

Package ‘RthroughExcelWorkbooksInstaller’

February 14, 2012

Type Package

Title Excel Workbooks supporting Statistics courses using 'R through Excel'

Version 1.2-6

Date 2011-02-06

Author Richard M. Heiberger <rmh@temple.edu> and Erich Neuwirth
<erich.neuwirth@univie.ac.at>

Maintainer Richard M. Heiberger <rmh@temple.edu>

Depends R (>= 2.12.1), rcom (>= 1.5-2), HH (>= 2.1-27)

OS_type windows

Imports RExcelInstaller (>= 3.1-12)

Description Workbooks in Excel that illustrate statistical concepts by accessing R functions from Excel. These workbooks use the automatic recalculation mode of Excel to update calculations and graphs in R. This R package downloads an executable which installs the workbooks on MS Windows systems where RExcel has already been installed.

SystemRequirements Windows, Excel >= 2002. Excel 2010. Excel 2007 requires MS Office 2007 SP >= 1. Excel 2002 and Excel 2003 require MS Office 2003 SP >= 3.

License LGPL

Repository CRAN

Date/Publication 2011-02-07 09:50:15

R topics documented:

RthroughExcelWorkbooksInstaller-package	2
AEdotplot	3
betaWeighted	5
Examples	6

linreg	7
normal.and.t	9
StudentData	11
var	12

Index	16
--------------	-----------

RthroughExcelWorkbooksInstaller-package
Install Excel Workbooks using R and Excel

Description

Install workbooks in Excel that illustrate statistical concepts by accessing R functions from Excel. These workbooks use the automatic recalculation mode of Excel to update calculations and graphs in R. This R package installs the workbooks on MS Windows systems where RExcel has already been installed.

Usage

```
installRthroughExcel()
```

Details

`installRthroughExcel` downloads and runs an executable file `'RthroughExcelWorkbooksInst.latest.exe'` constructed with the INNO Setup. It copies the Excel workbooks into the `'RExcel/R.and.Excel/'` directory, which by default is installed under `'C:/Program Files/'`.

`installRthroughExcel` also installs a certificate for code signing. This certificate is issued by "R and Excel book (Heiberger-Neuwirth)". All the Excel workbooks installed with this package are signed with this certificate and therefore can be run from any directory almost independently of the security settings of Excel.

Each of the workbooks is included with both extensions `' .xls'` (for Excel 2002 and Excel 2003) and `' .xslm'` (for Excel 2007 and 2010). The Workbooks are accessed by clicking "RthroughExcel Worksheets" from the RExcel menu item on the Excel 2002 or Excel 2003 menu bar or on the Add-Ins tab of the Excel 2007 or 2010 menu bar.

BookFilesTOC: Table of Contents, with an Excel button for each of the other workbooks and for documentation for each of the other workbooks.

StudentData: Measurements on 1126 Austrian undergraduates over the past ten years. Data collected by Erich Neuwirth. See [StudentData](#) for details.

linreg: Excel workbook with dynamic controls to illustrate simple linear regression, with emphasis on the concept of least squares, and the concept of leverage. The workbook uses the R graphical functions in [regr1.plot](#). The dynamic controls use RExcel to interface the controls in Excel with the graphics in R. See [linreg](#) for details.

normal.and.t: Workbook with dynamic controls to illustrate hypothesis testing and confidence intervals for the normal and t distributions. The workbook uses the R graphical functions in [norm.curve](#).

The dynamic controls use RExcel to interface the controls in Excel with the graphics in R. See [normal.and.t](#) for details.

AEdotplot: Workbook for the Adverse Effects Dotplot [AE.dotplot](#). When the user clicks a column of the data in the Excel worksheet, the RExcel interface immediately redraws the graph (in the R graphics window) sorted according to the clicked column. See [AEdotplot](#) for details.

Examples: Sample Excel Workbooks to Illustrate Data Transfer from Excel to R. See [Examples](#) for details.

betaWeighted: Simple Linear Regression Slope as Weighted Sum—Excel Workbook using R and Excel. See [betaWeighted](#) for details.

var: Variance Calculation using R and Excel (2002, 2003, 2007). See [var](#) for details.

