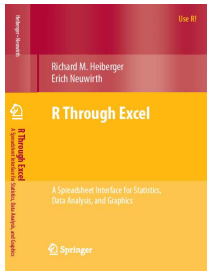
Dynamic Interaction of R Graphics and Excel

Richard M. Heiberger

Abstract

R provides powerful graphic tools. R also has a high startup cost for non-technical users. Excel is already on almost everyone's desk, provides a familiar interface, and has many control mechanisms (sliders, checkboxes, option buttons, double-clicking) with which users are comfortable. It is relatively easy to place complex R graphs into the the Excel automatic recalculation model, so the graphs are automatically updated when the data or the control mechanisms are changed on the spreadsheet. In this paper we present and discuss the behind-the-scenes details of several R graphical displays that are accessed and controlled through simple and familiar widgets.



Dynamic displays can be designed for different audience assumptions.

The normal and t plot in Section 1, designed for the introductory course, shows a graph of significance and power for the normal and t-tests. We adjust sliders to illustrate how the power changes as the sample mean $\bar{x}$ changes and as the location of the alternative value of the population mean μ_1 changes.

The linear regression plot in Section 2 shows what the term "least squares" means by drawing the squares associated with the least squares fit and comparing them to squares for a different model.

The Adverse Events Dotplot in Section 3, designed for the monitoring of safety data collected during clinical trials, shows the relative risk of various adverse events. We click the data array in Excel to change the display characteristics of the plot in R, for example, to emphasize the risk or the actual frequency of occurence of the types of events.

The simulated experiment example in Section 4 reverses the direction of control. This example uses clicks on an R graph to control the Excel display.

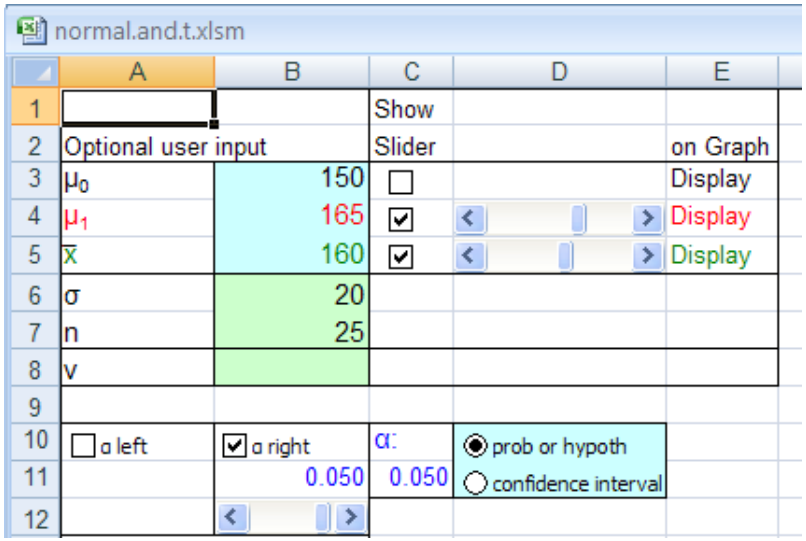
We illustrate and discuss the technical capabilities of the interface, the characteristics of the intended audience for these displays, and design decisions we made based on these considerations.

1 Normal and t

A typical homework exercise is as follows:

> We have an experiment from a normally distributed population with
>
> $$\begin{aligned} H_0&: \quad \mu = \mu_0 = 150 \\ H_1&: \quad \mu > 150 \end{aligned}$$
>
> We know $\sigma = 20$. We have observed $\bar{x}_{\text{obs}} = 160$ as the mean of $n = 25$ observations. Test at $\alpha = 0.05$. Determine the critical value. Under the alternate assumption that the population mean $\mu_1 = 165$, what is the probability of the Type II error and what is the power of the test? The answer is displayed in Fig. 1.

We enter the six numbers in the problem statement into the Normal and t worksheet and and immediately see the null and alternative distributions; the α, β, and p values; and all the relevant axes. In an introductory class we build up to this display one number at a time.



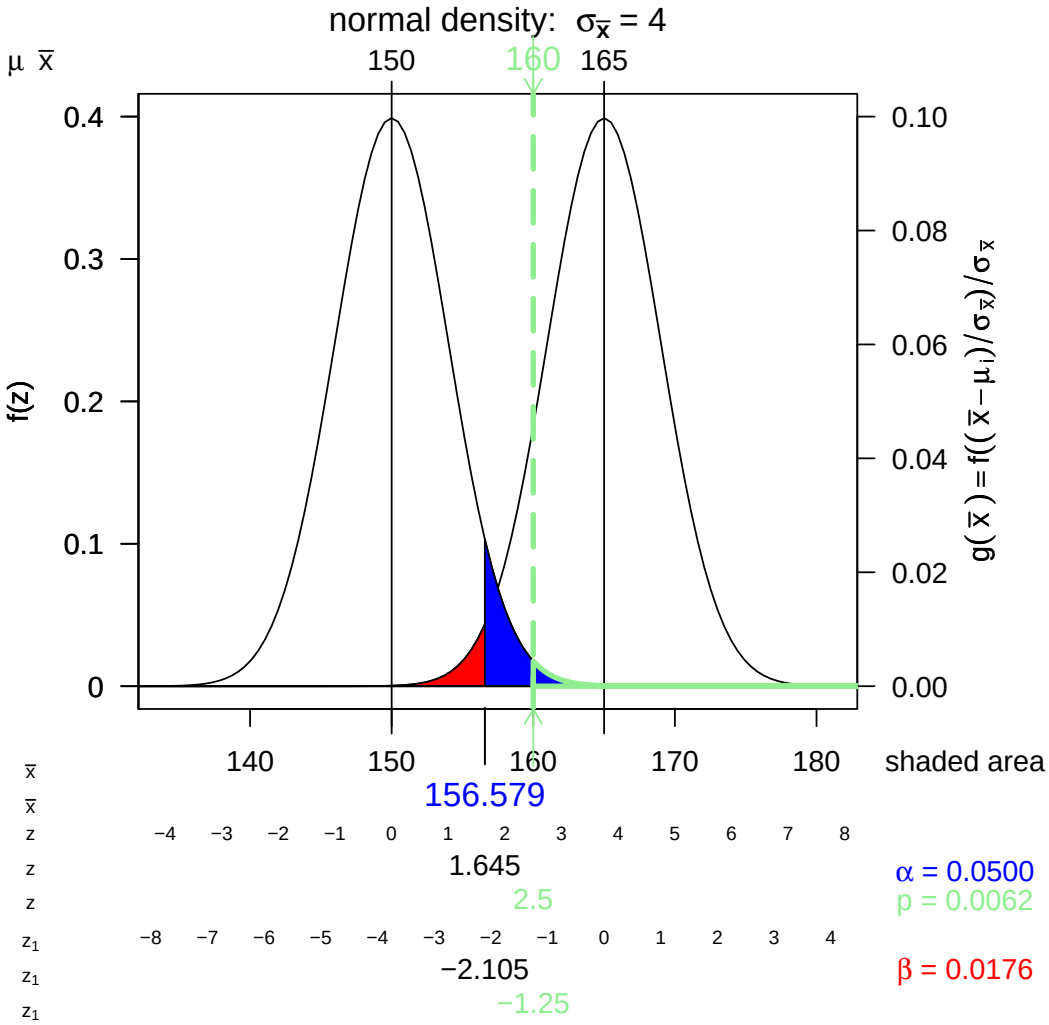


Figure 1: Evaluate the power at the alternative hypothesis mean $\mu_1 = 165$. We checked the checkbox in cell C4 to display the alternative distribution on the graph. When the checkbox is checked, the scroll bar can be used to dynamically adjust the value of μ_1. In this figure, we set the alternative mean to $\mu_1 = 165$ and see that $\beta(\mu_1 = 165) = 0.0176$ and power$(\mu_1 = 165) = (1 - 0.0176) = 0.9824$.

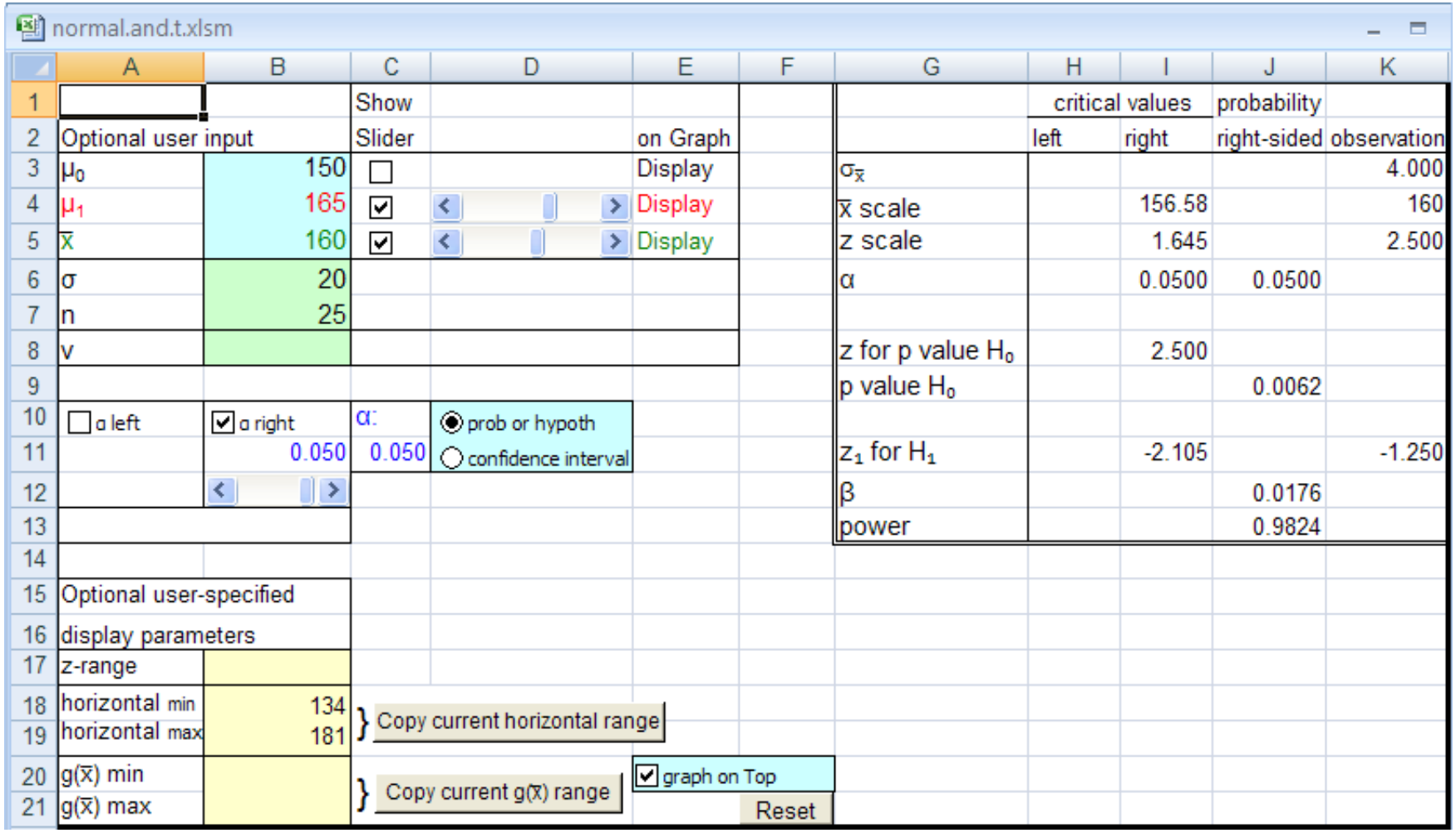


Figure 2: This is the full Excel display of the input values and controls along with the numerical output values. When any input value or slider is changed by the user, the Excel automatic recalculation sends a revised R command to R. The return values of the R command in turn trigger the automatic recalculation to revise the values displayed in the output area in cells G1:K13.

