

Applied Regression Modeling: A Business Approach

Computer software help: SPSS

SPSS (originally “Statistical Package for the Social Sciences”) is a commercial statistical software package with an easy-to-use graphical user-interface. Further information is available at www.spss.com. The following instructions are based on “SPSS 16.0 for Windows.” The book website contains supplementary material for other versions of SPSS.

Getting started and summarizing univariate data

- 1 If desired, change SPSS’s default **options** by selecting Edit > Options.

For example, to display variable names (in alphabetical order) rather than labels in dialog boxes, click the General tab; in the Variable Lists group select Display names and select Alphabetical. To show variable names rather than labels in output tables, click the Output Labels tab; under Pivot Table Labeling change Variables in labels shown as to Names. To display small numbers in tables without using scientific notation (which can make reading the numbers more difficult), click the General tab; under Output check No scientific notation for small numbers in tables.

To open a SPSS **data file**, select File > Open > Data.

To **recall** a previously used dialog box, hit the Dialog Recall tool (fourth button from the left in the Data Editor Window, sixth button from the left in the Viewer Window).

Output can be edited in the Viewer Window. Individual pieces of output (including tables and graphs) can be selected, edited, moved, deleted, and so on using both the Outline Pane (on the left) and the Display Pane (on the right). Text and headings can be entered using the Insert menu. Alternatively, copy and paste pieces of output from SPSS to a word processor like Microsoft Word.

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

For example, to find out about “boxplots,” click the Index tab, type boxplots in the first box, and select the index entry you want in the second box.

- 3 To **transform data** or compute a **new variable**, select Transform > Compute Variable.

Type a name (with no spaces) for the new variable in the Target Variable box, and type a mathematical expression for the variable in the Numeric Expression box. Current variables in the dataset can be moved into the Numeric Expression box, while the keypad and list of functions can be used to create the expression. Examples are LN(X) for the natural logarithm of X and X**2 for X^2 . Hit OK to create the new variable, which will be added to the dataset (check it looks correct in the Data Editor Window); it can now be used just like any other variable. If you get the error message “expression ends unexpectedly,” this means there is a syntax error in your Numeric Expression—a common mistake is to forget the multiplication symbol (*) between a number and a variable (e.g., 2*X represents 2X).

To create **indicator (dummy) variables** from a qualitative variable, select Transform > Recode into Different Variables.

Move the qualitative variable into the Input Variable -> Output Variable box, type a name for the first indicator variable in the Output Variable Name box, and press Change (the name should replace the question mark in the Input Variable -> Output Variable box). Next, press Old and New Values, type the appropriate category name/number into the Old Value box, type 1 into the New Value box, and press Add. Then select All other values, type 0 into the New Value box, and press Add. Click Continue to return to the previous dialog box, and hit OK (check that the correct indicator variable has been added to your spreadsheet in the Data Editor Window). Repeat for other indicator variables (if necessary).

- 4 Calculate **descriptive statistics** for quantitative variables by selecting Analyze > Descriptive Statistics > Frequencies.

Move the variable(s) into the Variable(s) list. Click Statistics to select the summaries, such as the Mean, that you would like. To avoid superfluous output uncheck Display frequency tables.

- 5 Create **contingency tables** or **cross-tabulations** for qualitative variables by selecting Analyze > Descriptive Statistics > Crosstabs.

Move one qualitative variable into the Row(s) list and another into the Column(s) list. Cell percentages (within rows, columns, or the whole table) can be calculated by clicking Cells.

- 6 If you have quantitative variables and qualitative variables, you can calculate **descriptive statistics** for cases grouped in different categories by selecting Analyze > Reports > Case Summaries.

Move the quantitative variable(s) into the Variables list and the qualitative variable(s) into the Grouping Variable(s) list. Click Statistics to select the summaries that you would like; the default is Number of Cases, but other statistics such as the Mean and Standard Deviation can also be selected. To avoid superfluous output uncheck Display cases.

- 7 To make a **stem-and-leaf plot** for a quantitative variable, select Analyze > Descriptive Statistics > Explore.

Move the variable into the Dependent List box. You can alter the statistics that are calculated and the plots that are constructed by clicking Statistics and Plots.

To make a **histogram** for a quantitative variable, select Graphs > Legacy Dialogs > Histogram.

Move the variable into the Variable box.

- 8 To make a **scatterplot** with two quantitative variables, select Graphs > Legacy Dialogs > Scatter/Dot.

Choose Simple Scatter and move the vertical axis variable into the Y Axis box and the horizontal axis variable into the X Axis box.

All possible scatterplots for more than two variables can be drawn simultaneously (called a **scatterplot matrix**) by choosing Matrix Scatter and moving the variables into the Matrix Variables list.

