

Applied Regression Modeling: A Business Approach

Computer software help: SAS JMP

JMP statistical discovery software is a SAS product that “dynamically links statistics with graphics right on your desktop, empowering you to explore data interactively and bring understanding to your world.” Further information is available at www.jmp.com. The following instructions are based on “SAS JMP 7.0 for Windows.”

Getting started and summarizing univariate data

- 1 If desired, change JMP’s default **options** by selecting File > Preferences.

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

You can also use File > Open to open text data files or Excel spreadsheets. For Excel spreadsheets, check the box labeled Always enforce Excel Row 1 as labels if the spreadsheet has the variable labels in the first row.

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

- 3 To **transform data** or compute a **new variable**, select Cols > New Column, type the new variable name in the Column Name box, and select Formula under Column Properties. In the resulting dialog box, select the variable to be transformed under Table Columns and build the formula using the various operations and functions. Examples are Transcendental > Log for the natural logarithm and x^y for powers such as 2 (“squared”).

The new variable should appear in the data spreadsheet (check that it looks correct) and can now be used just like any other variable.

To create **indicator (dummy) variables** from a qualitative variable, select the qualitative variable and select Cols > Recode.

Type the values 0 and 1 under New Value for the appropriate categories and change In Place to New Column. Check that the correct indicator variable has been created in the spreadsheet. Change the name and data/modeling type of the created variable by double-clicking the column heading (Data Type should be Numeric rather than Character and Modeling Type should be Continuous rather than Nominal). Repeat for other indicator variables (if necessary).

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

Analyze > Distribution.

Move the variable(s) into the Y, Columns list and click OK. In the resulting output window, you can select additional output by clicking on the red triangle next to each variable name.

- 5 Create **contingency tables** or **cross-tabulations** for qualitative variables by selecting Analyze > Fit Y by X.

Move one qualitative variable into the Y, Response list and another into the X, Factor list. Cell percentages (within rows, columns, or the whole table) are displayed automatically in the resulting table.

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

Select the quantitative variable(s) and then select the summaries that you would like from the Statistics menu. Move the qualitative variable(s) into the Group list.

- 7 To make a **stem-and-leaf plot** for a quantitative variable, select Analyze > Distribution. Move the variable(s) into the Y, Columns list and click OK. In the resulting output window, you can select Stem and Leaf by clicking on the red triangle next to each variable name.

To make a **histogram** for a quantitative variable, select Analyze > Distribution.

Move the variable(s) into the Y, Columns list and click OK. In the resulting output window, you can select various Histogram Options by clicking on the red triangle next to each variable name.

- 8 To make a **scatterplot** with two quantitative variables, select Analyze > Fit Y by X.

Move the vertical axis variable into the Y, Response box and the horizontal axis variable into the X, Factor box.

All possible scatterplots for more than two variables can be drawn simultaneously (called a **scatterplot matrix**) by selecting Graph > Scatterplot Matrix.

Move all the variables into the Y, Columns box.

- 9 You can **mark or label cases** in a scatterplot with different colors/symbols according to the categories in a qualitative variable by selecting Rows > Color or Mark by Column... before drawing the plot.

Select the column containing the variable you wish to mark by.

You can also **identify individual cases** in a scatterplot by hovering over individual points in the scatterplot. If you double-click a point, the corresponding row in the spreadsheet will be highlighted.

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

For frequency bar charts of one or two qualitative variables, move the variable(s) into the Categories, X, Levels box. The bars can also represent various summary functions for a quantitative variable. For example, to represent group means, select the quantitative variable and then select Mean from the Statistics menu.

- 11 To make **boxplots** for cases in different categories, select Analyze > Fit Y by X.

Move the quantitative variable into the Y, Response box and the qualitative variable into the X, Factor box.

In the resulting Oneway Analysis output window, click on the red triangle and select Quantiles.

To create clustered boxplots for two qualitative variables, first create a new qualitative variable consisting of all category combinations (using computer help #3 and the Character > Concat function). Then use this new variable as the X, Factor variable.

- 12 To make a **QQ-plot** (also known as a **normal probability plot**) for a quantitative variable, select Analyze > Distribution.
- Move the variable into the Y, Columns list and click OK. In the resulting output window, you can select Normal Quantile Plot by clicking on the red triangle next to the variable name.
- 13 To compute a **confidence interval** for a univariate population mean, select Analyze > Distribution.
- Move the variable into the Y, Columns list and click OK. In the resulting output window, you can select Confidence Interval by clicking on the red triangle next to the variable name. Enter the confidence level in the resulting Confidence Intervals dialog box and click OK.
- 14 To do a **hypothesis test** for a univariate population mean, select Analyze > Distribution.
- Move the variable into the Y, Columns list and click OK. In the resulting output window, you can select Test Mean by clicking on the red triangle next to the variable name. Enter the (null) hypothesized mean in the resulting Test Mean dialog box and click OK.