1.1 Mechanics of the Interaction

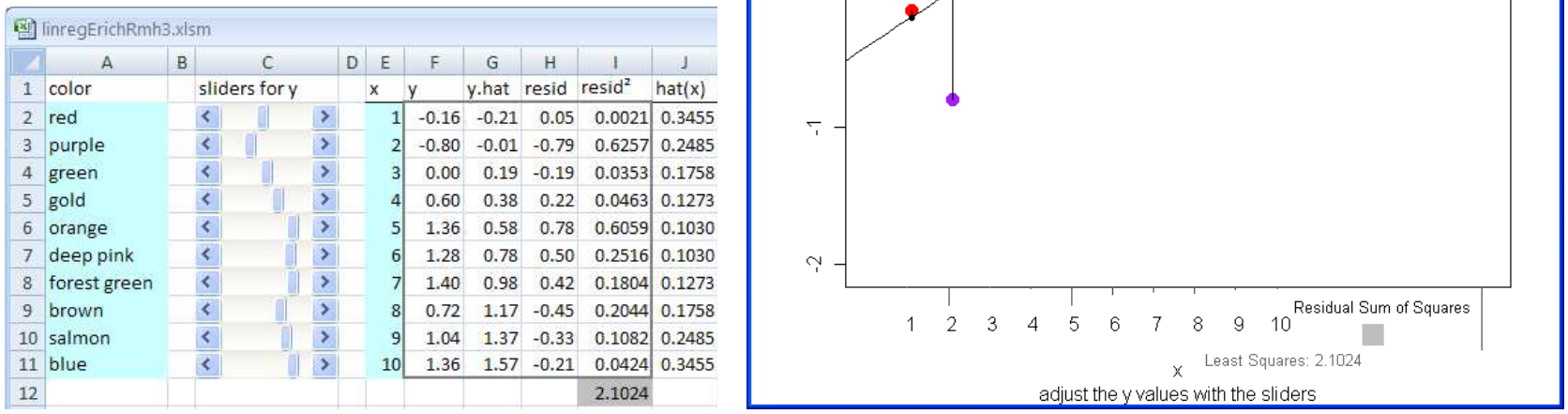
The normal.and.t workbook gives a user in Excel control over a complex graph constructed in R. It does so by placing the R functions inside the standard Excel automatic recalculation model. When a user changes a cell in the Excel workbook, a call to R is automatically generated using the revised data values.

Cells A1:K21 in Fig. 2 are designed for user input and output. This worksheet contains several shaded data entry fields and several standard Excel checkboxes and sliders for user control. It contains a region in cells G1:K13 for numerical output. It produces a graph in the R Graphics window.

The communication between R and Excel is done in the offscreen sections of the workbook, using RExcel's RApply function and several related functions. When the workbook detects that the user has changed a cell, it automatically updates all cells that depend on the value of the changed cell. When the cell containing the call to R detects that one of its data entry cells has been changed, it automatically issues a new call to the normal.and.t.dist.wrapper function in R with the revised argument values. The normal.and.t.dist.wrapper function calls the normal.curve function in the HH package. The return values from the function call are automatically displayed in the user output area in cells G1:K13.

2 Least Squares Regression

We use Excel control mechanisms for dynamic control of the R graph with the goal of explaining the terms "least squares" and "leverage".



linregErichRmh3.xlsm

	A	B	C	D	E	F	G	H	I	J
1	color		sliders for y		x	y	y.hat	resid	resid²	hat(x)
2	red				1	-0.16	-0.21	0.05	0.0021	0.3455
3	purple				2	-0.80	-0.01	-0.79	0.6257	0.2485
4	green				3	0.00	0.19	-0.19	0.0353	0.1758
5	gold				4	0.60	0.38	0.22	0.0463	0.1273
6	orange				5	1.36	0.58	0.78	0.6059	0.1030
7	deep pink				6	1.28	0.78	0.50	0.2516	0.1030
8	forest green				7	1.40	0.98	0.42	0.1804	0.1273
9	brown				8	0.72	1.17	-0.45	0.2044	0.1758
10	salmon				9	1.04	1.37	-0.33	0.1082	0.2485
11	blue				10	1.36	1.57	-0.21	0.0424	0.3455
12									2.1024	

Figure 3: Plot of artificial data in the spreadsheet in the left panel. 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.

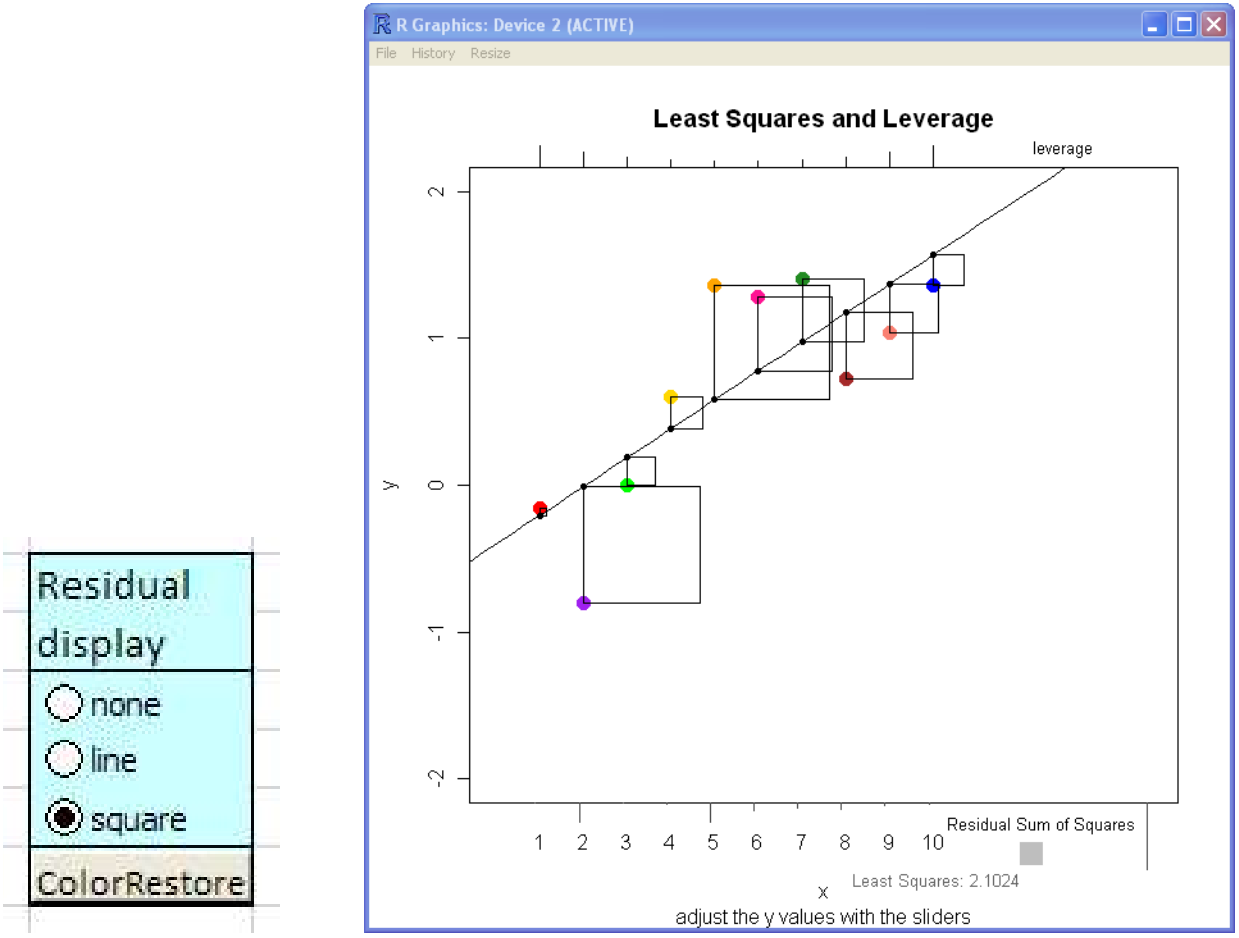


Figure 4: We click the square option button in the left panel to produce this figure, a standard regression line with the residuals indicated by squares, each of whose side is the length of the residual $e_i = (y_i - \hat{y}_i)$. The squares are visual squares; the number of inches used on the page or screen for the horizontal side is the same as the number of inches used by the vertical side $e_i = (y_i - \hat{y}_i)$. The bottom rug fringes have lengths proportional to the area of the squares. The top rug fringes have lengths proportional to the leverage of the points, smallest at the mean of the x values and larger towards the extremes.

