

Applied Regression Modeling: A Business Approach

Computer software help: Excel

Microsoft Excel is a commercial spreadsheet package with an easy-to-use graphical user-interface that is capable of carrying out a few basic statistical analyses. Further information is available at www.microsoft.com/excel/. The following instructions are based on “Microsoft Office Excel 2003 for Windows.” The book website contains supplementary material for other versions of Excel. This section contains less material than the preceding sections for SPSS, Minitab, SAS, and R because it is not possible to easily carry out many of the techniques discussed in this book using Excel. However, there are a number of add-on modules available for Excel that can improve its statistical analysis capabilities, for example, StatTools (available at www.palisade.com/stattools/) and Lumenaut Statistics Package (available at www.lumenaut.com/statistics.htm).

Getting started and summarizing univariate data

- 1 Change Excel’s default **options** by selecting Tools > Options.

Make sure Excel’s Data Analysis functions are available by selecting Tools > Add-Ins and checking Analysis ToolPak.

To open an Excel **data file**, select File > Open.

Output can be copied and pasted from Excel to a word processor like Microsoft Word. Graphs can also easily be copied and pasted to other applications.

- 2 You can access **help** by selecting Help > Microsoft Excel Help.

For example, to find out about “scatterplots,” type scatter plot in the search box.

- 3 To **transform data** or compute a **new variable**, type, for example, =LN(X) for the natural logarithm of X and =X^2 for X^2 .

To create **indicator (dummy) variables** from a qualitative variable, type, for example, =IF(X="level", 1, 0), where X is the qualitative variable and "level" is the name of one of the categories in X . Repeat for other indicator variables (if necessary).

- 4 Calculate **descriptive statistics** for quantitative variables by selecting

Tools > Data Analysis > Descriptive Statistics.

Select the Input Range to include the variable(s) of interest, check Labels in first row if appropriate, and check Summary statistics.

- 5 Create **contingency tables** or **cross-tabulations** for qualitative variables by selecting Data > PivotTable and PivotChart Report.

Select Microsoft Office Excel list or database as the data to be analyzed and PivotTable as the report to be created. Next, select an appropriate data range and put the PivotTable report in a new worksheet. Drag one qualitative variable to the Column Fields space, another qualitative variable to the Row Fields space, and drag some other convenient variable to the Data Items space. The resulting table should show sums of the variable in the Data Items for different combinations of the qualitative variable categories. To

change the sums to frequencies, double-click on cell A3 and change Sum to Count. To calculate row and column percentages, click the Options button.

6 If you have quantitative variables and qualitative variables, you can calculate **descriptive statistics** for cases grouped in different categories by creating a PivotTable (see computer help #5) and double-clicking on cell A3 to select different summary functions.

7 Excel does not appear to offer an automatic way to create a stem-and-leaf plot.

To make a **histogram** for a quantitative variable, select Tools > Data Analysis > Histogram.

Select the Input Range to include the variable of interest, check Labels if appropriate, and check Chart Output.

8 To make a **scatterplot** with two quantitative variables, select Insert > Chart > XY (Scatter).

In Step 2 of the Chart Wizard click on the Series tab, select the appropriate data ranges for the X Values, Y Values, and Name boxes, and click Finish.

Excel does not appear to offer an automatic way to create a scatterplot matrix.

9 Excel does not appear to offer an automatic way to create a scatterplot with different colors/symbols marking the categories in a qualitative variable.

However, you can **identify individual cases** in a scatterplot by hovering over them.

10 To make a **bar chart** for cases in different categories, first create a PivotTable (see computer help #5) of cell frequencies. Then select Insert > Chart to create a bar chart.

You may need to subsequently click the Chart Wizard tool to change the chart type (e.g., from stacked bars to clustered bars).

The bars can also represent various summary functions for a quantitative variable. For example, double-click on the cell that says Count of ... in the PivotChart worksheet and change it to Average to make the bar chart represent Means.

11 Excel does not appear to offer an automatic way to create boxplots.

12 Excel can make a **QQ-plot** (also known as a **normal probability plot**) for a quantitative variable automatically, but only through the Regression tool.

For example, select Tools > Data Analysis > Regression, then select the Input Y Range to include the response variable, select the Input X Range to include the predictor variable(s), and check Labels if appropriate. The predictor variables should be in adjacent columns in the spreadsheet for this to work. Finally, check Normal Probability Plots before hitting OK to produce a QQ-plot for the response variable in the regression.

13 To compute a **confidence interval** for a univariate population mean, select Tools > Data Analysis > Descriptive Statistics.

Select the Input Range to include the variable of interest, check Labels in first row if appropriate, check Summary statistics, check Confidence Level for Mean, and type the confidence level into the box.

