

Applied Regression Modeling: A Business Approach

Computer software help: SAS

SAS (originally “Statistical Analysis Software”) is a commercial statistical software package based on a powerful programming interface. It does, however, also have an easy-to-use graphical user-interface in its Analyst Application. Further information is available at www.sas.com. The following instructions are based on the Analyst Application for “SAS 9.1 for Windows.” The book website contains supplementary material for other versions of SAS.

Getting started and summarizing univariate data

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

Start the Analyst Application by selecting Solutions > Analysis > Analyst.

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

Output appears in a separate window each time you run an analysis. Select Edit > Copy to Program Editor to copy the output to a Program Editor Window. From there, output can be copied and pasted from SAS to a word processor like Microsoft Word. Graphs appear in separate windows and can also easily be copied and pasted to other applications. If you misplace any output, you can easily retrieve it by clicking on the Analyst Window and using the left-hand Outline Pane.

- 2 You can access **help** on SAS Analyst by selecting Help > Using This Window or clicking the Analyst Help tool.

For example, to find out about “boxplots” find Box Plots (under Creating Graphs in the main pane of the Help Window).

- 3 To **transform data** or compute a **new variable**, first select Edit > Mode > Edit to change the dataset from “browse” mode to “edit” mode. Then select Data > Transform > Compute.

Type a name (with no spaces) for the new variable in the top-left box, and type a mathematical expression for the variable in the large box just below this. Current variables in the dataset can be moved into this box, while the keypad and list of functions can be used to create the expression. Examples are $\log(X)$ 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 spreadsheet); it can now be used just like any other variable. If you get the error message “Unable to add a new column as specified,” this means there is a syntax error in your 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, first select Edit > Mode > Edit to change the dataset from “browse” mode to “edit” mode. Then select Data > Transform > Recode Values.

Select the qualitative variable in the Column to recode box, type a name for the first indicator variable in the New column name box, make sure New column type is Numeric,

and press OK. In the subsequent Recode Values dialog box, type 1 into the box next to the appropriate level, and type 0 into the boxes next to each of the other levels. Hit OK and check that the correct indicator variable has been added to your spreadsheet. Repeat for other indicator variables (if necessary).

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

Move the variable(s) into the Analysis list. Click Statistics to select the summaries, such as the Mean, that you would like.

- 5 Create **contingency tables** or **cross-tabulations** for qualitative variables by selecting Statistics > Table Analysis.

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

- 6 If you have quantitative variables and qualitative variables, you can calculate **descriptive statistics** for cases grouped in different categories by selecting Statistics > Descriptive > Summary Statistics.

Move the quantitative variable(s) into the Analysis list and the qualitative variable(s) into the Class list. Click Statistics to select the summaries that you would like.

- 7 SAS Analyst does not appear to offer an automatic way to make a stem-and-leaf plot for a quantitative variable.

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

Move the variable into the Analysis box.

- 8 To make a **scatterplot** with two quantitative variables, select Graphs > Scatter Plot > Two-Dimensional.

Move the vertical axis variable into the Y Axis box and the horizontal axis variable into the X Axis box.

SAS Analyst does not appear to offer an automatic way to make a scatterplot matrix.

- 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 Class box in the Scatterplot dialog.

SAS Analyst does not appear to offer an automatic way to change the colors/symbols used or to identify individual cases in a scatterplot.

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

For frequency bar charts of one qualitative variable, move the variable into the Chart box. For frequency bar charts of two qualitative variables, move one variable into the Chart box and the other into the Group By box. The bars can also represent various summary functions for a quantitative variable. For example, to represent Means, select Options, click the Bar Values tab, move the quantitative variable into the Analysis box, and select Average for Statistic to chart.

- 11 To make **boxplots** for cases in different categories, select Graphs > Box Plot.

Move the qualitative variable into the Class box. Move the quantitative variable into the Analysis box.

SAS Analyst does not appear to offer an automatic way to create clustered boxplots for two qualitative variables.
- 12 To make a **QQ-plot** (also known as a **normal probability plot**) for a quantitative variable, select Graphs > Probability Plot.