- 9 You can **mark or label cases** in a scatterplot with different colors/symbols according to the categories in a qualitative variable by moving the variable into the Set Markers by box in the Scatterplot dialog.

To change the colors/symbols used, edit the plot (double-click it in the Viewer Window) to bring up a Chart Editor Window, select the symbol you want to change by clicking on it in the legend at the right of the plot (the data points corresponding to this symbol should become highlighted when you do this), and select Edit > Properties. Select the color/symbol you want and hit Apply to see the effect. Hit Close to return to the plot; close the plot to return to the Viewer Window.

You can also **identify individual cases** in a scatterplot using labels by moving a qualitative text variable into the Label Cases by box in the Scatterplot dialog. This has no apparent effect on the plot when it is first drawn, but if you subsequently edit the plot (double-click it in the Viewer Window) to bring up a Chart Editor Window, you can then use the Point Identification tool (under Elements > Data Label Mode) to click on a point and the label for that point will be displayed.

- 10 To make a **bar chart** for cases in different categories, select Graphs > Legacy Dialogs > Bar.

For frequency bar charts of one qualitative variable, choose Simple and move the variable into the Category Axis box. For frequency bar charts of two qualitative variables choose Clustered and move one variable into the Category Axis box and the other into the Define Clusters by box. The bars can also represent various summary functions for a quantitative variable. For example, to represent Means, select Other statistic (e.g., mean) and move the quantitative variable into the Variable box.

- 11 To make **boxplots** for cases in different categories, select Graphs > Legacy Dialogs > Boxplot.

For just one qualitative variable, choose Simple and move the qualitative variable into the Category Axis box. Move the quantitative variable into the Variable box.

For two qualitative variables, choose Clustered and move one qualitative variable into the Category Axis box and the other into the Define Clusters by box. Move the quantitative variable into the Variable box.

- 12 To make a **QQ-plot** (also known as a **normal probability plot**) for a quantitative variable, select Analyze > Descriptive Statistics > Q-Q Plots.

Move the variable into the Variables box and leave the Test Distribution as Normal to assess normality of the variable. This procedure produces a regular QQ-plot (described in Section 1.2) as well as a “detrended” one.

- 13 To compute a **confidence interval** for a univariate population mean, select Analyze > Descriptive Statistics > Explore.

Move the variable for which you want to calculate the confidence interval into the Dependent List box and select Statistics for Display. Then click the Statistics button to bring up another dialog box in which you can specify the confidence level for the interval (among other things). Continue will take you back to the previous dialog box, where you can now hit OK.

- 14 To do a **hypothesis test** for a univariate population mean, select Analyze > Compare Means > One-Sample T Test.

Move the variable for which you want to do the test into the Test Variable(s) box and type the (null) hypothesized value into the Test Value box. The p-value calculated (displayed as “Sig.”) is a two tail p-value; to obtain a one tail p-value you will either need to divide this value by two or subtract it from one and then divide by two (draw a picture to figure out which).

Simple linear regression

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

Move the response variable into the Dependent box and the predictor variable into the Independent(s) box. Just hit OK for now—the other items in the dialog box are addressed below. In the output, ignore the column headed “Standardized Coefficients.”

- 16 To add a **regression line** or **least squares line** to a scatterplot, edit the plot (double-click it in the Viewer Window) to bring up a Chart Editor Window and select Elements > Fit Line at Total.

This brings up another dialog in which you need to make sure Linear is selected under Fit Method. Hit Close to add the least squares line and return to the plot; close the plot to return to the Viewer Window.

- 17 To find 95% **confidence intervals for the regression parameters** in a simple linear regression model, select Analyze > Regression > Linear.

Move the response variable into the Dependent box and the predictor variable into the Independent(s) box. Before hitting OK, click the Statistics button and check Confidence intervals (under Regression Coefficient) in the subsequent Linear Regression: Statistics dialog box. Click Continue to return to the main Linear Regression dialog box, and then hit OK. The confidence intervals are displayed as the final two columns of the “Coefficients” output.

This applies more generally to multiple linear regression also.

- 18 To find a **confidence interval for the mean of Y** at a particular value of X in a simple linear regression model, select Analyze > Regression > Linear.

Move the response variable into the Dependent box and the predictor variable into the Independent(s) box. Before hitting OK, click the Save button and check Mean (under Prediction Intervals) in the subsequent Linear Regression: Save dialog box. Type the value of the confidence level that you want in the Confidence Interval box (the default is 95%), click Continue to return to the main Linear Regression dialog box, and then hit OK.