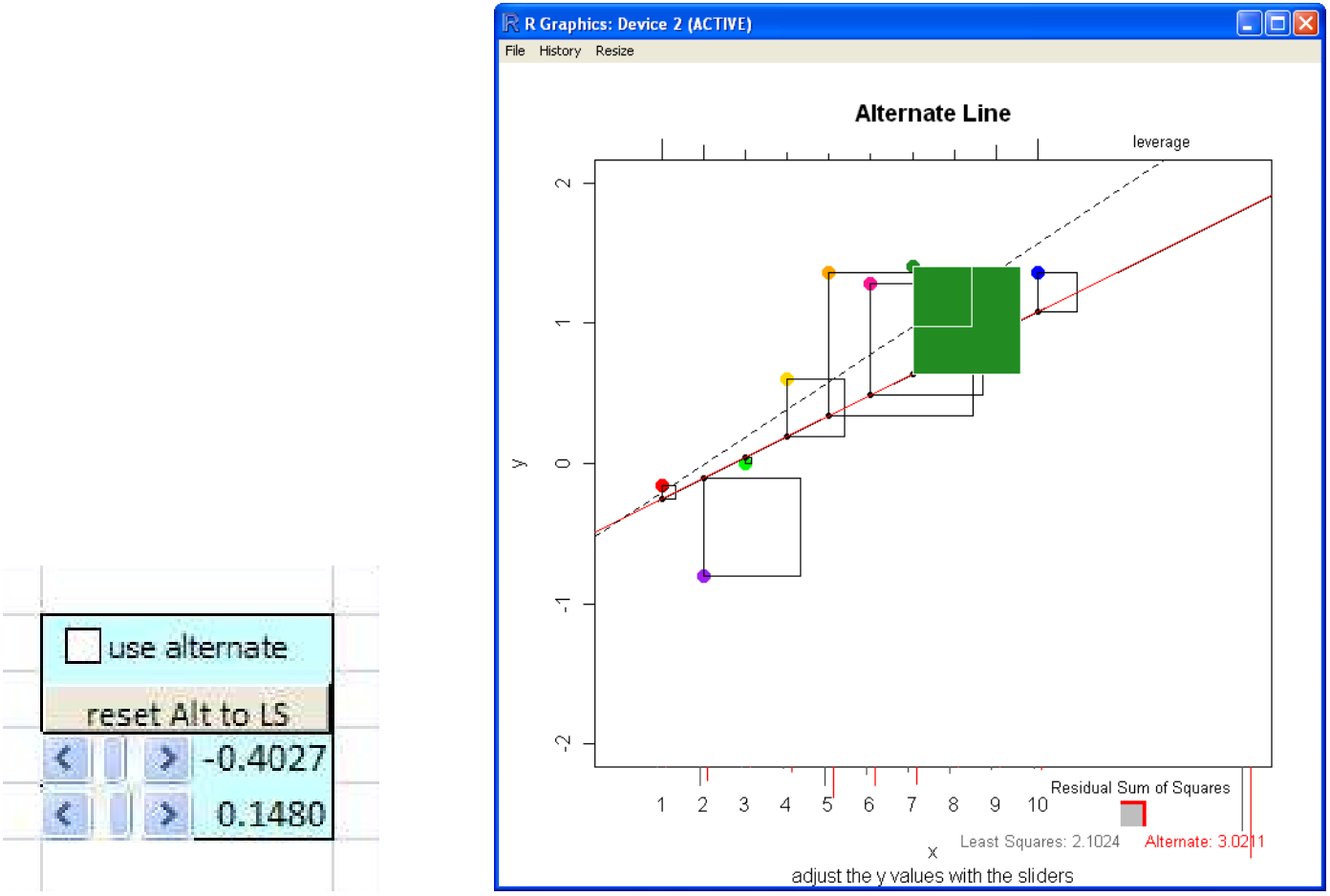
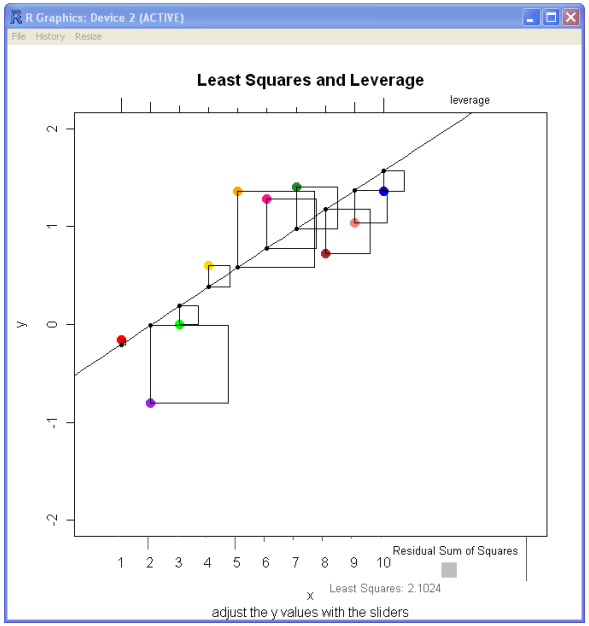


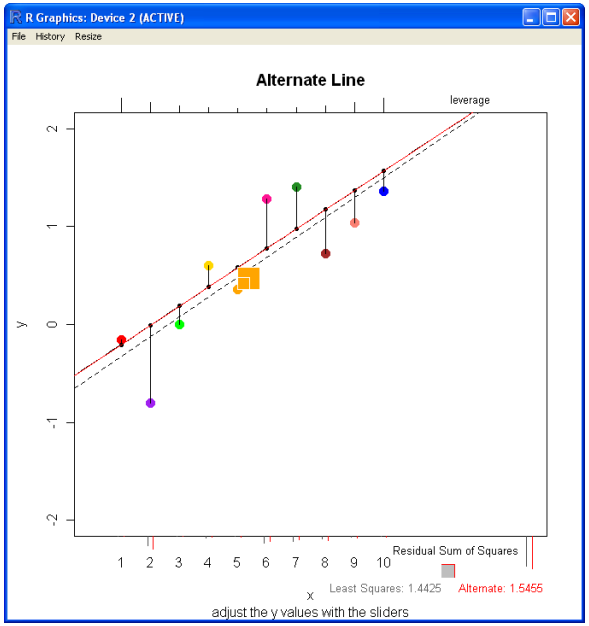
Figure 5: In the left panel, we click to form the squares from an arbitrary (solid) line instead of the (dotted) least squares line. The squared residuals from both lines are shown colored for point 7. In this example, we immediately see that the alternate squared residual is larger than the least-squares squared residual for this point at $x = 7$. The bottom red rugs are proportional to the squared alternate residuals. The alternate sum of squared residuals is shown on the graph both numerically and as a red square that is always larger than the gray square for the residual sum of squares calculated by least squares.

a. Fig. 4
$y_5 = y_{10} = 1.36$



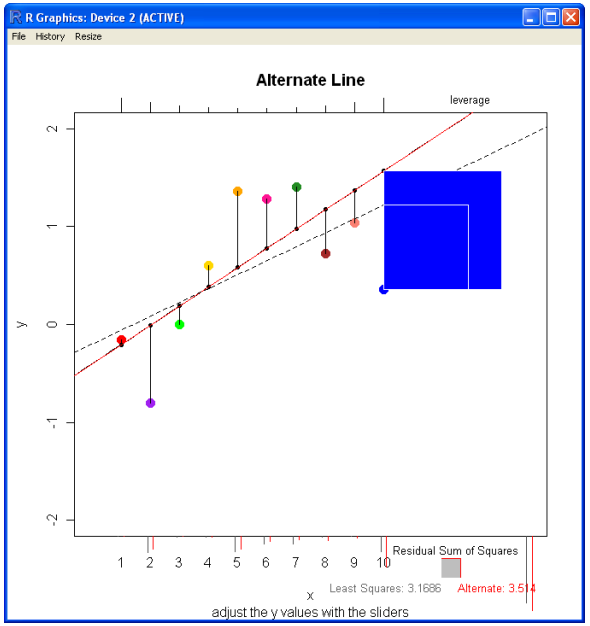
$y = -0.4027 + 0.1968x$

b.
$y_5 = 0.36,\ y_{10} = 1.36$



$y = -0.5360 + 0.2029x$

c.
$y_5 = 1.36,\ y_{10} = 0.36$



$y = -0.2027 + 0.1423x$

Figure 6: In this set of three plots, so we can easily compare the regression lines and the sizes of the squared residuals after changing y values of several points. The regression lines for the first two panels, original data and with point 5 changed, are similar. The line for the third panel, with point 10 changed, is different. In the right two panels, the original line is shown as a solid red line and the the new lines are dashed gray lines. In the second panel, the residuals of the new point from both lines are similar. Point $x = 5$ is in the center of the range of x-values. Therefore, changing its y-value does not have a large effect on the line. In the third panel, the residual of the new point from the original line is larger than from the new line. This is to be expected because the new line follows the change in the y-value of point $x = 10$, which is on the extreme of the x-values.

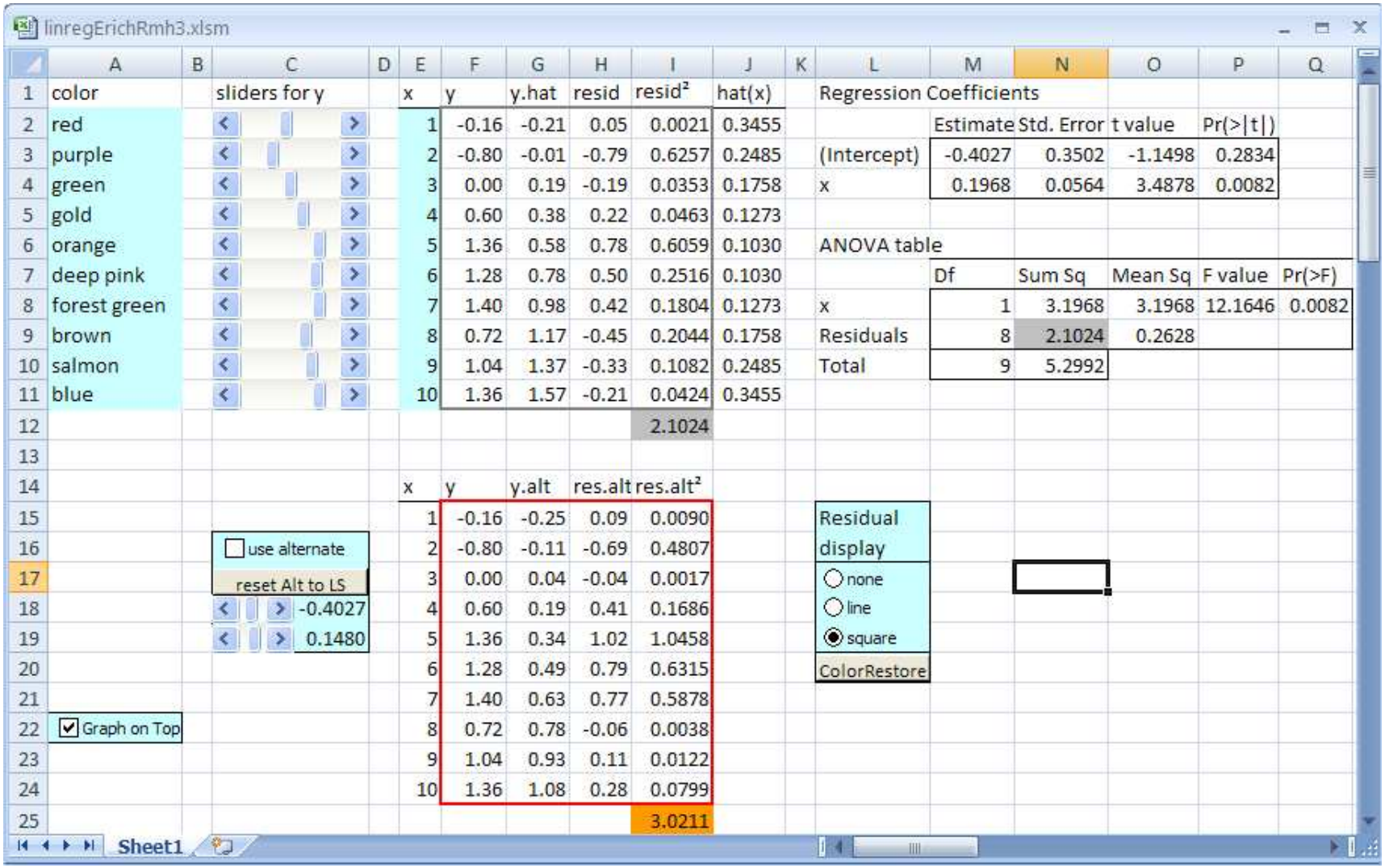


Figure 7: This is the complete worksheet. We choose the color of the points by typing any of the 657 color names known to R (`colors()`) in cells A2:A11. The sliders in cells C2:C11 control the y values. The regression coefficients and ANOVA table are shown in cells L1:Q10. The residual sum of squares in ANOVA table cell N9 is identical to the sum of the squared residuals in cell I12. The detail for the arbitrary alternate straight line is in cells E14:I25, with the sum of the squared alternate residuals in cell I25.