Author(s)

Richard M. Heiberger, Temple University, Philadelphia <rmh@temple.edu> and Erich Neuwirth, University of Vienna, <erich.neuwirth@univie.ac.at>

References

Thomas Baier, Richard Heiberger, Kerstin Schinagl, Erich Neuwirth (2006). "Using R for teaching statistics to nonmajors: Comparing experiences of two different approaches." UserR! Conference, Vienna June 2006. <<http://www.r-project.org/useR-2006/Slides/BaierEtAl.pdf>>.

Thomas Baier and Erich Neuwirth (2007). Excel :: COM :: R. Computational Statistics, Volume 22, Number 1/April 2007. Physica Verlag.

Richard M. Heiberger and Erich Neuwirth (2009). *R through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer. <http://www.springer.com/978-1-4419-0051-7>

See Also

[RExcelInstaller-package](#)

AEdotplot

AE (Adverse Effects) Dotplot of incidence and relative risk using R and Excel

Description

RExcel interface to the AE.dotplot in the R **HH** package.

Details

From the RExcel menu item (on the main menubar in Excel 2003, or on the Add-Ins menubar in Excel 2007 or 2010), click 'RthroughExcel Workbooks', then click the 'AE dotplot' button. All Excel workbooks in the 'RthroughExcel Workbooks' directory are signed with a certificate issued by "R and Excel book (Heiberger-Neuwirth)". The worksheet in the workbook contains a copy of the data in the example in [AE.dotplot](#). The worksheet also contains RExcel code that sends the data from the worksheet to R and draws the Adverse Effects Dotplot described in [AE.dotplot](#). On

doubleclick of a column name (in cells 'A5:N5'), the data is sorted by that column and then sent over to R for the graph to be redrawn in the new sort order. The variable name used for the sort is place into the main title of the graph. The names of Treatments A and B are placed by the user into cells 'B2:B3'. The names of the treatments and the patient counts in each are placed into the legend of the plot.

The data in the columns of the "data" worksheet are in wide format. They are automatically reshaped by the R code in the "code" worksheet to the long format that is used by the `ae.dotplot` function. The variables are:

'Event' Name of Adverse Event

'PCT A' Percent of patients in Treatment A for whom the event was observed.

'PCT B' Percent of patients in Treatment B for whom the event was observed.

'N A' number of patients in treatment group A.

'AE A' number of patients in treatment group A for whom the event was observed.

'N B' number of patients in treatment group B.

'AE B' number of patients in treatment group B for whom the event was observed.

'Relative Risk' Relative risk defined as 'PCT B' treatment divided by 'PCT A'.

'logrelrisk' natural logarithm of 'Relative Risk'

'ase.logrelrisk' asymptotic standard error of 'Relative Risk'.

'logrelriskCI.lower, logrelriskCI.upper' confidence interval for 'logrelrisk'.

'relriskCI.lower, relriskCI.upper' back transform of the CI for the log relative risk into the relative risk scale.

Author(s)

Richard M. Heiberger, Temple University, Philadelphia <rmh@temple.edu> and Erich Neuwirth, University of Vienna, <erich.neuwirth@univie.ac.at>

References

Ohad Amit, Richard M. Heiberger, and Peter W. Lane. (2008) "Graphical Approaches to the Analysis of Safety Data from Clinical Trials". *Pharmaceutical Statistics*, 7, 1, 20–35.

<http://www3.interscience.wiley.com/journal/114129388/abstract>

See Also

[AE.dotplot](#)

betaWeighted

Simple Linear Regression Slope as Weighted Sum—Excel Workbook using R and Excel

Description

The slope coefficient in simple linear regression can be written as a weighted sum of the slopes of the lines connecting each point with the centroid. We graphically display the individual slopes and the regression line. The R graph is updated dynamically as the location of the points is changed in Excel.

