobs on your HPC Server cluster, extract all data from them and delete them before installing Revolution R Enterprise 5.1.

Known Issues:

- The ***rxGlm*** function does not handle a model containing only an intercept.