2.1 Mechanics of the Interaction

Anytime a number or value in the worksheet is changed, either by typing or by using the sliders or checkboxes or options, the Excel cell containing the call to R detects that its inputs have changed and automatically calls R to revise the graph.

3 Adverse Events Dotplot of incidence and relative risk

Evaluation of adverse experience data is a critical aspect of all clinical trials. Figure 9 is a two-panel display of the most frequently occurring AEs in the active arm of the study. The first panel displays their incidence by treatment group, with different symbols for each group. The second panel displays the relative risk of an event on the active arm relative to the placebo arm, with 95% confidence intervals as defined by [Agresti, 1990] for a 2×2 table.

The AEs are ordered by relative risk so that events with the largest increases in risk for the active treatment are prominent at the top of the display. We do not recommend ordering alphabetically by preferred term, which is the likely default with routine programming, because that makes it more difficult to see the crucial information of relative importance of the AEs.

AEdotplot.xlsm

	A	B	C	D	E	F	G	H
1	Double click a column name in row 5	to sort the data and plot the graph						
2	Treatment A name:	Treatment A						
3	Treatment B name:	Treatment B						
4								
5	Event	PCT A	PCT B	N A	AE A	N B	AE B	Relative Risk
6	ARTHRALGIA	0.46	3.48	216	1	431	15	7.52
7	NAUSEA	4.63	19.03	216	10	431	82	4.11
8	ANOREXIA	0.93	3.48	216	2	431	15	3.76
9	HEMATURIA	0.93	3.25	216	2	431	14	3.51
10	INSOMNIA	1.85	6.03	216	4	431	26	3.26
11	VOMITING	2.78	8.58	216	6	431	37	3.09
12	DYSPEPSIA	3.70	9.74	216	8	431	42	2.63
13	WEIGHT DECREASE	0.93	2.09	216	2	431	9	2.26
14	PAIN	1.85	3.94	216	4	431	17	2.13
15	FATIGUE	1.85	3.71	216	4	431	16	2.00
16	DIARRHEA	10.65	20.88	216	23	431	90	1.96
17	FLATULENCE	2.78	4.64	216	6	431	20	1.67
18	DIZZINESS	4.17	6.73	216	9	431	29	1.61
19	ABDOMINAL PAIN	9.26	14.15	216	20	431	61	1.53
20	RESPIRATORY DISORDER	1.85	2.55	216	4	431	11	1.38
21	HEADACHE	6.48	8.35	216	14	431	36	1.29
22	INJURY	5.56	6.96	216	12	431	30	1.25

Figure 8: Data on adverse events in an Excel spreadsheet.

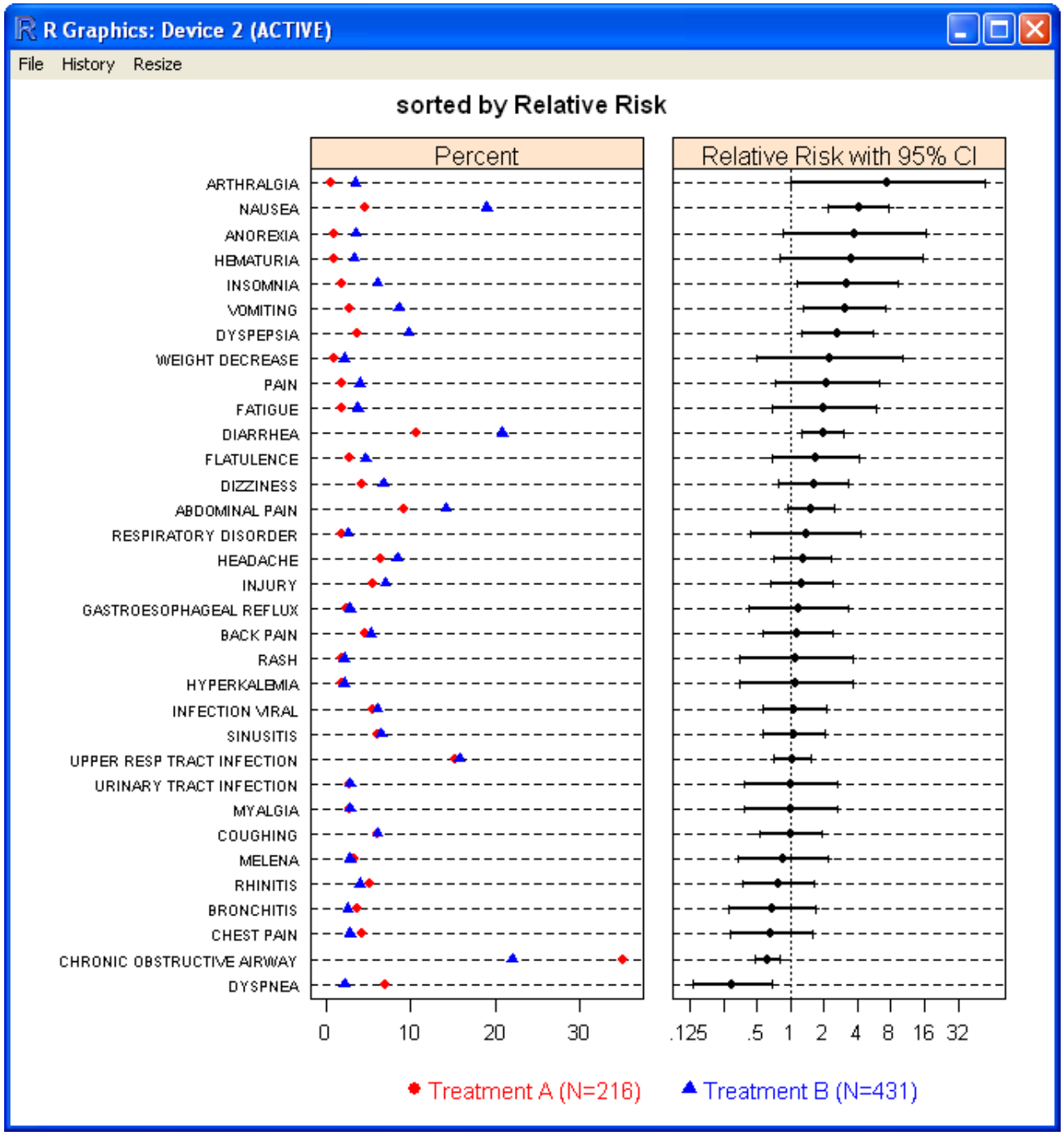


Figure 9: Most frequent on-therapy adverse events sorted by relative risk

AEdotplot.xlsm

	A	B	C	D	E	F	G	H
1	Double click a column name in row 5	to sort the data and plot the graph						
2	Treatment A name:	Treatment A						
3	Treatment B name:	Treatment B						
4								
5	Event	PCT A	PCT B	N A	AE A	N B	AE B	Relative Risk l
6	WEIGHT DECREASE	0.93	2.09	216	2	431	9	2.26
7	VOMITING	2.78	8.58	216	6	431	37	3.09
8	URINARY TRACT INFECTION	2.78	2.78	216	6	431	12	1.00
9	UPPER RESP TRACT INFECTION	15.28	15.78	216	33	431	68	1.03
10	SINUSITIS	6.02	6.50	216	13	431	28	1.08
11	RHINITIS	5.09	3.94	216	11	431	17	0.77
12	RESPIRATORY DISORDER	1.85	2.55	216	4	431	11	1.38
13	RASH	1.85	2.09	216	4	431	9	1.13
14	PAIN	1.85	3.94	216	4	431	17	2.13
15	NAUSEA	4.63	19.03	216	10	431	82	4.11
16	MYALGIA	2.78	2.78	216	6	431	12	1.00
17	MELENA	3.24	2.78	216	7	431	12	0.86
18	INSOMNIA	1.85	6.03	216	4	431	26	3.26
19	INJURY	5.56	6.96	216	12	431	30	1.25
20	INFECTION VIRAL	5.56	6.03	216	12	431	26	1.09
21	HYPERKALEMIA	1.85	2.09	216	4	431	9	1.13
22	HEMATURIA	0.93	3.25	216	2	431	14	3.51

Figure 10: Double-click the spreadsheet on a column title, in this case alphabetical by event name (silly, yes, but the data may have been given to you in that sort order).

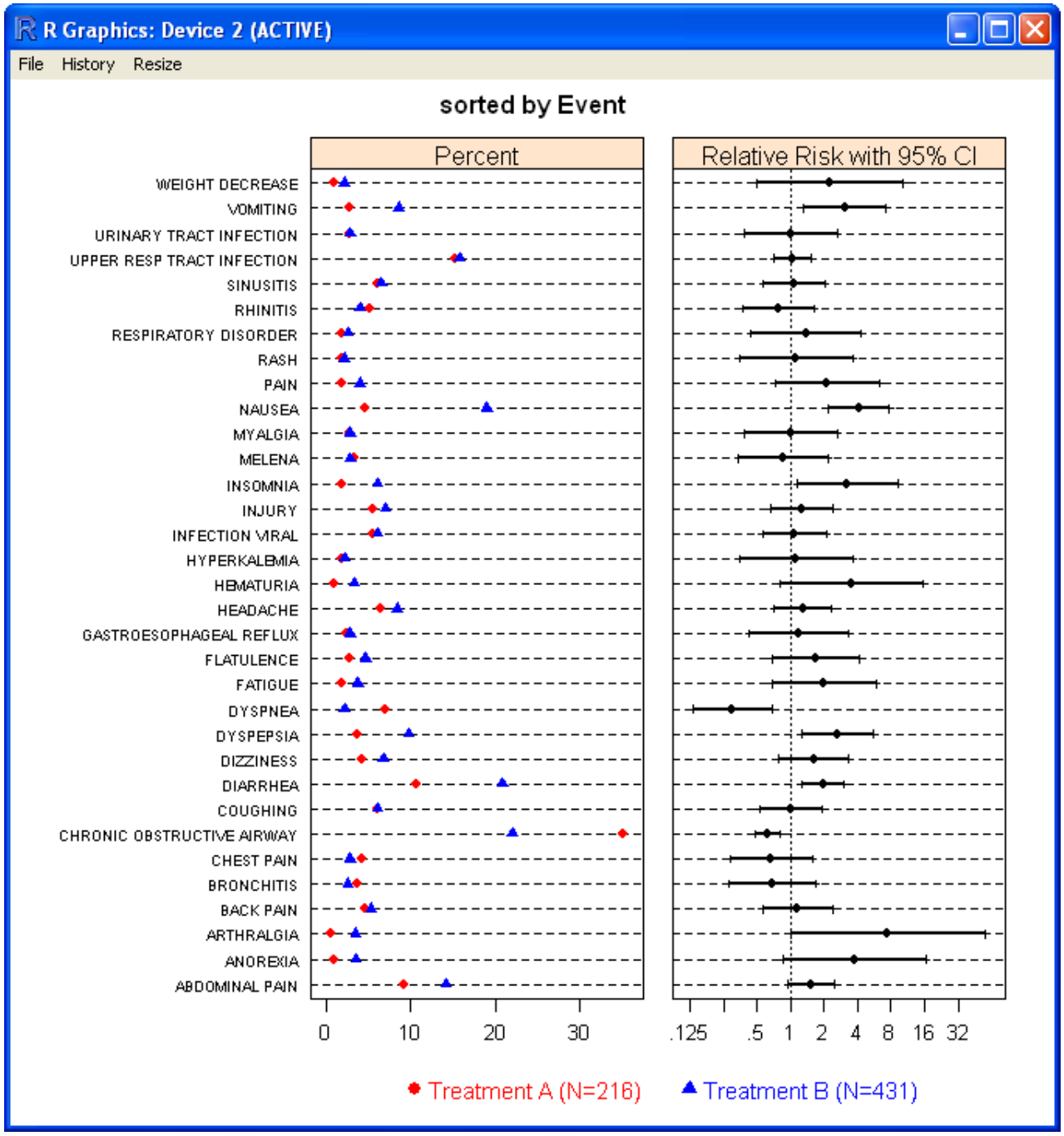


Figure 11: The graph immediately sorts itself to match.