Details

From the RExcel menu item (on the main menubar in Excel 2003, or on the Add-Ins menubar in Excel 2007 or 2010), click ‘RthroughExcel Workbooks’, then click the ‘beta Weighted Average’ button. All Excel workbooks in the ‘RthroughExcel Workbooks’ directory are signed with a certificate issued by “R and Excel book (Heiberger-Neuwirth)”.

$$S_x = \sum_i (x_i - \bar{x})^2 \quad \text{This is uppercase } S$$

$$\begin{aligned} \hat{\beta}_1 = b_1 &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ &= \sum_i \left(\frac{(x_i - \bar{x})}{S_x} \right) ((y_i - \bar{y})) \\ &= \sum_i \left(\frac{(x_i - \bar{x})}{S_x} \right) \left(\frac{(x_i - \bar{x})}{(x_i - \bar{x})} \right) ((y_i - \bar{y})) \\ &= \sum_i \left(\frac{(x_i - \bar{x})^2}{S_x} \right) \left(\frac{(y_i - \bar{y})}{(x_i - \bar{x})} \right) \\ &= \sum_i w_i \left(\frac{(y_i - \bar{y})}{(x_i - \bar{x})} \right) \end{aligned}$$

Closing the Workbook

Closing the workbook closes the R Graphics window.

Author(s)

Richard M. Heiberger, Temple University, Philadelphia <rmh@temple.edu> and Erich Neuwirth, University of Vienna, <erich.neuwirth@univie.ac.at>

References

Richard M. Heiberger and Erich Neuwirth (2009). *R through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer. <http://www.springer.com/978-1-4419-0051-7>

Examples

Sample Excel Workbooks to Illustrate Data Transfer from Excel to R

Description

Four workbooks containing trivial sample datasets in different formats. The datasets are designed for the primary purpose of illustrating the transfer of differently structured data from Excel to R using the RExcel 'Put R DataFrame' context menu command. There is a secondary goal of illustrating decimal alignment with the RExcel 'Prettyformat Numbers' context menu command.

Details

From the RExcel menu item (on the main menubar in Excel 2003, or on the Add-Ins menubar in Excel 2007 or 2010), click 'RthroughExcel Workbooks', then click one of the dataset buttons. All Excel workbooks in the 'RthroughExcel Workbooks' directory are signed with a certificate issued by "R and Excel book (Heiberger-Neuwirth)".

TwoColumns 'x' and 'y' columns. In R, we can use the data to illustrate the use of the scatterplot and linear regression commands.

NoHeader Same data values as 'TwoColumns' but without column headers. We need to place headers on the columns with Excel commands before sending the data to R.

Long 'y' numerical column and 'group' character column. In R, we illustrate the use of the one-way ANOVA command.

Wide Same data values as 'Long' but in the wide format, with each column containing data values for one level of the classification factor. We need to convert it to the long format with the RExcel 'Paste as Stacked' context menu command before sending the data to R.

Author(s)

Richard M. Heiberger, Temple University, Philadelphia <rmh@temple.edu> and Erich Neuwirth, University of Vienna, <erich.neuwirth@univie.ac.at>

References

Richard M. Heiberger and Erich Neuwirth (2009). *R through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer. <http://www.springer.com/978-1-4419-0051-7>

Description

Excel workbook with dynamic controls to illustrate simple linear regression, with emphasis on the concept of least squares and the concept of leverage. The workbook uses the R graphical functions in the R package **HH**. The dynamic controls use RExcel to interface the controls in Excel with the graphics in R.

Details

From the RExcel menu item (on the main menubar in Excel 2003, or on the Add-Ins menubar in Excel 2007 or 2010), click ‘RthroughExcel Workbooks’, then click the ‘Linear Regression’ button. All Excel workbooks in the ‘RthroughExcel Workbooks’ directory are signed with a certificate issued by “R and Excel book (Heiberger-Neuwirth)”.

This worksheet uses the automatic recalculation mode of Excel to update an R graph as numerical values or control tools are changed on the worksheet. The worksheet opens with a display of artificial data, the table of coefficients, and the ANOVA table from a linear regression of that data. The worksheet operates by automatically sending a dataframe `xy.aov.total` to R and calculating in model `xy.lm` the revised regression analysis on that dataframe every time the user changes a value with a slider, or changes a checkbox or button. It also opens an R graphics window showing the graph of the data along with the least squares line, the predicted values, and the residuals.

Worksheet

The worksheet opens with artificial data with x_i and y_i in columns ‘E’ and ‘F’, and a color name in column ‘A’. The y_i values in column ‘F’ are controlled by the sliders in column ‘C’. The table of regression coefficients is in region ‘L1:P4’ and the ANOVA table for the regression is in region ‘L6:Q10’. The predicted values \hat{y}_i are in column ‘G’ and the residuals $e_i = (y_i - \hat{y}_i)$ are in column ‘H’.