Move the variable into the Analysis box and leave the Distribution as Normal to assess normality of the variable.
- 13 To compute a **confidence interval** for a univariate population mean, select Statistics > Hypothesis Tests > One-Sample t-test for a Mean.

Move the variable for which you want to calculate the confidence interval into the Variable box and click the Tests button to bring up another dialog box in which you can select Interval under Confidence intervals and specify the confidence level for the interval. OK 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 Statistics > Hypothesis Tests > One-Sample t-test for a Mean.

Move the variable for which you want to do the test into the Variable box and type the (null) hypothesized value into the Mean = box. Specify a lower tailed (“less than”), upper tailed (“greater than”), or two tailed (“not equal”) alternate hypothesis. OK will take you back to the previous dialog box, where you can now hit OK.

Simple linear regression

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

Move the response variable into the Dependent box and the predictor variable into the Explanatory box. Just hit OK for now—the other items in the dialog box are addressed below.
- 16 To include a **regression line** or **least squares line** on a scatterplot, select Statistics > Regression > Simple.

Move the response variable into the Dependent box and the predictor variable into the Explanatory box. Before hitting OK, click the Plots button, and check Plot observed vs independent under Scatterplots. Click OK to return to the main Simple Linear Regression dialog box, and then hit OK. Click on the Analyst Window, and double click on Scatter plot under Simple Linear Regression in the left-hand Outline Pane to find the resulting graph.

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

Move the response variable into the Dependent box and the predictor variable into the Explanatory box. Before hitting OK, click the Statistics button, and check Confidence limits for estimates under Parameter estimates. Click OK to return to the main Simple Linear Regression dialog box, and then hit OK. The confidence intervals are displayed as the final two columns of the “Parameter Estimates” output.

This applies more generally to multiple linear regression also.

- 18 To find a 95% **confidence interval for the mean of Y** at a particular value of X in a simple linear regression model, select Statistics > Regression > Simple.

Move the response variable into the Dependent box and the predictor variable into the Explanatory box. Before hitting OK, click the Save Data button and add L95M and U95M to the empty box in the subsequent Simple Linear Regression: Save Data dialog box. Check the Create and save diagnostics data box, click OK to return to the main Simple Linear Regression dialog box, and then hit OK.

Click on the Analyst Window, and double click on Diagnostics Table under Simple Linear Regression > Diagnostics in the left-hand Outline Pane to find the results. The confidence intervals for the mean of Y at each of the X-values in the dataset are displayed as two columns headed .L95M and .U95M.

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, create a dataset containing (just) the X-value in question (with the same variable name as in the original dataset), and save this dataset. Then fit the regression model and follow the steps above, but before hitting OK, click the Prediction button, click Predict additional data under Prediction input and locate the dataset you just saved under Data set name. Then check List predictions and Add prediction limits under Prediction output. Click OK to return to the main Simple Linear Regression dialog box, and then hit OK. Click on the Analyst Window, and double click on Predictions under Simple Linear Regression in the left-hand Outline Pane to find the results.

This applies more generally to multiple linear regression also.

- 19 To find a 95% **prediction interval** for an individual value of Y at a particular value of X in a simple linear regression model, select Statistics > Regression > Simple.

Move the response variable into the Dependent box and the predictor variable into the Explanatory box. Before hitting OK, click the Save Data button and add L95 and U95 to the empty box in the subsequent Simple Linear Regression: Save Data dialog box. Check the Create and save diagnostics data box, click OK to return to the main Simple Linear Regression dialog box, and then hit OK.

Click on the Analyst Window, and double click on Diagnostics Table under Simple Linear Regression > Diagnostics in the left-hand Outline Pane to find the results. The prediction intervals for an individual value of Y at each of the X-values in the dataset are displayed as two columns headed .L95 and .U95.

This applies more generally to multiple linear regression also.

SAS Analyst does not appear to offer an automatic way to create a prediction interval for an individual Y-value at an X-value that is not in the dataset.