3.1 Mechanics of the Interaction

We detect a double-click event in the column-header row of the Excel worksheet. That event triggers a macro that sends a plot command to R to redraw the graph sorted according to the values in the clicked column.

4 Control of Excel from the R Graphics Window

This example is abstracted from a simulated experiment. We have several experimental scenarios, each defined by the values of a set of parameters. We press an Excel button to tell R to calculate the response value for each scenario under two different strategies. We display in Excel a single number summary of a more detailed response and display an R graph of the set of summaries. Based on our review of the summary graph, we click on the graph to tell Excel to get the detailed information from R.

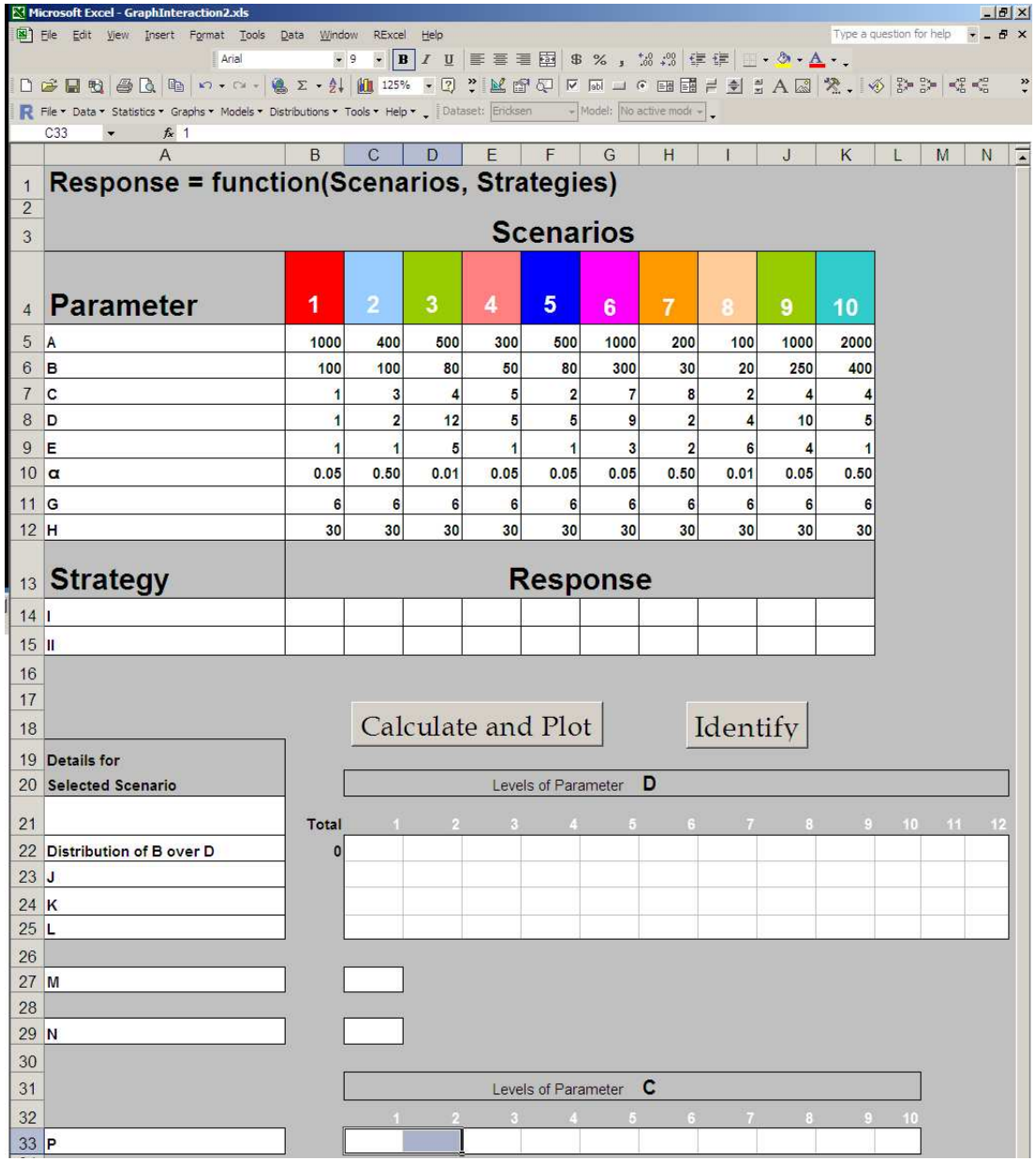


Figure 12: Initialization Parameters for 10 Scenarios.

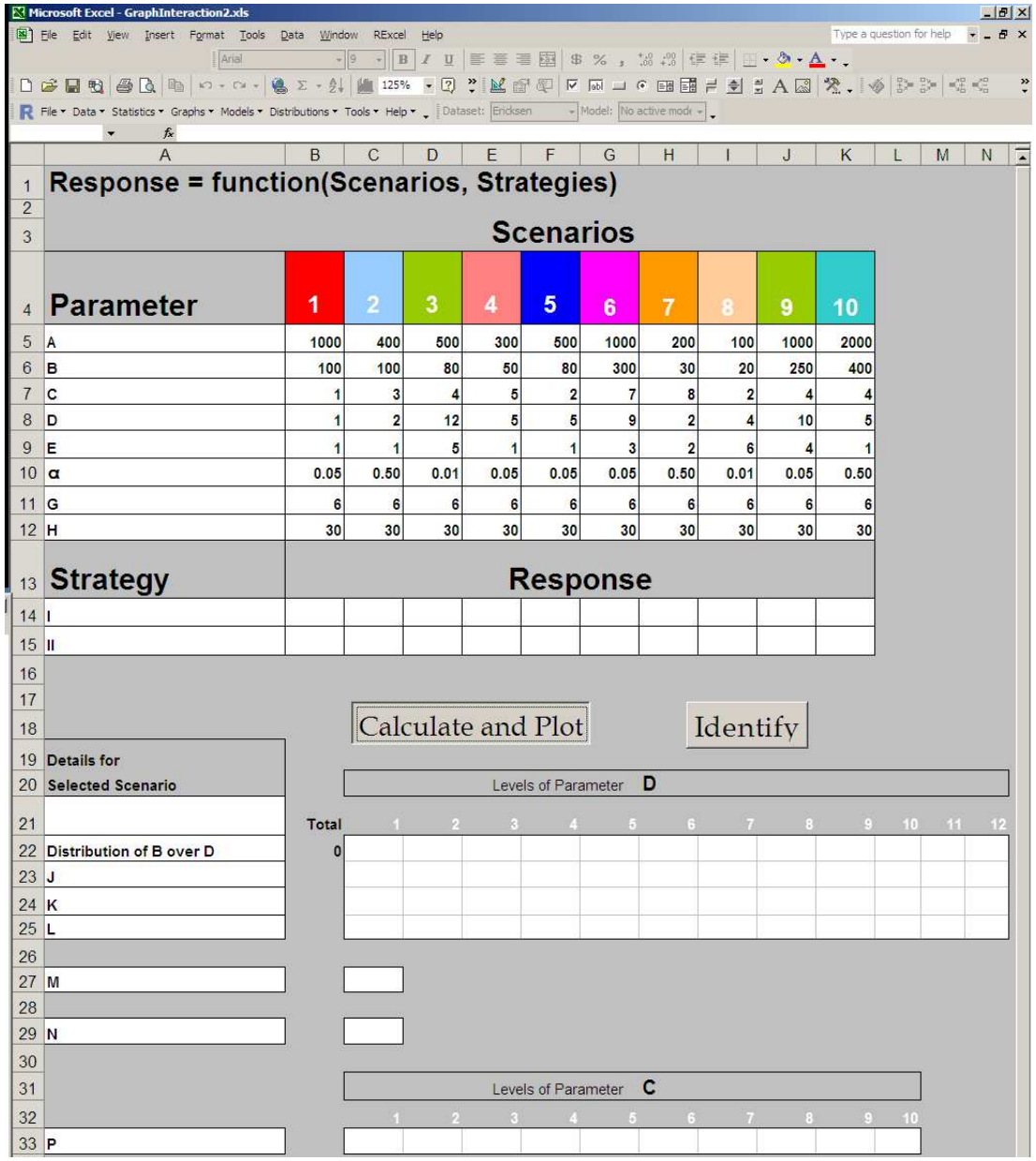


Figure 13: Click the Calculate and Plot button to get Figures 14 and 15.