The arithmetic for calculation of the regression coefficients is displayed in region ‘E1:I12’. The residuals e_i in column ‘H’ are squared to e_i^2 and displayed in column ‘I’. Their sum $\sum e_i^2$, is displayed in cell ‘I12’. This is the same number as is displayed in the ANOVA table as the “Sum of Squares for Residuals” in cell ‘N9’.

The term least squares means that the regression coefficients $\hat{\beta}_0$ in cell ‘M3’ and $\hat{\beta}_1$ in cell ‘M4’ are the values that minimize the sum of squared differences between the observed and predicted y values. That is,

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

is at its minimum value when $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$. The differences, labeled *residuals*, are in column ‘H’, and the squared differences are in column ‘I’. The sum of squared differences is in cell ‘I12’ and in the ANOVA table in cell ‘N9’.

Plot of Data, Lines, and Residuals

Plot of artificial data in the spreadsheet. Each observed point (x_i, y_i) from columns ‘E’ and ‘F’ is plotted in the color specified in column ‘A’. The least squares line for this data is in black. Each predicted value \hat{y}_i is marked with a small black dot on the least squares line. Residuals are indicated with the vertical lines $e_i = (y_i - \hat{y}_i)$ at each value of x_i .

Additional features on the graph are

- Subtitle: *adjust the y values with the sliders*. Reminder that this graph is directly connected to the worksheet.
- Bottom rug (black). The lengths at the tick marks are proportional to the squared residuals and their sum (cells ‘I2:I11, I12’).
- Residual Sum of Squares. The numerical value of the sum of squared residuals (cell ‘I12’) is displayed.
- Gray box. The area is proportional (with a different factor) to the sum of squared residuals (cell ‘I12’).
- Top rug: leverage. The lengths are proportional (yet another proportionality factor) to the ‘hat(x)’ values in cells ‘J2:J11’.

We can see that the least squares line minimizes the sum of squared residuals by looking at the individual squares in the sum. Click ‘square’ in cell ‘L19’ to display the squares of each residual. The squares are visual squares, the number of inches used on the page or screen for the horizontal side is the same as the number of inches used by the vertical side $e_i = (y_i - \hat{y}_i)$.

You may construct a similar plot for your own data using the Rcmdr menu item ‘Graphs > Squared Residuals’ menu and dialog box.

Alternate Line

We click ‘use alternate’ in cell ‘C16’ to change the base line for the residuals to the distances from the arbitrary line specified by the coefficients in cells ‘C18:C19’. The alternate line goes through the alternate points ‘y.alt’ in cells ‘G15:G24’. The alternate residuals in cells ‘H15:H24’ are squared in ‘I15:I24’. The sum of squares of the alternate residuals are shown in cell ‘I25’. The alternate sum of squares in cell ‘I25’ is always greater than or equal to the Residual Sum of Squares in cell ‘I12’. The least squares line, based on the least squares coefficients in cells ‘M3:M4’ is still visible as a dashed gray line. Comparison of the two lines is very nice on a live screen where it is possible to toggle between them.

We can make a direct graphical comparison of the squares associated with the two lines. Double-click the resid² value of the point at $x = 7$ in both the least squares and the alternate displays (cells ‘I8’ and ‘I21’). The cell values are now colored the associated color in cell ‘A8’. The squared residuals from both lines are also now colored the associated color.

Hat Diagonals and Leverage

We can adjust the sliders and see the least squares line shift a lot for extreme x_i and not very much for intermediate x_i . The amount of shift in \hat{y}_i for a unit shift in y_i is called leverage and is given by the *hat* value h_i in cells ‘J2:J11’.

Additional Controls

Click ‘reset Alt to LS’ in cell ‘C17’ to set the alternate coefficients to match the least squares coefficients.

Check ‘Graph on Top’ to keep the graph window always visible. Uncheck it to allow other windows to be on top.

The color names in cells 'A2:A11' can be changed. Any color name that R knows about (type `colors()` to see the list of about 700 color names) can be typed into one of the cells.

