Script for RegressIt video 1: descriptive statistics

1. This video will walk you through the descriptive analysis tools that RegressIt provides. See the instruction handout for details of loading the files and starting up the program. This has already been done here, so the data worksheet is in view and the RegressIt ribbon appears at the top of the screen. There are **40 buttons** on the ribbon, and they provide some analysis and navigation tools that you won't find in any other software. 40 buttons may sound like a lot, but you will find good use for most of them in the course of 20 minutes of work, as you will see in this demonstration. When doing regression analysis, or evaluating work done by others, you will often study and compare the table and chart output of many models, and RegressIt is designed to make that easy and systematic.

2. There are also two alternatives to the ribbon interface: the **vertical and horizontal floating toolbars**. Clicking the **Toolbar V** button (or hitting **Control-G**) launches the vertical toolbar, which has the same button arrangement turned sideways. This is usually a better option for use on small screens or touchscreens, because it maximizes vertical space for output, it has bigger buttons, and your hand doesn't block your view when using it on a touchscreen.

    The horizontal floating toolbar [launched from the **Toolbar H** button] has the same button layout as the ribbon and it also has larger buttons. **The floating toolbars can be placed outside the Excel window, and that needs to be done on a Mac to keep them in view at all times.** [If you launch a floating toolbar on the Mac and it later disappears and seemingly cannot be re-launched, it is probably active but hiding behind the window of Excel or another program.] Here's what the vertical toolbar looks like when placed to the right of the Excel window... and here's what the horizontal toolbar looks like when placed below it.

3. Before running any new analysis, let's try out a tool that provides an aerial view of the contents of the file. Click the **History** button on the RegressIt ribbon. This opens a window that lists all of the worksheets, with some information about their contents, including dependent variable names and bottom-line statistics for regression models. The run time of each analysis and the name of the computer on which it was run are also included, very useful if your work is carried out on more than one computer [at different points in time] or you are an instructor grading student work.

4. This window provides not only an audit trail but also a random access navigation tool. You can click on the line for any worksheet in the file, and then jump directly to it from here by clicking the **Move-To-Sheet** button... This box can be left open while you are working, but I will close it for now.

5. Now let's go to the data [...by hitting the **Data** button on the ribbon]. The data worksheet is always the **first** worksheet in the file. This data set has **6 variables**, one called Y that will serve as the dependent variable in the regression models, and five others called X1 through X5 that are candidates for independent variables. [There is also a date index variable that will be used for plotting.] There are **250 observations** of them, except that **the last 5 values of Y are missing, for purposes of forecasting**. The file contains **3 descriptive statistics analyses and 5 regression models** previously fitted to the data. In this test we will re-run them all.

6. Next let's go to the **Stats 1** worksheet by hitting the **Right** button on the RegressIt ribbon. This descriptive analysis includes all 6 variables and all the table and chart options, which consist **of 3 tables and 48 charts** in this case. Also, Y has been designated as the **first variable** so that it will appear first in tables and in chart arrays, out of alphabetical order. Usually you will want to do this for your dependent variable if you are planning to fit regression models later.

7. Re-run this analysis by going to the RegressIt ribbon and clicking the **Descriptive Statistics** button, which will open the dialog box for this procedure The same variable selections and analysis options will be pre-selected by default, and the default name for the new analysis will be Stats 4, the next number in the sequence of stats sheets. [You can enter a different name if you want.] Just hit the **Run** button to re-run exactly the same analysis.

8. As the output is generated, you will see progress messages in the lower left corner of the Excel window. They're a bit hard to read on a PC because of the dark green background. [When the analysis is complete, the message "Finished with output from Stats 4" will appear.] On this computer the analysis runs in 5 seconds, but it could take 20 seconds or more on yours to produce all the charts.

9. Initially all you will see are the title rows above the tables and chart groups. Hit the **Show All** button on the ribbon to open them all up. There's a lot of output. If you go to the RegressIt ribbon and click the **Zoom Out** button a few times, you can see how much there is…   Now click **Zoom 100%** to restore the original view…   then click the **Last Stats** button a few times. This will toggle you back and forth between this analysis and the original one. They should look identical. Normally you will use this tool to jump back to the last stats sheet from a regression model sheet.