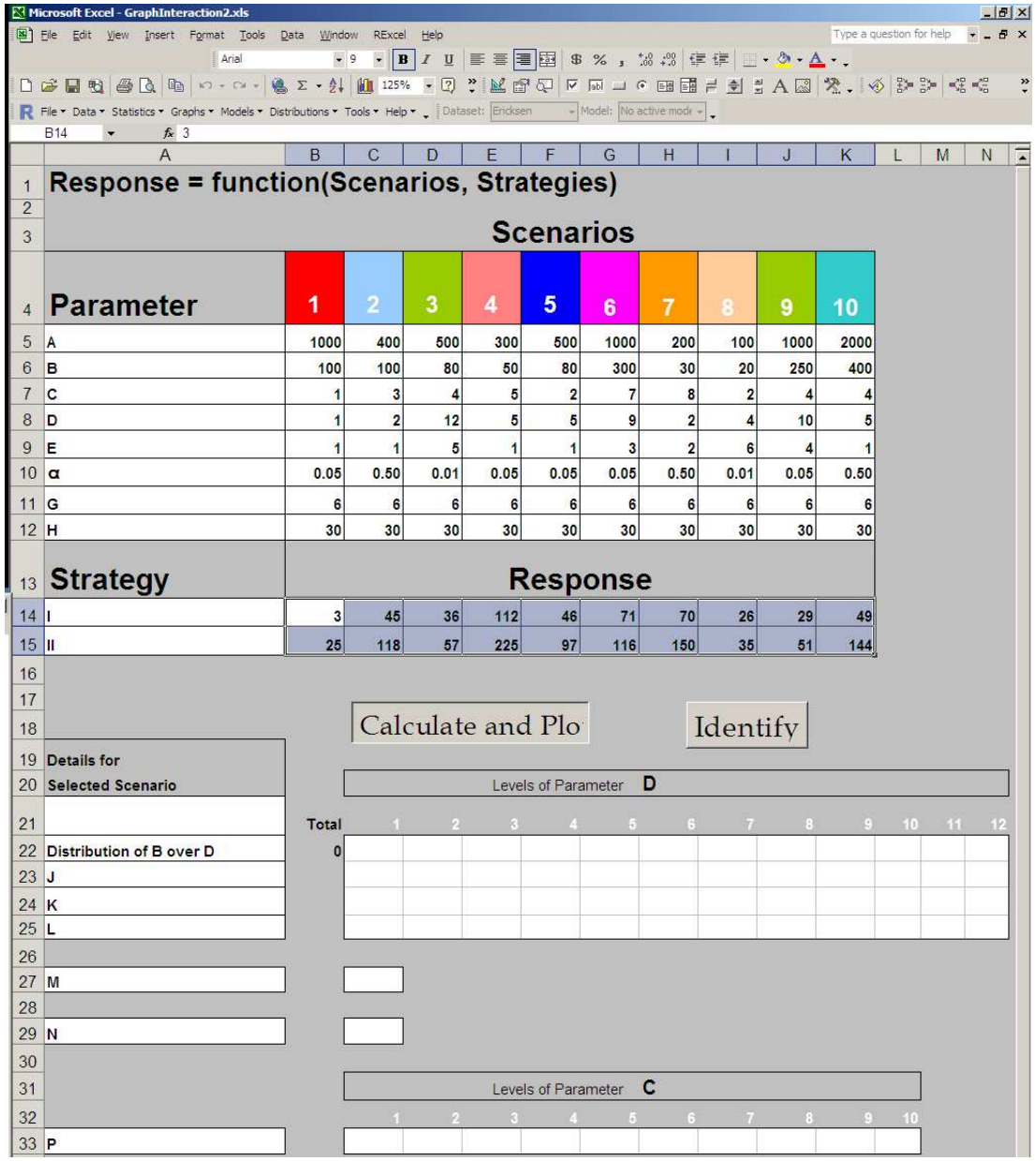


Figure 14: The Response values are calculated and displayed: Numerical Values.

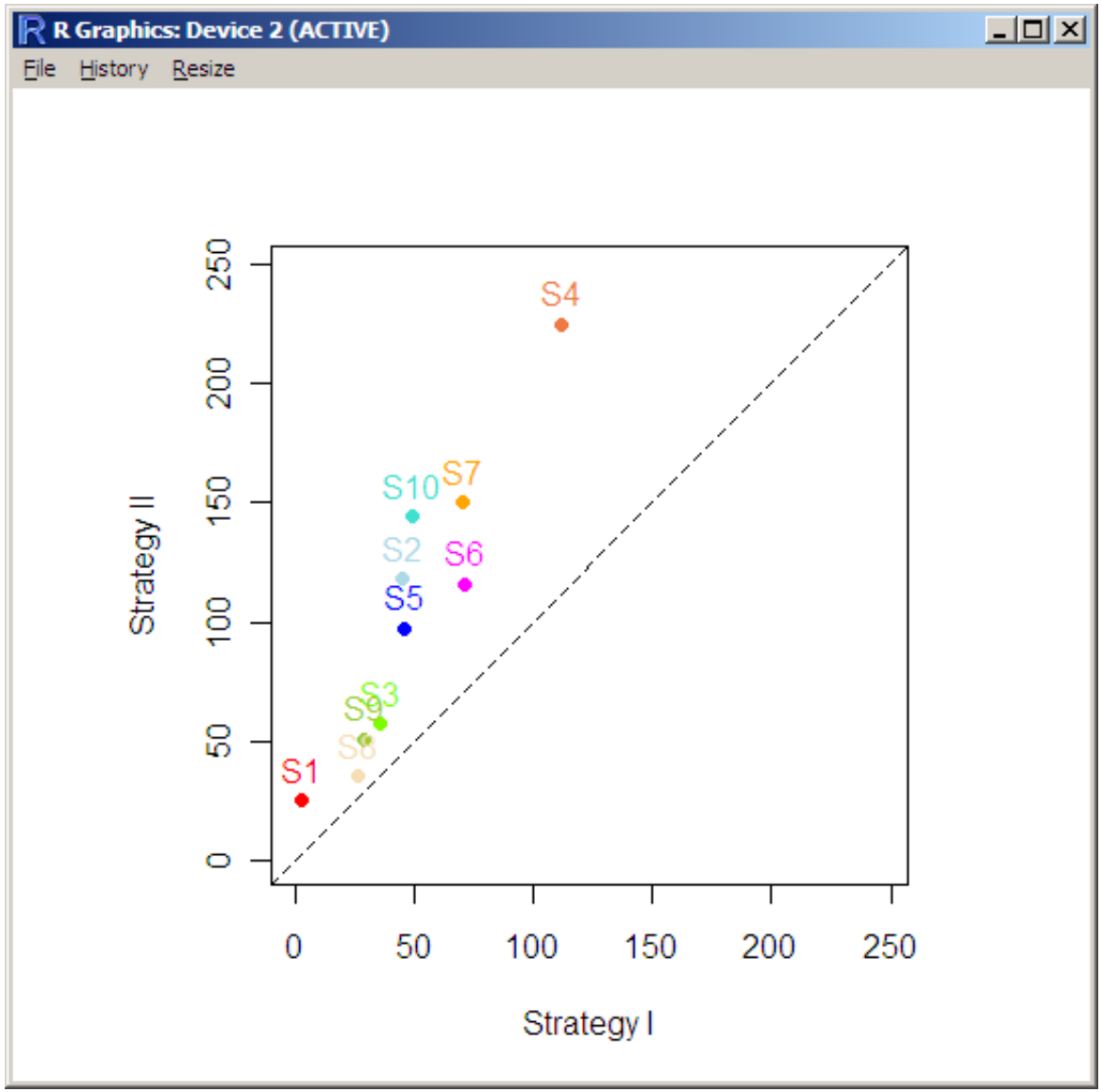


Figure 15: The Response values are calculated and displayed: Display. It is very clear from the plot that Strategy II has larger response values than Strategy I. The dots for the Scenarios are colored to match the column headers in the table in Figure 12.

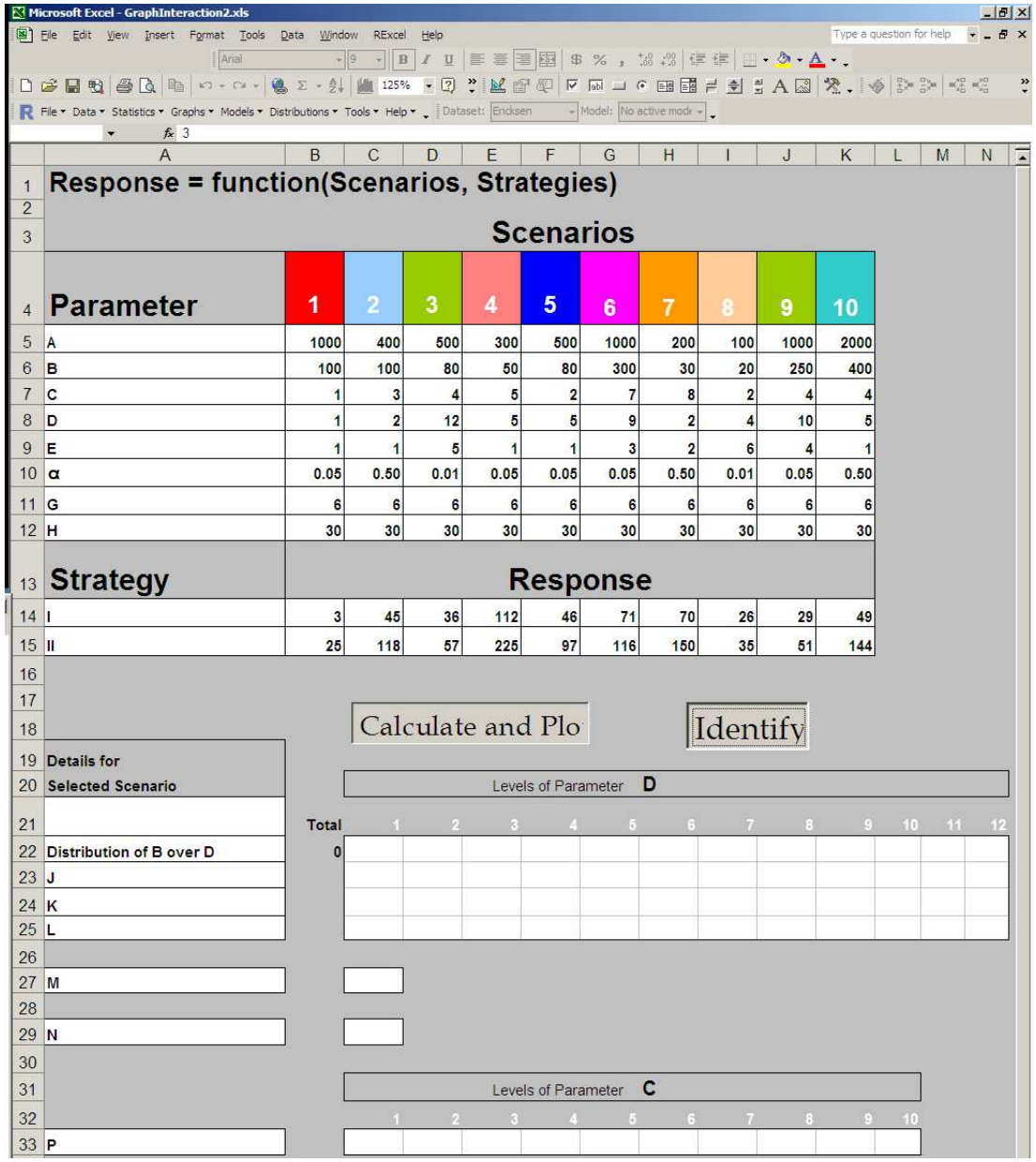


Figure 16: Let us investigate detail about Scenario 4. Click the **Identify** button to get the selection cross-hairs in Figure 17.