The 'Residual Display' section is in cells 'A15:A20'. The option buttons change the form of display of residuals. Double clicking the *resid*² cells 'I2:I11' and 'I15:I24' highlights the squared residual for the associated cell in the clicked cell and on the graph. Clicking the 'Color Restore' button in cell 'A20' restores the colors in the squared residual cells to black and removes the coloring from the squares in the graph.

The 'Reset' section in cells 'A22:A23' is used to restore the workbook to one of three states. Click cell 'A23' for a dropdown list of the three states: 'even spacing' (the opening state), 'uneven x spacing', 'negative slope'. Double click one of them to restore that state.

Closing the Workbook

Closing the workbook closes the R Graphics window.

Author(s)

Richard M. Heiberger, Temple University, Philadelphia <rmh@temple.edu> and Erich Neuwirth, University of Vienna, <erich.neuwirth@univie.ac.at>

References

Richard M. Heiberger and Erich Neuwirth (2009). *R through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer. <http://www.springer.com/978-1-4419-0051-7>

Richard M. Heiberger (2008). HH: Statistical Analysis and Data Display: Heiberger and Holland. R package version 2.1-15.

normal.and.t

Normal and t Excel Workbook using R and Excel

Description

Excel workbook with dynamic controls to illustrate hypothesis testing and confidence intervals for the normal and t distributions. The workbook uses the R graphical functions in the R package **HH**. The dynamic controls use RExcel to interface the controls in Excel with the graphics in R.

Details

From the RExcel menu item (on the main menubar in Excel 2003, or on the Add-Ins menubar in Excel 2007 or 2010), click 'RthroughExcel Workbooks', then click the 'Normal and t' button. All Excel workbooks in the 'RthroughExcel Workbooks' directory are signed with a certificate issued by "R and Excel book (Heiberger-Neuwirth)".

This worksheet uses the automatic recalculation mode of Excel to update an R graph as numerical values or control tools are changed on the worksheet.

When 'prob or hypoth' in cell 'D10' is checked, the worksheet settings display a graph in the R graphics window showing the normal or *t* distributions centered on the null and/or alternate hypothesis means. The horizontal axis is marked with the *z* or *t* values under both hypotheses and

under the observed data scale. The rejection region with probability α , the p -value region, and/or the Type II Error region with probability β are shaded.

When ‘confidence interval’ in cell ‘D11’ is checked, the worksheet settings display a graph in the R graphics window showing a normal or t confidence interval centered on the observed value \bar{x} .

Calculated values are display in the worksheet in cells ‘G1:K13’.

Arguments in cells ‘B3:B8’

There are 6 numbers that define the distribution under study.

μ_0 Mean under the null hypothesis.

μ_1 Mean under the alternate hypothesis.

z, t, \bar{x} For standard distributions ($\sigma = 1$ or $s = 1$) and blank n), the lookup value of z (if ν is blank) or t (if ν is a positive integer) is specified. For specified values of ($\sigma = 1$ or $s = 1$) and n , the observed \bar{x} is specified.

σ Standard deviation taken from problem specification.

n Number of observations.

ν Blank for normal, positive integer for t distribution.

Arguments in cells ‘A10:B13’

Check α ‘right’ and/or α ‘left’ to specify which side or sides are to be shaded. The α -level for each specified side is displayed by the sliders in cells ‘A12:B12’. If cell ‘A10’ is checked, cell ‘A13’ shows a checkbox specifying whether the left α is to be controlled by the right slider (default is yes).

Dynamic Controls

The checkboxes in cells ‘C3:C5’ turn on sliders in cells ‘D3:D5’. The sliders are used to investigate the effect of moving μ_1 away from μ_0 and the effect of changing the value of the observed mean \bar{x} .

Additional Controls

Cells ‘A15:B21’ may be used to take control of the horizontal and vertical ranges of the graph. If they are specified, they control the range. If they are not specified, the range is determined by the values of cells ‘B3:B8’.

Click ‘Reset’ in cell ‘F21’ to restore the worksheet to the opening state.

Check ‘Graph on Top’ in cell ‘E20’ to keep the graph window always visible. Uncheck it to allow other windows to be on top.

Closing the Workbook

Closing the workbook closes the R Graphics window.