The confidence intervals for the mean of Y at each of the X-values in the dataset are displayed as two columns headed LMCI.1 and UMCI.1 in the Data Editor Window (not in the Viewer Window). The “LMCI” stands for “lower mean confidence interval,” while the “UMCI” stands for “upper mean confidence interval.” Each time you ask SPSS to calculate confidence intervals like this it will add new columns to the dataset and increment the end digit by one; for example, the second time you calculate confidence intervals for the mean of Y the end points will be called LMCI.2 and UMCI.2.

You can also obtain a confidence interval for the mean of Y at an X-value that is not in the dataset by doing the following. Before fitting the regression model, add the X-value to the dataset in the Data Editor Window (go down to the bottom of the spreadsheet, and type the X-value in the appropriate cell of the next blank row). Then fit the regression model and follow the steps above. SPSS will ignore the X-value you typed when fitting the model (since there is no corresponding Y-value), so all the regression output (such as the estimated regression parameters) will be the same. But SPSS will calculate a confidence interval for the mean of Y at this new X-value based on the results of the regression. Again, look for it in the dataset; it will be displayed in the two columns headed LMCI and UMCI in the Data Editor Window (not in the Viewer Window).

This applies more generally to multiple linear regression also.

- 19 To find a **prediction interval** for an individual value of Y at a particular value of X in a simple linear regression model, select Analyze > Regression > Linear.

Move the response variable into the Dependent box and the predictor variable into the Independent(s) box. Before hitting OK, click the Save button and check Individual (under Prediction Intervals) in the subsequent Linear Regression: Save dialog box. Type the value of the confidence level that you want in the Confidence Interval box (the default is 95), click Continue to return to the main Linear Regression dialog box, and then hit OK.

The prediction intervals for an individual Y-value at each of the X-values in the dataset are displayed as two columns headed LICI.1 and UICI.1 in the Data Editor Window (not in the Viewer Window). The “LICI” stands for “lower individual confidence interval,” while the “UICI” stands for “upper individual confidence interval.” We call them prediction (not confidence) intervals. Each time you ask SPSS to calculate prediction intervals like this it will add new columns to the dataset and increment the end digit by one; for example, the second time you calculate prediction intervals for an individual value of Y the end points will be called LICI.2 and UICI.2.

You can also obtain a prediction interval for an individual Y-value at an X-value that is not in the dataset by doing the following. Before fitting the regression model, add the X-value to the dataset in the Data Editor Window (go down to the bottom of the spreadsheet, and type the X-value in the appropriate cell of the next blank row). Then fit the regression model and follow the steps above. SPSS will ignore the X-value you typed when fitting the model (since there is no corresponding Y-value), so all the regression output (such as the estimated regression parameters) will be the same. But SPSS will calculate a prediction interval for an individual Y at this new X-value based on the results of the regression. Again, look for it in the dataset; it will be displayed in the two columns headed LICI and UICI in the Data Editor Window (not in the Viewer Window).

This applies more generally to multiple linear regression also.

Multiple linear regression

- 20 To fit a **multiple linear regression model**, select Analyze > Regression > Linear.
Move the response variable into the Dependent box and the predictor variables into the Independent(s) box.
- 21 To add a **quadratic regression line** to a scatterplot, edit the plot (double-click it in the Viewer Window) to bring up a Chart Editor Window and select Elements > Fit Line at Total.
This brings up another dialog in which you need to check the Quadratic option under Fit Method. Hit Apply and Close to add the quadratic regression line and return to the plot; close the plot to return to the Viewer Window.
- 22 Categories of a qualitative variable can be thought of as defining **subsets** of the sample. If there are also a quantitative response and a quantitative predictor variable in the dataset, a regression model can be fit to the data that represents separate regression lines for each subset. First use computer help #8 and #9 to make a scatterplot with the response variable on the vertical axis, the quantitative predictor variable on the horizontal axis, and the cases marked with different colors/symbols according to the categories in the qualitative predictor variable. To add a **regression line for each subset** to this scatterplot, edit the plot (double-click it in the Viewer Window) to bring up a Chart Editor Window and select Elements > Fit Line at Subgroups.
This brings up another dialog in which you need to make sure Linear is selected under Fit Method. Hit Close to add the least squares lines for each subset of selected points and return to the plot. Close the plot to return to the Viewer Window.
- 23 To find the F-statistic and associated p-value for a **nested model F-test** in multiple linear regression, select Analyze > Regression > Linear.
Move the response variable into the Dependent box and the predictor variables in the *reduced* model into the Independent(s) box. Click the Next button to the right of where it says Block 1 of 1; it should now say Block 2 of 2 and the Independent(s) box should have been cleared. Move the *additional* predictors in the *complete* model (i.e., the predictors whose usefulness you are assessing) into this Block 2 Independent(s) box. You should now have the predictors that are in *both* the reduced and complete models in Block 1, and the predictors that are *only* in the complete model in Block 2. Then click Statistics and check R squared change. Finally click Continue to return to the Regression dialog and OK to obtain the results. The F-statistic is in the second row of the “Model Summary” in the column headed F Change, while the associated p-value is in the column headed Sig. (Ignore the numbers in the first rows of these columns.)
- 24 To save **studentized residuals** in a multiple linear regression model, select Analyze > Regression > Linear.
Move the response variable into the Dependent box and the predictor variables into the Independent(s) box. Before hitting OK, click the Save button and check Studentized under Residuals in the subsequent Linear Regression: Save dialog box. Click Continue to return to the main Linear Regression dialog box, and then hit OK. The studentized residuals are saved as a variable called SRE.1 in the Data Editor Window; they can now be used just like any other variable, for example, to construct residual plots. Each time you ask SPSS to save studentized residuals like this it will add a new variable