Multiple linear regression

- 20 To fit a **multiple linear regression model**, select Statistics > Regression > Linear.
Move the response variable into the Dependent box and the predictor variables into the Explanatory box.
- 21 To include a **quadratic regression line** on a scatterplot, select Statistics > Regression > Simple.
Move the response variable into the Dependent box and the predictor variable into the Explanatory box, and change Model from Linear to Quadratic. Before hitting OK, click the Plots button, and check Plot observed vs independent under Scatterplots. Click OK to return to the main Simple Linear Regression dialog box, and then hit OK. Click on the Analyst Window, and double click on Scatter plot under Simple Linear Regression in the left-hand Outline Pane to find the resulting graph.
- 22 SAS Analyst does not appear to offer an automatic way to create a scatterplot with separate regression lines for subsets of the sample.
- 23 SAS Analyst 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 SAS Analyst regression output and appropriate percentiles from a F-distribution.
- 24 To save **studentized residuals** in a multiple linear regression model, select Statistics > Regression > Linear.
Move the response variable into the Dependent box and the predictor variables into the Explanatory box. Before hitting OK, click the Save Data button and add STUDENT to the empty box in the subsequent Linear Regression: Save Data dialog box. Check the Create and save diagnostics data box, click OK to return to the main Linear Regression dialog box, and then hit OK.
Click on the Analyst Window, and double click on Diagnostics Table under Linear Regression > Diagnostics in the left-hand Outline Pane to find the results. The studentized residuals are displayed as .STUDENT. (SAS can also calculate “deleted studentized residuals,” which it calls RSTUDENT.)
- 25 SAS Analyst does not appear to offer an automatic way to add a loess fitted line to a scatterplot.
- 26 To save **leverages** in a multiple linear regression model, select Statistics > Regression > Linear.
Move the response variable into the Dependent box and the predictor variables into the Explanatory box. Before hitting OK, click the Save Data button and add H to the empty box in the subsequent Linear Regression: Save Data dialog box. Check the Create and save diagnostics data box, click OK to return to the main Linear Regression dialog box, and then hit OK.
Click on the Analyst Window, and double click on Diagnostics Table under Linear Regression > Diagnostics in the left-hand Outline Pane to find the results. The leverages are displayed as .H.

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

Move the response variable into the Dependent box and the predictor variables into the Explanatory box. Before hitting OK, click the Save Data button and add COOKD to the empty box in the subsequent Linear Regression: Save Data dialog box. Check the Create and save diagnostics data box, click OK to return to the main Linear Regression dialog box, and then hit OK.

Click on the Analyst Window, and double click on Diagnostics Table under Linear Regression > Diagnostics in the left-hand Outline Pane to find the results. Cook's distances are displayed as `._COOKD`.

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

Move the response variable into the Dependent box and the predictor variables into the Explanatory box. Before hitting OK, click the Plots button, and click the Residual tab in the subsequent Linear Regression: Plots dialog box. Check Plot residuals vs variables under Residual plots, and select Standardized for Residuals and Predicted Y for Variables to create a scatterplot of the studentized residuals on the vertical axis versus the standardized predicted values on the horizontal axis. You could also check Independents for Variables to create residual plots with each predictor variable on the horizontal axis. Click OK to return to the main Linear Regression dialog box, and then hit OK. Click on the Analyst Window, and double click on the resulting graphs under Linear Regression > Residual Plots in the left-hand Outline Pane.

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 Statistics > Descriptive > Correlations.

Move the variables into the Correlate box and hit OK.

- 30 To find **variance inflation factors** in multiple linear regression, select Statistics > Regression > Linear.

Move the response variable into the Dependent box and the predictor variables into the Explanatory box. Before hitting OK, click the Statistics button, and the Tests tab in the resulting Linear Regression: Statistics dialog box. Check Variance inflation factors under Collinearity, click OK to return to the main Linear Regression dialog box, and then hit OK. The variance inflation factors are in the last column of the "Parameter Estimates" output under "Variance Inflation."

- 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 select Graphs > Scatter Plot > Two-Dimensional.

Move the "X1effect" variable into the Y Axis box and X1 into the X Axis box. Before hitting OK, click on Display and select Connect points with straight lines in the resulting 2-D Scatter Plot: Display dialog box. Click OK to return to the main 2-D Scatter Plot dialog box, and then hit OK.

SAS Analyst 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).