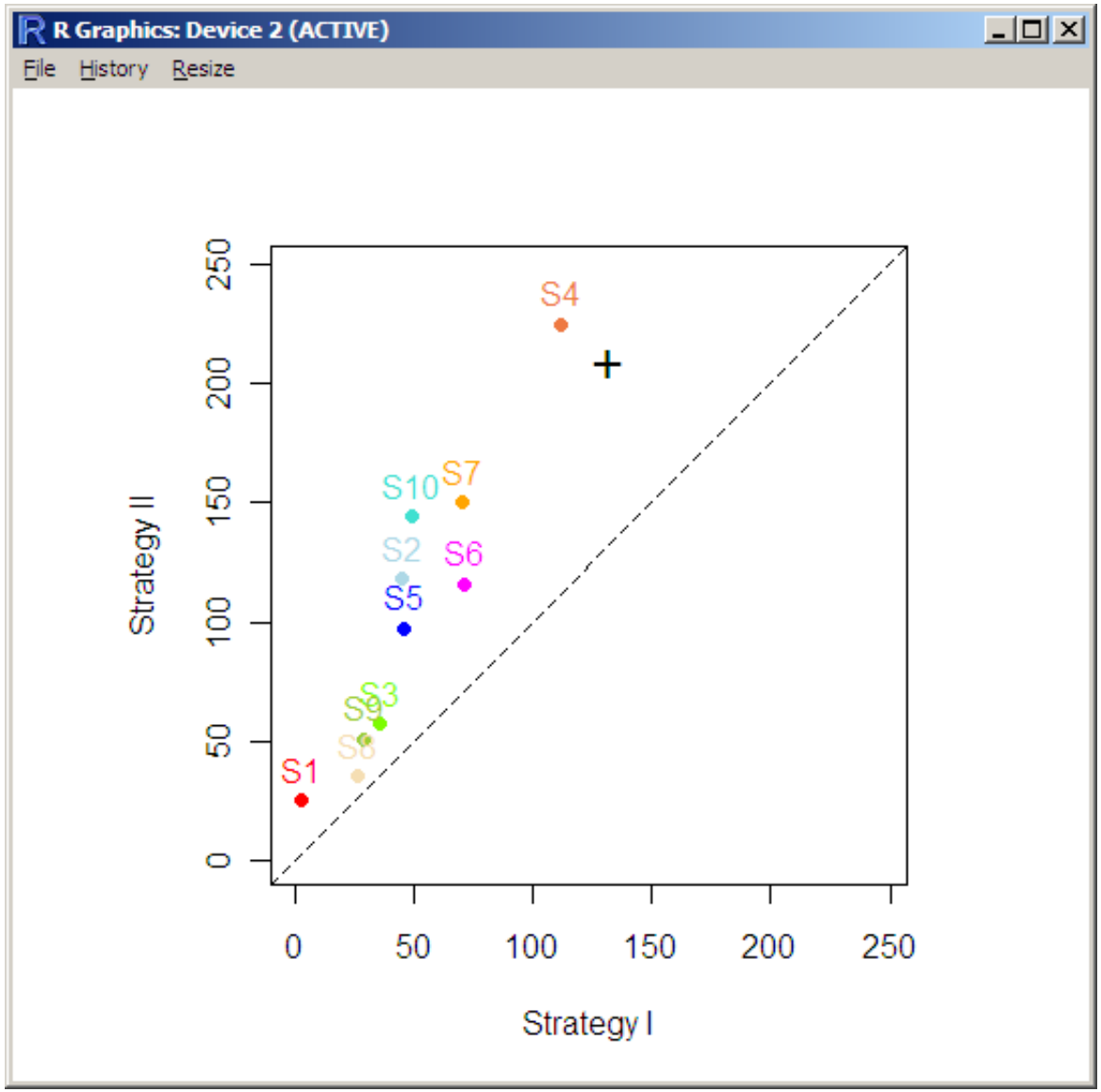


Figure 17: Click on the the point associated with Scenario 4 to get Figures 17 and 18.

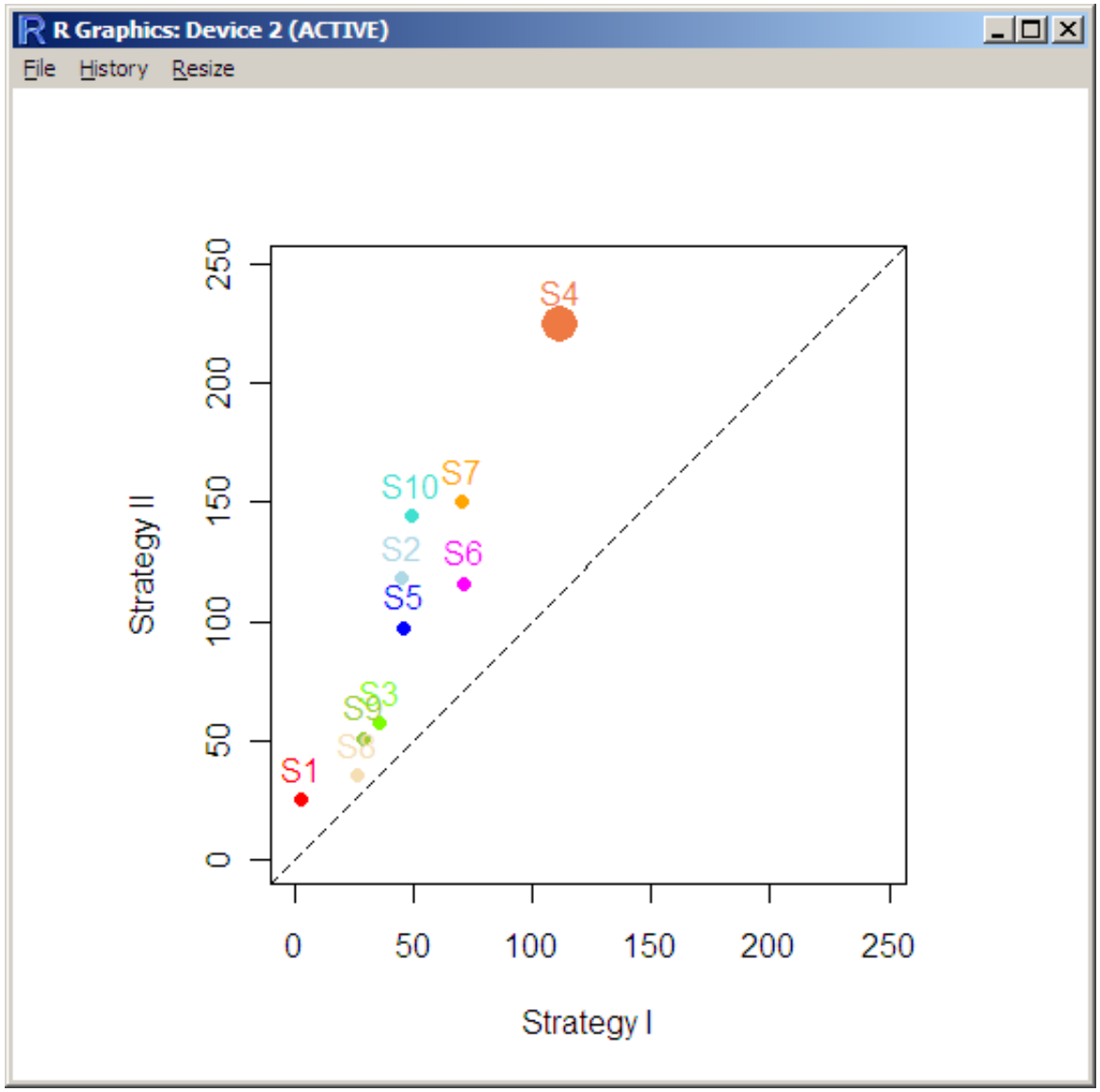


Figure 18: The dot for Scenario 4 is enlarged (in the same color).

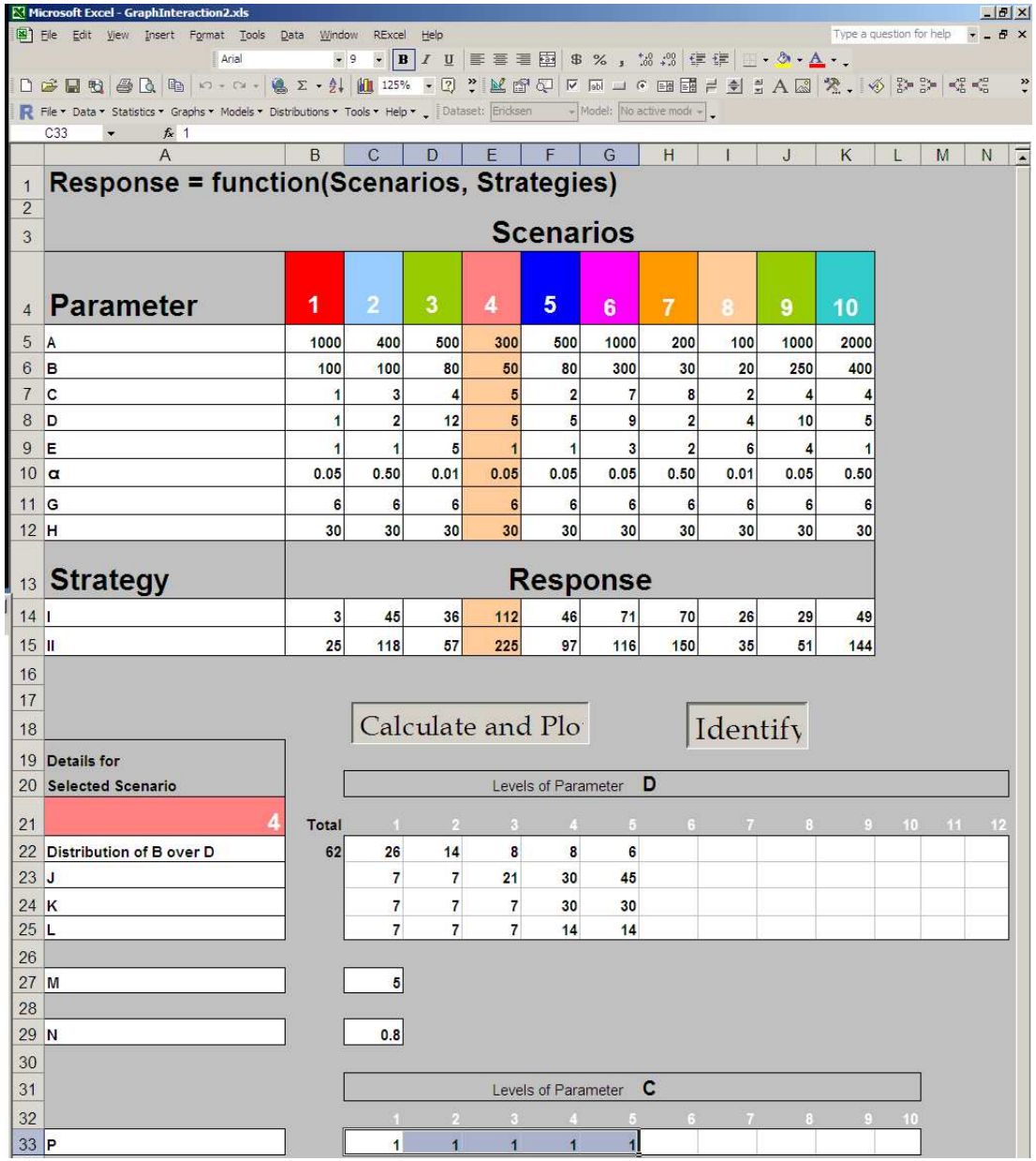


Figure 19: The input parameters and the response values for Scenario 4 are highlighted in a lighter shade of the color displayed in the column header and in the plot. The details for Scenario 4 are displayed in the bottom sections of the worksheet.