10. Now let's take a closer look at the tables and charts on the Stats 4 output worksheet. Use the **Down** key on the ribbon to move down the worksheet… You could use Excel's own worksheet navigation tools for this, but RegressIt's Up and Down buttons page up and down the worksheet by **whole table** and **whole chart**, so that the topmost one is always perfectly aligned with the top of the screen, and no fine adjustments are needed.   Moving from top to bottom you see the **descriptive statistics table… autocorrelation table… series plots… histogram plots… correlation matrix… and scatterplots.** Each chart is a native Excel chart that is subject to editing if you have chosen the editable-chart option at run time, and they are formatted for presentation with variable names and key statistics in their titles.

11. Use the **Up** arrow to move back up to the **correlation matrix** from the scatterplot matrix, and hit the **Zoom In** button a few times to enlarge it. The correlation matrix has Y in the first column to make it easy to focus on its correlations with the X variables. If you toggle the **Colors** button on the ribbon, color coding will be applied to the values, blue for positive and red for negative, with darker colors for larger magnitudes. The **Fonts** button highlights the magnitudes of numbers by font color and boldness. The same color and font scheme applies to the autocorrelation table above. Notice that both X1 and X4 are highly correlated with Y—X1 a little more so—but they are even more highly correlated with each other, so perhaps there is some redundancy.

12. Click the **Top** button and then the **Zoom 100%** button to restore the original view of this worksheet, then then click the **Left** button a couple of times to move to the **Stats 2** sheet. [Note: the zoom buttons re-set *all* worksheets to the new zoom level.]

13. Click the **Descriptive Statistics** button, followed by **Run,** to re-run the Stats 2 analysis. It differs from the first one in that it includes **high-resolution fixed format charts** rather than editable charts, the series charts are formatted with **vertical bars** instead of points and lines, and the only scatterplots are those of **Y versus the other 5 variables**, a preview of simple linear regression models that might be fitted. [Use the Down button to move down the worksheet to the scatterplots, then hit the **Top** button to return to the top.] Hit **Zoom Out** a few times to see all the output at once, then hit **Zoom 100%,** then toggle the **Last Stats** button to go back and forth to the original.

14. Now hit the **Left** [or right] button a couple of times to move to the **Stats 3** worksheet and hit **Descriptive Statistics** and **Run** to re-run its analysis. This one consists of only a descriptive statistics table and a table of correlations and **squared** correlations versus the first variable. You can use this table to help identify the most important predictors of Y. Here too it is interesting to toggle the **Colors** and **Fonts** buttons on the ribbon to highlight the most significant correlations.

15. As you know, the **squared correlation** between two variables is the value of **R-squared** that you would get in a simple regression of one on the other. That's why it is called R-squared. You can sort this table on squared correlations vs. Y as follows. Click any cell in the table and then click the **Filter** button on the RegressIt ribbon. This turns on arrows that allow you to sort on any column. Click the arrow above the squared correlation column, and choose the largest-to-smallest option. The variables with the smallest squared correlations with Y now appear at the bottom.

16. If you launch a regression model from this sheet, by clicking the **Linear Regression** button, the initial selections of dependent and independent variables will be pulled from it, with the first variable as the dependent variable. All of the other variables will be **pre-selected** as independent variables unless you have **de-selected** them first. Let's cancel out and return to the **Stats 3** sheet and see how this works.

17. To deselect one or more variables, **select a cell or a range of cells that spans their rows in the squared correlation table** and hit the **Remove** button on the RegressIt ribbon. This will cause their lines in the table to be **grayed out**. If you now launch the regression procedure from here, those variables will **not** be in the default selections. Here I am de-selecting variable X2, whose squared correlation with Y is virtually zero. If you click the **Linear Regression** button again, you can see that it is no longer pre-selected. Don't bother to run this model, though. Just hit **Cancel**.

18. This example is rather trivial because there are only 5 candidates for independent variables, but in a real application you might start with 20 or 50 or more variables, in which case sorting on squared correlations would be more helpful in quickly identifying the ones that are potentially the most useful.

19. By the way, it's not *required* to screen out candidate independent variables on the basis of small correlations with the dependent variable, prior to fitting any models. In a multiple regression model, a variable with whose individual correlation with the dependent variable is small could perhaps be useful in combination with other variables, and there could be a logical reason for including it, for example, when it is part of a matched set of dummy variables for mutually exclusive conditions. But in general you should be careful about throwing in a lot of weakly correlated variables without good reason. Don't expect magic to happen.

Now go on to the second video: **linear regression.**