Script for RegressIt video 2: linear regression

1. To continue with the demonstration of RegressIt, let's play with the regression models in the test file. Start on the **Model 1.0** worksheet. Before re-running it, click the **Summaries** button on the ribbon. This will jump you to the **Model Summaries worksheet**, which is always the rightmost [RegressIt] sheet in the file. [There can also be non-RegressIt worksheets to the right of it, such as the "R code" sheet here.] This worksheet is a very distinctive feature of RegressIt. It contains a **journal-article-style table** of side-by-side statistics for all regression models that have ever been fitted in this workbook, even those whose sheets were later deleted. It is a tool for directly comparing model parameters when doing your analysis and when creating reports for presentation. It also provides an even more detailed view of the audit trail for regression models. It contains the summary statistics of the models (R-squared and all that) as well as some diagnostic tests of model assumptions and the estimated coefficients of the variables. Here you see the results for the five regression models that have already been fitted. [Note: the formatting of coefficients is not consistent across these models because different display formats were used for testing. Normally you will stick to one format.]

2. If you toggle the **Show-All** button while positioned on this sheet, you can collapse the display to show only the summary statistics of each model, not the diagnostic stats and coefficients.

3. The **Colors** button can be used to selectively apply color and coding to the coefficient estimates, with blue for positive values and red for negative values, and darker colors for more significant ones, as in the correlation matrix. The **Fonts** button similarly applies darker fonts to more significant values.

4. Now let's check out the linear regression options and fit some new models. Click the **Last Model** button to return to **Model 1.0**, then click the **Linear Regression** button, which opens up the dialog box for running a new model. There are quite a few options for producing output, and by default the same ones that were used in Model 1.0 sheet are checked by default.

5. Before hitting Run, note that there is a **plus sign** next to each of the check-boxes. If you click the plus sign, you will get **a pop-up note that explains its use.** For example, here's the note that explains the normal quantile plot. So, if the options look complicated at first, you can learn about their functions by reading the notes.

6. One more thing to do before hitting Run: **check the box for the table of residuals and influence measures.** It was included in the Model 1.0 sheet, but this feature does not have memory because the table could be huge for a large data set.

7. Now hit the **Run** button without changing anything else, in order to re-run exactly the same analysis.

8. After the new model sheet is produced, click the **Zoom Out** button a few times to see all the output, which includes 7 charts in this case [plus one more that is currently hidden]… then **zoom back to 100%.** Now hit the **Last Model** button to toggle back and forth between the new model and the old one to see that they are the same. Normally you will use this tool to toggle between models that are different.

9. This model is a simple regression of Y on X1. The **line fit plot** appears near the top of the worksheet, and it includes confidence bands around the regression line. The editable chart option has been used, and you can vary the confidence level by clicking the **Confidence-plus** and **Confidence-minus** buttons on the ribbon, and watch the effects. The bands get wider and narrower as you do this.

10. Now use the **Down** button on the RegressIt ribbon to page down the sheet by table and chart. This analysis includes **out-of-sample forecasts for the 5 missing values of Y**. The forecasts appear in both a table and chart whose confidence limits are interactive. The **Confidence-plus** and **Confidence-minus** button make the limits wider and narrower here too, in both the table and chart.

11. If you jump further down the worksheet with the **Down** button, you will see a number of different plots of the predictions and errors of the model. At the very bottom you will find the **residual table**, which includes predictions and several error and influence measures. [The residual table is hidden by default, because it could be huge: you will need to click its plus sign in the sidebar to unhide it. The residual-vs-independent-variable plots are also hidden by default, because there could be a large number of them, and the check-box to display them is above the one for the residual table. Here we have a single such plot. *Hit the Up button to return to other charts above*.]

12. All of the chart titles on the worksheet include the **model name, dependent variable, number of variables and sample size** by default. This is useful information to have in view when flipping back and forth between worksheets or when copying charts to reports. (The titles can also be customized—more about that later.) Now hit the **Top** button to return to the top of the worksheet.