Author(s)

Richard M. Heiberger, Temple University, Philadelphia <rmh@temple.edu> and Erich Neuwirth, University of Vienna, <erich.neuwirth@univie.ac.at>

References

- Richard M. Heiberger and Erich Neuwirth (2009). *R through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer. <http://www.springer.com/978-1-4419-0051-7>
- Richard M. Heiberger (2008). *HH: Statistical Analysis and Data Display*: Heiberger and Holland. R package version 2.1-15.

 StudentData

Student Data Excel Workbook using R and Excel

Description

Measurements on 1126 Austrian undergraduates over the past ten years. Data collected by Erich Neuwirth.

Details

From the RExcel menu item (on the main menubar in Excel 2003, or on the Add-Ins menubar in Excel 2007 or 2010), click 'RthroughExcel Workbooks', then click the 'Student Data' button. All Excel workbooks in the 'RthroughExcel Workbooks' directory are signed with a certificate issued by "R and Excel book (Heiberger-Neuwirth)". There are two worksheets in the workbook. The 'Studentdata_raw' worksheet contains raw data. The 'Studentdata' worksheet excludes one observation that has obviously incorrect size and weight information, and corrects other inaccuracies. In the 'Studentdata' worksheet, the grade variables are formatted as text. This way they will automatically become factors in the dataframe after transfer to R. The annoyance is that Excel, in standard configuration, marks this with a green error mark in the upper left corner of the cells. The error marks can be switched off in Excel 2002 and 2003 with "Tools '>' Options... '>' Error Checking '>' Number stored as text", in Excel 2007 with "MS Office Button '>' Excel Options '>' Formulas '>' Error Checking Rules '>' Number formatted as text or preceded by an apostrophe", and in Excel 2010 with "File '>' Options '>' Formulas '>' Error Checking Rules '>' Number formatted as text or preceded by an apostrophe".

Gender Student's Gender: m for man and w for woman.

Weight Student's weight in kg.

Size Student's height in cm.

Eyes Student's eye color.

Hair Student's hair color.

Shoesize Student's shoe size (European sizes: 1.5(length of foot in centimetres + 2 centimetres)).

Mathgrade Discrete values (1, 2, 3, 4) with 1 as the best grade.

Germangrade Discrete values (1, 2, 3, 4) with 1 as the best grade.

Englishgrade Discrete values (1, 2, 3, 4) with 1 as the best grade.

Smoker Student's smoking status. yes for smoker, no for non-smoker.

EduMother Educational levels of the student's Mother: (ordered Secondary, Upper Secondary, Degree).

SizeMother Mother's height in cm.

SmokeMother Mother's smoking status. yes for smoker, no for non-smoker.

EduFather Educational levels of the student's Father: (ordered Secondary, Upper Secondary, Degree).

SizeFather Father's height in cm.

SmokeFather Father's smoking status. yes for smoker, no for non-smoker.

ZodiacSign • AriesMarch 21–April 20

- TaurusApril 21–May 21
- GeminiMay 22–June 21
- CancerJune 22–July 22
- LeoJuly 23–August 21
- VirgoAugust 22–September 23
- LibraSeptember 24–October 23
- ScorpioOctober 24–November 22
- SagittariusNovember 23–December 22
- CapricornDecember 23–January 20
- AquariusJanuary 21–February 19
- PiscesFebruary 20–March 20

Author(s)

Richard M. Heiberger, Temple University, Philadelphia <rmh@temple.edu> and Erich Neuwirth, University of Vienna, <erich.neuwirth@univie.ac.at>

References

Richard M. Heiberger and Erich Neuwirth (2009). *R through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer. <http://www.springer.com/978-1-4419-0051-7>

var

Variance Calculation using R and Excel (2002, 2003, 2007, 2010)

Description

Excel workbook illustrating numerically accurate calculations of the variance function in finite-precision arithmetic.

Details

From the RExcel menu item (on the main menubar in Excel 2002 or 2003, or on the Add-Ins menubar in Excel 2007 or 2010), click 'RthroughExcel Workbooks', then click the 'Variance' button. All Excel workbooks in the 'RthroughExcel Workbooks' directory are signed with a certificate issued by "R and Excel book (Heiberger-Neuwirth)".