to the dataset and increment the end digit by one; for example, the second time you save studentized residuals they will be called SRE.2.

- 25 To add a **loess fitted line** to a scatterplot (useful for checking the zero mean regression assumption in a residual plot), edit the plot (double-click it in the Viewer Window) to bring up a Chart Editor Window and select Elements > Fit Line at Total.

This brings up another dialog in which you need to check the Loess option under Fit Method. The default value of 50 for % of points to fit tends to be a little on the low side: I would change it to 75. Hit Apply and Close to add the loess fitted line and to return to the plot; close the plot to return to the Viewer Window.

- 26 To save **leverages** in a multiple linear regression model, select Analyze > Regression > Linear.

Move the response variable into the Dependent box and the predictor variables into the Independent(s) box. Before hitting OK, click the Save button and check Leverage values under Distances in the subsequent Linear Regression: Save dialog box. Click Continue to return to the main Linear Regression dialog box, and then hit OK. This results in “centered” leverages (= ordinary leverage $-1/n$, where ordinary leverage is defined in Section 5.1.2 and n is the sample size) being saved as a variable called LEV.1 in the Data Editor Window; they can now be used just like any other variable, for example, to construct scatterplots. Each time you save leverages like this, SPSS will add a new variable to the dataset and increment the end digit by one; for example, the second set of leverages will be called LEV.2.

- 27 To save **Cook’s distances** in a multiple linear regression model, select Analyze > Regression > Linear.

Move the response variable into the Dependent box and the predictor variables into the Independent(s) box. Before hitting OK, click the Save button and check Cook’s under Distances in the subsequent Linear Regression: Save dialog box. Click Continue to return to the main Linear Regression dialog box, and then hit OK. Cook’s distances are saved as a variable called COO.1 in the Data Editor Window; they can now be used just like any other variable, for example, to construct scatterplots. Each time you save Cook’s distances like this, SPSS will add a new variable to the dataset and increment the end digit by one; for example, the second set of Cook’s distances will be called COO.2.

- 28 To create a **residual plot** automatically in a multiple linear regression model, select Analyze > Regression > Linear.

Move the response variable into the Dependent box and the predictor variables into the Independent(s) box. Before hitting OK, click the Plots button and move *SRESID into the Y box and *ZPRED into the X box to create a scatterplot of the studentized residuals on the vertical axis versus the standardized predicted values on the horizontal axis. Click Continue to return to the main Linear Regression dialog box, and then hit OK.

To create residual plots manually, first create studentized residuals (see computer help #24), and then construct scatterplots with these studentized residuals on the vertical axis.

- 29 To create a **correlation matrix** of quantitative variables (useful for checking potential **multicollinearity** problems), select Analyze > Correlate > Bivariate.

Move the variables into the Variables box and hit OK.

- 30 To find **variance inflation factors** in multiple linear regression, select Analyze > Regression > Linear.
- Move the response variable into the Dependent box and the predictor variables into the Independent(s) box. Then click Statistics and check Collinearity diagnostics. Click Continue to return to the Regression dialog and OK to obtain the results. The variance inflation factors are in the last column of the “Coefficients” output under “VIF.”
- 31 To draw a **predictor effect plot** for graphically displaying the effects of transformed quantitative predictors and/or interactions between quantitative and qualitative predictors in multiple linear regression, first create a variable representing the effect, say, “X1effect” (see computer help #3). Then select Graphs > Legacy Dialogs > Interactive > Line.
- Move the “X1effect” variable into the vertical axis box and X1 into the horizontal axis box.
- If the “X1effect” variable just involves X1 (e.g., $1+3X1+4X1^2$), you can hit OK at this point.
- Otherwise, if the “X1effect” variable also involves a qualitative variable (e.g., $1-2X1+3D2X1$, where D2 is an indicator variable), you should move the qualitative variable into the Legend Variables Color or Style box before hitting OK. See Section 5.4 for an example.