Simple linear regression

- 15 To fit a **simple linear regression model** (i.e., find a least squares line), select Analyze > Fit Model.
- Move the response variable into the Y box, select the predictor variable and Add it to the Construct Model Effects box, and click Run Model.
- 16 To add a **regression line** or **least squares line** to a scatterplot, select Analyze > Fit Y by X.
- Move the response variable into the Y, Response box, move the predictor variable into the X, Factor box, and click OK. Click on the red triangle in the resulting Fit Y by X output window, and select Fit Line.
- 17 To find **95% confidence intervals for the regression parameters** in a simple or multiple linear regression model, fit the model using computer help #15 or #20, right-click in the body of the Parameter Estimates table in the resulting Fit Least Squares output window, and select Columns > Lower 95% and Columns > Upper 95%.
- 18 To find a **confidence interval for the mean of Y** at a particular value of X in a simple linear regression model, fit the model using computer help #15 or #20, click on the red triangle next to Response in the resulting Fit Least Squares output window, and select Save Columns > Mean Confidence Interval.
- This will produce 95% intervals for each of the X-values in the dataset by default (in columns labeled “Lower 95% Mean Y” and “Upper 95% Mean Y”). Each time you ask JMP to calculate confidence intervals like this it will add new columns to the dataset and append a number to the column headers (e.g., “2” for the second time). If you hold down the Shift key and then select Save Columns > Mean Confidence Interval you’ll be prompted to enter a significance level (e.g., enter 0.10 for 90% intervals).

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 (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. JMP will ignore the X-value you types 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 JMP 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 in the columns labeled “Lower 95% Mean Y” and “Upper 95% Mean Y.”

This applies more generally to multiple linear regression also.

- 19 To find a **prediction interval** for an individual Y-value at a particular X-value in a simple linear regression model, fit the model using computer help #15 or #20, click on the red triangle next to Response in the resulting Fit Least Squares output window, and select Save Columns > Indiv Confidence Interval.

This will produce 95% intervals for each of the X-values in the dataset by default (in columns labeled “Lower 95% Indiv Y” and “Upper 95% Indiv Y”). Each time you ask JMP to calculate confidence intervals like this it will add new columns to the dataset and append a number to the column headers (e.g., “2” for the second time). If you hold down the Shift key and then select Save Columns > Indiv Confidence Interval you’ll be prompted to enter a significance level (e.g., enter 0.10 for 90% intervals).

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 (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. JMP will ignore the X-value you types 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 JMP 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 in the columns labeled “Lower 95% Indiv Y” and “Upper 95% Indiv Y.”

This applies more generally to multiple linear regression also.

Multiple linear regression

- 20 To fit a **multiple linear regression model**, select Analyze > Fit Model.

Move the response variable into the Y box, select the predictor variables and Add them to the Construct Model Effects box, and click Run Model.

- 21 To add a **quadratic regression line** to a scatterplot, select Analyze > Fit Y by X.

Move the response variable into the Y, Response box, move the predictor variable into the X, Factor box, and click OK. Click on the red triangle in the resulting Fit Y by X output window, and select Fit Polynomial > 2,quadratic.

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 according to the categories in the qualitative predictor variable. To add a **regression line for each subset** to this scatterplot, first click on the red triangle in the resulting Fit Y by X output window, select Group By ..., select the qualitative predictor variable, and click OK. Then click on the red triangle again and select Fit Line.

23 To find the F-statistic and associated p-value for a **nested model F-test** in multiple linear regression, fit the model using computer help #20, click on the red triangle next to Response in the resulting Fit Least Squares output window, and select Custom Test....

The resulting Custom Test output will have a list of regression parameters that has a column of zeroes next to it; click the zero next to the first parameter in the nested F-test null hypothesis and change the value to "1." Then click Add Column and repeat for the second parameter in the null hypothesis. Repeat for each of the parameters in the null hypothesis, then click Done.

24 To save **studentized residuals** in a multiple linear regression model, fit the model using computer help #20, click on the red triangle next to Response in the resulting Fit Least Squares output window, and select Save Columns > Studentized Residuals.

The studentized residuals are saved as a variable called Studentized Resid *, where the star represents the response variable name; they can now be used just like any other variable, for example, to construct residual plots.

25 JMP does not appear to offer a way to add a **loess fitted line** to a scatterplot, but it can add a similar **smoothing spline fitted line** (useful for checking the zero mean regression assumption in a residual plot). To do so, select Analyze > Fit Y by X.

Move the vertical axis variable (e.g., the studentized residuals) into the Y, Response box, move the horizontal axis variable into the X, Factor box, and click OK. Click on the red triangle in the resulting Fit Y by X output window, and select Fit Spline; you can experiment to find a value for the smoothing parameter "lambda" that captures the major trends in the scatterplot without being overly "wiggly," but typically a value of 1 or 10 should work well.

26 To save **leverages** in a multiple linear regression model, fit the model using computer help #20, click on the red triangle next to Response in the resulting Fit Least Squares output window, and select Save Columns > Hats.

The leverages are saved as a variable called h *, where the star represents the response variable name; they can now be used just like any other variable, for example, to construct scatterplots.

- 27 To save **Cook's distances** in a multiple linear regression model, fit the model using computer help #20, click on the red triangle next to Response in the resulting Fit Least Squares output window, and select Save Columns > Cook's D Influence.

The Cook's distances are saved as a variable called Cook's D Influence *, where the star represents the response variable name; they can now be used just like any other variable, for example, to construct scatterplots.

- 28 JMP will automatically create a **residual plot** in a multiple linear regression model, specifically one with the (ordinary) residuals on the vertical axis versus the predicted values on the horizontal axis.

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 > Multivariate Methods > Multivariate.

Move all the variables into the Y, Columns box and hit OK.

- 30 To find **variance inflation factors** in multiple linear regression, fit the model using computer help #20, right-click in the body of the Parameter Estimates table in the resulting Fit Least Squares output window, and select Columns > 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).

If the "X1effect" variable just involves X1 (e.g., $1+3X1+4X1^2$), then use computer help #16 to create the line plot.

Otherwise, if the "X1effect" variable also involves a qualitative variable (e.g., $1-2X1+3D2X1$, where D2 is an indicator variable), you should then use computer help #22 to create the line plot.