This worksheet uses both the Excel built-in ‘var’ function and the R var function to calculate the variance of the three numbers $10^i + 1$, $10^i + 2$, $10^i + 3$ for $i = 1 \dots 17$.

With standard IEEE double-precision arithmetic data storage (53 binary digits in the mantissa, see both `?.Machine` in R, and R FAQ 7.31 "Why doesn't R think these numbers are equal?"), the correct variance for $i = 1 \dots 15$ is 1. This is the value that R reports. For $i = 16$, the full-precision value of the numbers requires 54 binary digits. Since there are only 53 binary digits available, the numbers are rounded back to 53 binary digits and their decimal representation is (1000000000000000, 1000000000000002, 1000000000000004). The correct variance of these numbers is 4. This is the value R reports. For $i = 17$, the full-precision value of the numbers requires 57 binary digits. The numbers are rounded back to 53 binary digits and their decimal representation is (10000000000000000, 10000000000000000, 10000000000000000). These three 53-bit numbers are identical and their variance is 0.

Column ‘H’ in the Excel worksheet, labeled ‘R’, shows the variance values calculated by R with the RExcel function ‘RApply’.

Column ‘G’ in the Excel worksheet, labeled ‘Excel’, shows the variance values calculated by the builtin Excel ‘var’ function. The values in Column ‘G’ depend on the version of Excel. We show the results of four versions of Excel in the table.

R		1	2	3	R
1	1e+01	11	12	13	1
2	1e+02	101	102	103	1
3	1e+03	1001	1002	1003	1
4	1e+04	10001	10002	10003	1
5	1e+05	100001	100002	100003	1
6	1e+06	1000001	1000002	1000003	1
7	1e+07	10000001	10000002	10000003	1
8	1e+08	100000001	100000002	100000003	1
9	1e+09	1000000001	1000000002	1000000003	1
10	1e+10	10000000001	10000000002	10000000003	1
11	1e+11	100000000001	100000000002	100000000003	1
12	1e+12	1000000000001	1000000000002	1000000000003	1
13	1e+13	10000000000001	10000000000002	10000000000003	1
14	1e+14	100000000000001	100000000000002	100000000000003	1
15	1e+15	1000000000000001	1000000000000002	1000000000000003	1
16	1e+16	10000000000000000	10000000000000002	10000000000000004	4
17	1e+17	100000000000000000	10000000000000000	10000000000000000	0
18	1e+18	100000000000000000	10000000000000000	10000000000000000	0
19	1e+19	100000000000000000	10000000000000000	10000000000000000	0
20	1e+20	100000000000000000	10000000000000000	10000000000000000	0

Excel		1	2	3
1	1e+01	11	12	13
2	1e+02	101	102	103

3	1e+03	1001	1002	1003
4	1e+04	10001	10002	10003
5	1e+05	100001	100002	100003
6	1e+06	1000001	1000002	1000003
7	1e+07	10000001	10000002	10000003
8	1e+08	100000001	100000002	100000003
9	1e+09	1000000001	1000000002	1000000003
10	1e+10	10000000001	10000000002	10000000003
11	1e+11	100000000001	100000000002	100000000003
12	1e+12	1000000000001	1000000000002	1000000000003
13	1e+13	10000000000001	10000000000002	10000000000003
14	1e+14	100000000000001	100000000000002	100000000000003
15	1e+15	1000000000000000	1000000000000000	1000000000000000
16	1e+16	10000000000000000	10000000000000000	10000000000000000
17	1e+17	100000000000000000	100000000000000000	100000000000000000
18	1e+18	1000000000000000000	1000000000000000000	1000000000000000000
19	1e+19	10000000000000000000	10000000000000000000	10000000000000000000
20	1e+20	100000000000000000000	100000000000000000000	100000000000000000000

	Excel	Excel 2002	Excel 2003 Excel 2007	Excel 2010	\
1	1e+01	1	1	1	
2	1e+02	1	1	1	
3	1e+03	1	1	1	
4	1e+04	1	1	1	
5	1e+05	1	1	1	
6	1e+06	1	1	1	
7	1e+07	1	1	1	
8	1e+08	0	1	1	
9	1e+09	0	1	1	
10	1e+10	0	1	1	
11	1e+11	0	1	1	
12	1e+12	0	1	1	
13	1e+13	0	1	1	
14	1e+14	0	1	1	
15	1e+15	0	1	1	
16	1e+16	36028797018964000	36028797018964000	4	
17	1e+17	0	0	0	
18	1e+18	0	0	0	
19	1e+19	18889465931478600000000	18889465931478600000000	0	
20	1e+20	0	0	0	

There are three things to notice about the Excel representation of the numbers.