The resulting Confidence Level value represents the “uncertainty” in the intervals. In other words, the interval goes from the sample mean minus this uncertainty up to the sample mean plus this uncertainty.

- 14 Excel does not appear to offer an automatic way to do a hypothesis test for a univariate population mean.

It is possible to do the test by hand calculation using Excel descriptive statistics output and appropriate percentiles from a t-distribution.

Simple linear regression

- 15 To fit a **simple linear regression model** (i.e., find a least squares line), select Tools > Data Analysis > Regression.

Select the Input Y Range to include the response variable, select the Input X Range to include the predictor variable, and check Labels if appropriate. Just hit OK for now—the other items in the dialog box are addressed below.

- 16 To add a **regression line** or **least squares line** to a scatterplot, select the plot by clicking on it and select Chart > Add Trendline.

This brings up another dialog in which you need to make sure Linear is selected under Trend/Regression type. Hit OK to add the least squares line to the plot.

- 17 In fitting a simple linear regression model (see computer help #15), Excel automatically finds 95% **confidence intervals for the regression parameters**.

This applies more generally to multiple linear regression also.

- 18 Excel does not appear to offer an automatic way to find a confidence interval for the mean of Y at a particular value of X in a linear regression model.

- 19 Excel does not appear to offer an automatic way to find a prediction interval for an individual value of Y at a particular value of X in a linear regression model.

Multiple linear regression

- 20 To fit a **multiple linear regression model**, select Tools > Data Analysis > Regression.

Select the Input Y Range to include the response variable, select the Input X Range to include the predictor variables, and check Labels if appropriate. The predictor variables should be in adjacent columns in the spreadsheet for this to work.

- 21 To add a **quadratic regression line** to a scatterplot, select the plot by clicking on it, and select Chart > Add Trendline.

This brings up another dialog in which you need to make sure Polynomial with Order 2 is selected under Trend/Regression type. Hit OK to add the quadratic regression line to the plot.

- 22 Excel does not appear to offer an automatic way to create a scatterplot with separate regression lines for subsets of the sample.

- 23 Excel does not appear to offer an automatic way to find the F-statistic and associated p-value for a nested model F-test in multiple linear regression.

It is possible to calculate these quantities by hand using Excel regression output and appropriate percentiles from a F-distribution.

- 24 Excel does not appear to offer an automatic way to save studentized residuals in a multiple linear regression model. However, it does calculate crude **standardized residuals**, which it defines as ordinary residuals divided by their standard deviation.

In particular, select Tools > Data Analysis > Regression. Select Input Y Range to include the response variable, select Input X Range to include the predictor variables, and check Labels if appropriate. Predictor variables should be in adjacent columns in the spreadsheet for this to work. Finally, check Standardized Residuals before hitting OK.

- 25 Excel does not appear to offer an automatic way to add a loess fitted line to a scatterplot.

- 26 Excel does not appear to offer an automatic way to save leverages in a multiple linear regression model.

- 27 Excel does not appear to offer an automatic way to save Cook's distances in a multiple linear regression model.

- 28 To create some **residual plots** automatically in a multiple linear regression model, select Tools > Data Analysis > Regression.

Select the Input Y Range to include the response variable, select the Input X Range to include the predictor variables, and check Labels if appropriate. The predictor variables should be in adjacent columns in the spreadsheet for this to work. Finally, check Residual Plots before hitting OK to create residual plots with each predictor variable on the horizontal axis.

To create residual plots manually, first create standardized residuals (see computer help #24), and then construct scatterplots with these standardized residuals on the vertical axis. In particular, you should plot them against the Excel-provided fitted (predicted) values for the regression.

- 29 To create a **correlation matrix** of quantitative variables (useful for checking potential **multicollinearity** problems), select Tools > Data Analysis > Correlation.

Select the Input Range to include the variables of interest and check Labels in First Row if appropriate. The variables should be in adjacent columns in the spreadsheet.

- 30 Excel does not appear to offer an automatic way to find variance inflation factors in a multiple linear regression model.

- 31 To draw a **predictor effect plot** for graphically displaying the effects of quantitative predictors in multiple linear regression, first create a variable representing the effect, say, "X1effect" (see computer help #3)—this variable must just involve X1 (e.g., $1+3X1+4X1^2$).

Then sort the X1 variable in ascending order using Data > Sort, and select Insert > Chart > XY (Scatter)—select the plot with data points connected by smoothed lines without markers.

v

In Step 2 of the Chart Wizard click on the Series tab, select the appropriate data ranges for the X Values (sorted X1 variable), Y Values (“X1 effect” variable), and Name boxes, and click Finish.

Excel does not appear to offer an automatic way to create more complex predictor effect plots (say, with separate lines representing different subsets of the sample).