4.1 Summary

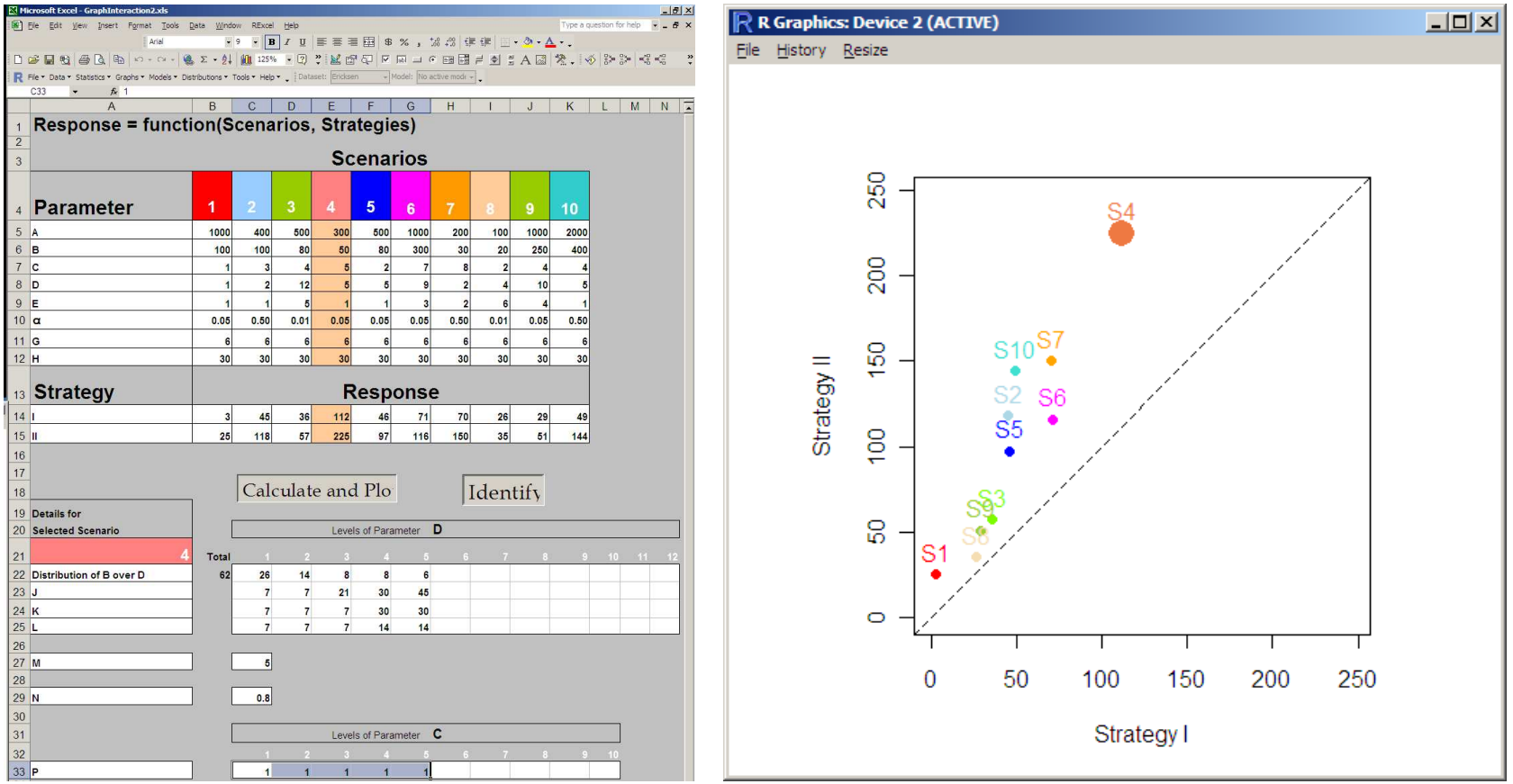


Figure 20: Repeat of Figures 19 and 18. The input parameters and the response values for Scenario 4 are highlighted in a lighter shade of the color displayed in the column header and in the plot. The details for Scenario 4 are displayed in the bottom sections of the worksheet. The dot for Scenario 4 is enlarged in the plot and is the same color as the column header.

4.2 Mechanics of the Interaction

The Calculate and Plot button runs a macro that

1. Sends the Parameter×Scenarios data to R
2. Tells R to run the calculations
3. Brings the Summary information back to Excel
4. Tells R to plots the Summary information

The Identify button runs a macro that

1. runs the `identify()` function in R, allowing the user to click on a point
2. returns the selected point to Excel
3. Brings the detailed information on the selected summary back to Excel.

Acknowledgements

The interaction between R and Excel uses the RExcel package [Neuwirth et al., 2009] and [Baier and Neuwirth, 2007].

The normal.and.t example and the least squares regression example are described in [Heiberger and Neuwirth, 2009a] and are included in the R package [Heiberger and Neuwirth, 2009b]. The HH package accompanies the book [Heiberger and Holland, 2004].

The AEdotplot is described in [Amit et al., 2008]. The R/S-Plus function is included in the HH package [Heiberger, 2009a] and [Heiberger, 2009b]. The RExcel example is included in the RthroughExcel package [Heiberger and Neuwirth, 2009b].

The interaction features of the example in Section 4 are abstracted from a project under development at the GlaxoSmithKline Research Statistics Unit by Sourish Saha and Vladimir Anisimov. This example benefited from discussions with Erich Neuwirth on implementation detail.

References

[Agresti, 1990] Agresti, A. (1990). *Categorical Data Analysis*. Wiley.

[Amit et al., 2008] Amit, O., Heiberger, R. M., and Lane, P. W. (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`.

[Baier and Neuwirth, 2007] Baier, T. and Neuwirth, E. (2007). Excel :: Com :: R. *Computational Statistics*, 22(1):91–108.

[Heiberger, 2009a] Heiberger, R. M. (2009a). HH: Statistical analysis and data display: Heiberger and Holland. R package, `http://www.r-project.org`; contributions from Burt Holland.

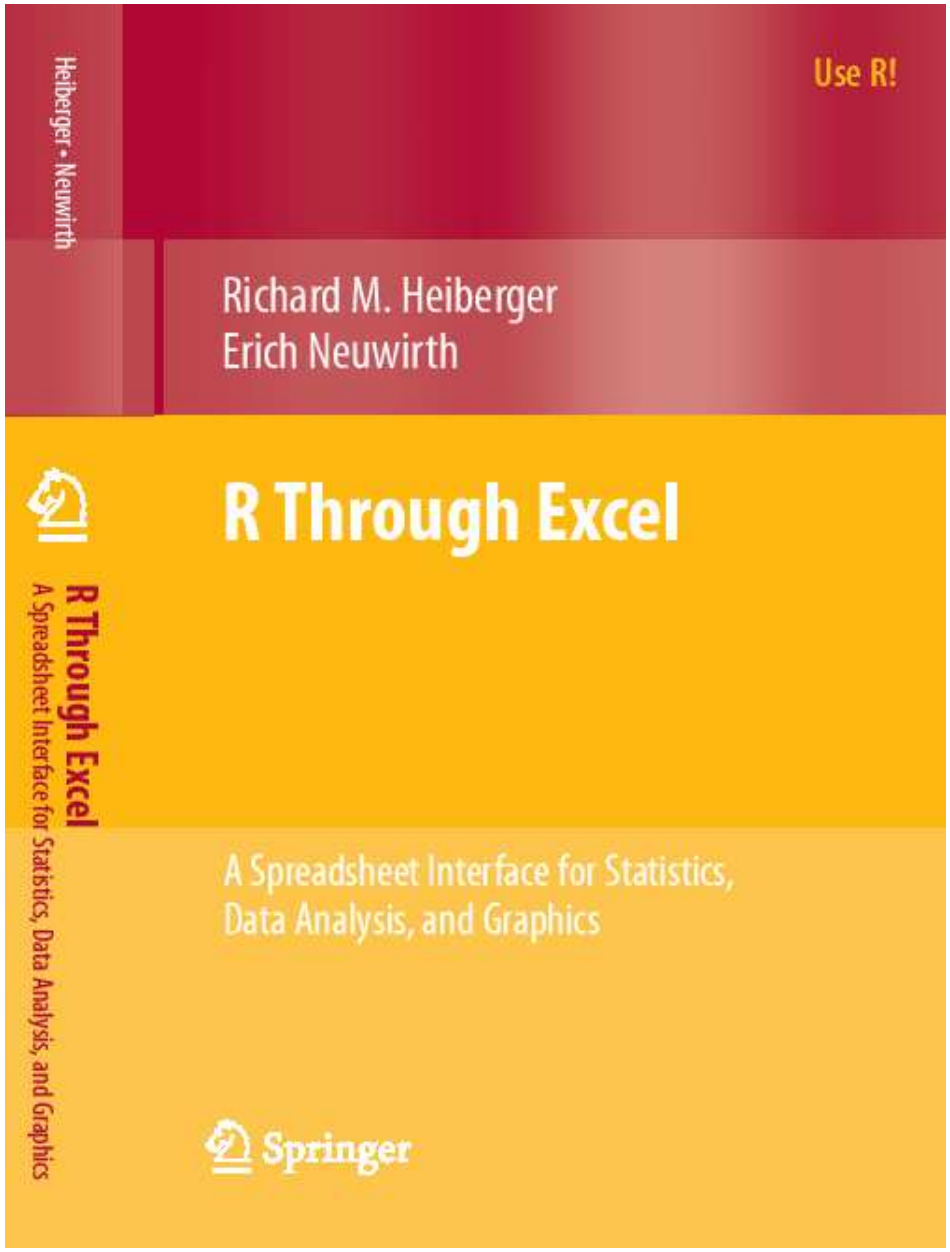
[Heiberger, 2009b] Heiberger, R. M. (2009b). HH: Statistical Analysis and Data Display: Heiberger and Holland. S-Plus package, `http://csan.insightful.com`; contributions from Burt Holland.

[Heiberger and Holland, 2004] Heiberger, R. M. and Holland, B. (2004). *Statistical Analysis and Data Display: An Intermediate Course: Accompanying Online Files*. Springer-Verlag, New York. `http://springeronline.com/0-387-40270-5`.

[Heiberger and Neuwirth, 2009a] Heiberger, R. M. and Neuwirth, E. (2009a). *R Through Excel*. Use R. Springer Verlag.

[Heiberger and Neuwirth, 2009b] Heiberger, R. M. and Neuwirth, E. (2009b). *RthroughExcelWorkbooksInstaller: Excel Workbooks supporting Statistics courses using 'R through Excel'*. R package version 1.1-13.

[Neuwirth et al., 2009] Neuwirth, E., with contributions by Richard Heiberger, Ritter, C., Pieterse, J. K., , and Volkering, J. (2009). *RExcelInstaller: Integration of R and Excel, (use R in Excel, read/write XLS files)*. R package version 3.0-12.


Heiberger • Neuwirth
Use R!
Richard M. Heiberger
Erich Neuwirth
R Through Excel
A Spreadsheet Interface for Statistics, Data Analysis, and Graphics
Springer
R Through Excel
A Spreadsheet Interface for Statistics, Data Analysis, and Graphics