1. Excel does not display the numbers in lines 15 and 16 correctly (it has zeroed out the last digit), even though it stores them correctly.

2. Excel 2002 does not calculate the variance correctly on lines 8 through 15. The error is consistent with Excel 2002 using the numerically inaccurate difference between the sum of the squared values and the squared mean value.

Excel 2003 and 2007 calculate lines 8 through 15 correctly. Here, Excel 2003 and 2007 are using the numerically accurate algorithm of taking the difference of each individual value and the mean, and then summing those squared differences.

Excel 2010 calculate all 20 lines displayed above correctly.

3. Excel on line 16 is doing something amazing. We think it is mishandling the rounding of the numbers to 53 binary digits. Excel 2002, 2003, and 2007 do something similar on lines 19 22 23 31 42 44 45 46 48 50 52, after which we stopped looking. We show the problem in lines 16 and 19 in the table above. Excel 2003 and 2007 also has a problem on lines 36 and 37. Excel 2010, both 32-bit and 64-bit, displays the problem beginning on line 36.

References

Heiberger, Richard~M. and Holland, Burt (2004). *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R, and SAS*. Springer Texts in Statistics. Springer. ISBN 0-387-40270-5. See Section E.5.3.

Richard M. Heiberger and Erich Neuwirth (2009). *R through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer. <http://www.springer.com/978-1-4419-0051-7>

Index

- *Topic **arith**
 - var, [12](#)
- *Topic **datasets**
 - betaWeighted, [5](#)
 - Examples, [6](#)
 - StudentData, [11](#)
- *Topic **distribution**
 - normal.and.t, [9](#)
- *Topic **dynamic**
 - AEdotplot, [3](#)
 - linreg, [7](#)
 - normal.and.t, [9](#)
 - RthroughExcelWorkbooksInstaller-package,
 - [2](#)
- *Topic **hplot**
 - AEdotplot, [3](#)
- *Topic **htest**
 - AEdotplot, [3](#)
 - normal.and.t, [9](#)
- *Topic **interface**
 - AEdotplot, [3](#)
 - linreg, [7](#)
 - normal.and.t, [9](#)
 - RthroughExcelWorkbooksInstaller-package,
 - [2](#)
 - var, [12](#)
- *Topic **regression**
 - linreg, [7](#)
- AE.dotplot, [3, 4](#)
- AE.dotplot (AEdotplot), [3](#)
- ae.dotplot (AEdotplot), [3](#)
- AEdotplot, [3, 3](#)
- aedotplot (AEdotplot), [3](#)
- betaWeighted, [3, 5](#)
- BookFilesTOC
 - (RthroughExcelWorkbooksInstaller-package),
 - [2](#)
- dynamic graphics
 - (RthroughExcelWorkbooksInstaller-package),
 - [2](#)
- Examples, [3, 6](#)
- installRthroughExcel
 - (RthroughExcelWorkbooksInstaller-package),
 - [2](#)
- installRthroughExcelWorkbooks
 - (RthroughExcelWorkbooksInstaller-package),
 - [2](#)
- linreg, [2, 7](#)
- Long (Examples), [6](#)
- NoHeader (Examples), [6](#)
- norm.curve, [2](#)
- normal.and.t, [3, 9](#)
- regr1.plot, [2](#)
- RExcelInstaller-package, [3](#)
- RthroughExcelWorkbooksInstaller
 - (RthroughExcelWorkbooksInstaller-package),
 - [2](#)
- RthroughExcelWorkbooksInstaller-package,
 - [2](#)
- StudentData, [2, 11](#)
- TwoColumns (Examples), [6](#)
- var, [3, 12](#)
- Wide (Examples), [6](#)
- xls
 - (RthroughExcelWorkbooksInstaller-package),
 - [2](#)