13. The **tables** have the same information in **title rows** above them. These rows are hidden by default when the model is first run, but you can display them by toggling the **Titles** button on the ribbon, as shown here. You should include this row for audit trail purposes when copying a table elsewhere.

14. If you turn on the **Titles** and then toggle the **Show-All** button to hide all the tables and charts, you will see just an **outline view** of the contents of the worksheet, from which you can selectively open a single kind of output if you wish… [Open the line fit plot.] Now hit the **Show-All** button again to re-open all the output, and toggle the **Titles** button to re-hide the titles.

15. By the way, one of the other outputs at the top of the sheet is the **model equation** in text form, which is sometimes useful for copying to reports.

16. Notice that when we re-ran model 1.0, the default name for the new model was **Model 1.1**. This illustrates the default model numbering scheme used by RegressIt. **If the model name ends in a number, preceded by a space or period, the default name of the next model is the same except with that number increased by 1.** You can also type any name you want when you run a new model.

17. Here's another important feature to look at before we continue: the **teaching notes** feature. Notice that many cells on the model sheet have little **red flags** in their upper right corners. This

means that they contain comments, which could consist of many paragraphs of text**. If you move the mouse over a cell with a red flag, the comment will pop up.**

18. Most of the comments are teaching notes that explain how to interpret the various statistics and charts. For example here's the note that explains R-squared… . Other comments contain more layers of detail about the model. Here's what's underneath the **model name** cell... it includes audit trail information such as the **run time, current file name, and computer name**.

19. You can also display the comment by clicking on the cell and hitting the **View** button on the ribbon. This will cause the comment to be displayed in the top center of the Excel window, which is better in narrow windows or on touchscreens. The red flags can be toggled on and off by hitting the Notes button.

20. At present RegressIt contains about 10,000 words of pop-up teaching notes and check-box instructions, and much more is planned. These notes could be customized to fit a course or program—let us know if you are interested.

21. Now hit the **Last Model** button to return to the Model 1.0 sheet… then hit the **Right** button to move to **Model 2.0**. This is a simple regression model with no constant, which means that the regression line is forced to pass through the origin of coordinates, which is not a good idea in this case. Normally you don't remove the constant. This model is included only for purposes of testing all model options.

22. Before re-running it, notice that something funny is going on. Model 2.0 obviously doesn't fit the data as well as Model 1.0, yet its R-squared is much bigger, 95%, compared to about 50% for Model 1. [Toggle the **Last Model** button to go back and forth.] Does this mean Model 2.0 is better? No! When there is no constant, R-squared does not have the same definition. It measures the improvement over a model that predicts all values of Y to be zero, whereas in a model with a constant, R-squared measures the improvement over a model that predicts all values of Y to equal the mean. Don't judge a model by R-squared! Keep your eye on the bottom line, which is usually the standard error of the regression.

23. Now hit the **Run** button to re-run Model 2.0, just to make sure it works OK for you. You'll get a new sheet called Model 2.1.

24. Next, hit the **Last Model** button followed by the **Right** button to go to the worksheet for **Model 3.0**, which is a multiple regression model with all 5 X variables. Click the **Linear Regression** button… then hit **Run** to re-run this model… and you should get the same output on a new sheet called **Model 3.1**

25. You can toggle the **Colors** and **Fonts** buttons here to apply color and font coding to the t-statistics and standardized coefficients of the variables, to highlight the ones that are most significant. The formatting of the t-stats is the same as on the Model Summaries sheet.

26. Not all of the independent variables are statistically significant in this model. In particular, variables X2 and X4 have P-values well above 0.05. Recall that in the descriptive analysis we found that X2 had almost zero correlation with Y, and X4 seemed like it might be redundant with X1.

27. You can **sort the coefficient table on P-values**, which is helpful for finding the least significant ones when there is a large number of variables. Let's do that now: position the cursor anywhere in the coefficient table… click the **Filter** button on the ribbon… then click the arrow above the P-value column to sort from smallest to largest. This will send X2 and X4 to the bottom.

28. If you hit the **Linear Regression** button to launch another regression model from this sheet, its defaults will include the same variables and output options by default, as we saw before.

29. If you want to remove insignificant variables before re- launching the regression procedure, click on their rows in the coefficient table and hit the **Remove** button on the ribbon. Toggling the Remove button causes a variable's line in the coefficient table to be grayed out, as I am doing here with X4, and if you launch another regression model when it is grayed out… it will no longer be preselected. This method of sequentially removing variables is usually more efficient than looking them up in the original list of all variables and unchecking their boxes.

30. Let's go ahead and run the model without X4, getting a new sheet called Model 3.2. Here we see that X2 remains very insignificant. Let's **click on its row** and hit the **Remove** button to flag it for removal from the next model, and turn one another model option: customized titles. [Click **the Linear Regression** button and then click the **Customize Titles** check-box next to the model name in the dialog box.] This is something you will occasionally want to do for presentation. After hitting **Run**, you get a prompt for a custom model name to be used in table and chart titles. Let's enter **"Final model for Y".** Here's the output, and if you click the **Titles** button, you can can see that the custom title is used in both tables and charts.

31. If you click the **History** button to display the updated history list, you will see that it now includes the 3 new stats analyses and 5 new regression models, and you will see the **custom name** among the model parameters for **Model 3.3**.

32. While the History box is still open, use it to jump to **Model 4.0**, which is the same as Model 3.0 except with no constant. The no-constant option is included only for completeness in testing model types. As in Model 2.0, it has a suspiciously high value of R-squared, but this doesn't mean it is a good model. Hit Run to re-run it, just to make sure it works for you.

33. Finally, re-run **Model 5.0**. Hit the **Last Model** button and then the **Right** button to get to its sheet. This is a constant-only model, which merely predicts all values of Y to equal the mean. It has an R-squared of zero by definition. This too is a model that you don't normally bother to fit, but it provides a baseline against to which compare the models that also include one or more independent variables. [Hit the **Linear Regression** button followed by **Run** to re-run it.]

34. When you fit a constant-only model [and choose the line fit plot option as has been done here], you have the option to specify an independent variable against which to plot the regression line, which is a horizontal line rather than a sloping line in this case. Let's **specify X1 as the variable to plot against**, for comparison to the original simple regression model. [Then click **Continue-With-Selected-Variable**.]

35. Here's the output, which includes the horizontal regression line. Use the **History** tool to jump back to Model 1.0…. Then toggle the **Last Model** button to toggle back and forth to **Model 5.1** and see the difference in their line fit plots: horizontal versus sloping. Both lines pass through the center of mass of the data: the means of both variables. If you go from Model 1.0 to model **2.0**, you see the third possibility for a regression line: one that passes through the origin.

36. Another tool that can be used to move directly from one regression model to another is the **Relatives** button. If you click this button, you see a box that shows the parents and children of this model, if any. This provides you with yet another view of the audit trail, as well as a tool for further exploring the model space. Go back to Model **1.0**… and hit this button. **Model 1.0 has three children, Models 2.0 and 5.0 as well as 1.1.** This means that the other models were originally created by starting from the model 1.0 sheet, launching a new model from there, and perhaps making some changes in the variables or output options. If you click on one of the children and hit the **Go To** button, you jump to its model sheet. [Go to Model 2.0.] You can similarly back up to the parent of a model. [Go back to Model 1.0 here.] Sometimes you will want to back up to an earlier point in a model sequence and branch off in a new direction from there.

37. One more important tool on the ribbon is the **Compare** button. It re-sets the viewpoint of all model worksheets in the file to agree with the one currently in view so that the same table or chart appears at very top. If you click this button and then jump from one model sheet to another, you can easily compare particular charts or tables in their output. Let's use the **Down** button to jump down to **the forecast table and chart** on the **Model 1.0** sheet, and then hit the **Compare** button. If we now move to the sheets of other models, we can directly compare their forecasts.

38. That's all for this demonstration. I encourage you to do some more playing around in the file to get a better feel for the modeling options and navigation options that are provided by RegressIt's procedure menus and ribbon interface. And of course, try it out on your own data.