Applied Regression Modeling: A Business Approach

Computer software help: Minitab

Minitab is a commercial statistical software package with an easy-to-use graphical user-interface. Further information is available at `www.minitab.com`. The following instructions are based on "Minitab Release 14 for Windows." The book website contains supplementary material for other versions of Minitab.

Getting started and summarizing univariate data

1 If desired, change Minitab's default **options** by selecting Tools > Options.

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

To **edit last dialog** box, select Edit > Edit Last Dialog or hit the Edit Last Dialog tool (ninth button from the left).

Output appears in the Session Window and can be copied and pasted from Minitab to a word processor like Microsoft Word. Graphs appear in separate windows and can also easily be copied and pasted to other applications.

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

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 Calc > Calculator.

Type a name (with no spaces) for the new variable in the Store result in variable box, and type a mathematical expression for the variable in the Expression box. Current variables in the dataset can be moved into the Expression box, while the keypad and list of functions can be used to create the expression. Examples are LOGE('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 Worksheet Window); it can now be used just like any other variable. If you get the error message "Completion of computation impossible," 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, select Calc > Make Indicator Variables.

Move the qualitative variable into the Indicator variables for box, type a range of columns in which to store the variables (e.g., C5-C6) in the Store results in box, and hit OK (check that the correct indicator variables have been added to your spreadsheet in the Worksheet Window).

4 Calculate **descriptive statistics** for quantitative variables by selecting Stat > Basic Statistics > Display Descriptive Statistics.

Move the variable(s) into the Variables 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 Stat > Tables > Cross Tabulation and Chi-Square.

Move one qualitative variable into the rows box and another into the columns box. Cell percentages (within rows, columns, or the whole table) can be calculated by checking the appropriate boxes under Display.

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

Move the qualitative variable into the rows box (and another qualitative variable into the columns box if there is more than one). Click Associated Variables to select the quantitative variable for which you would like descriptive statistics, and the descriptive statistics to display; the default is the number of cases, but other statistics such as the Mean and Standard Deviation can also be selected.

7 To make a **stem-and-leaf plot** for a quantitative variable, select Graph > Stem-and-Leaf.

Move the variable into the Graph variables box.

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

Choose Simple and move the variable into the Graph variables box.

8 To make a **scatterplot** with two quantitative variables, select Graph > Scatterplot.

Choose Simple and move the vertical axis variable into the first row of the Y variables column and the horizontal axis variable into the first row of the X variables column.

All possible scatterplots for more than two variables can be drawn simultaneously (called a **scatterplot matrix**) by selecting Graph > Matrix Plot, choosing Matrix of plots, Simple, and moving the variables into the Graph 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 selecting Graph > Scatterplot and choosing With Groups. After moving the vertical axis variable into the first row of the Y variables column and the horizontal axis variable into the first row of the X variables column, move the grouping variable into the Categorical variables for grouping box.

To change the colors/symbols used, select the symbols you want to change by clicking on one of the points with that symbol twice (all the data points should become highlighted on the first click, and just the points in that group should remain highlighted on the second click). Then select Editor > Edit Symbols.

Select the color/symbol you want and hit OK to see the effect.

You can also **identify individual cases** in a scatterplot by hovering over them.

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

For frequency bar charts of one qualitative variable, choose Simple with Bars represent: Counts of unique values and move the variable into the Categorical variables box. For frequency bar charts of two qualitative variables, choose Cluster with Bars represent: Counts of unique values and move the variables into the Categorical variables box. The bars can also represent various summary functions for a quantitative variable; for example, to represent means, select Bars represent: A function of a variable and select Mean for the function.

11 To make **boxplots** for cases in different categories, select Graph > Boxplot.

Choose One Y, With Groups, move the quantitative variable into the Graph variables box, and move the qualitative variable(s) into the Categorical variables box.

12 To make a **QQ-plot** (also known as a **normal probability plot**) for a quantitative variable, select Graph > Probability Plot.

Choose Single and move the variable into the Graph variables box.

13 To compute a **confidence interval** for a univariate population mean, select Stat > Basic Statistics > 1-Sample t.

Move the variable for which you want to calculate the confidence interval into the Samples in columns box. Then click the Options button to bring up another dialog box in which you can 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 Stat > Basic Statistics > 1-Sample t.

Move the variable for which you want to calculate the confidence interval into the Samples in columns box, check Perform hypothesis test, and type the (null) hypothesized value into the Hypothesized mean box. Then click the Options button to bring up another dialog box in which you can specify a lower tailed ("less than"), upper tailed ("greater than"), or two tailed ("not equal") alternative 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 Stat > Regression > Regression.

Move the response variable into the Response box and the predictor variable into the Predictors box. 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 Editor > Add > Regression Fit, and Linear for the Model Order.

You can create a scatterplot with a regression line superimposed by selecting Graph > Scatterplot.

Choose With Regression and move the response variable into the first row of the Y variables column and the predictor variable into the first row of the X variables column.

17 Minitab does not appear to offer an automatic way to find 95% confidence intervals for the regression parameters in a linear regression model.

It is possible to calculate these intervals by hand using Minitab regression output and appropriate percentiles from a t-distribution.

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

Move the response variable into the Response box and the predictor variable into the Predictors box. Before hitting OK, click the Options button and move the predictor variable into the Prediction intervals for new observations box in the subsequent Regression: Options dialog box. Check Confidence limits and type the value of the confidence level that you want in the Confidence level box (the default is 95%). Click OK to return to the main 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 in the Session Window and also as two columns headed CLIM1 and CLIM2 in the Worksheet Window. Each time you ask Minitab 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 CLIM3 and CLIM4.

You can also obtain a confidence interval for the mean of Y at an X-value that is not in the dataset by typing the value into the Prediction intervals for new observations box in the Regression: Options dialog box. In this case, the confidence interval for the mean of Y at this X-value is displayed only in the Session Window (and not in the Worksheet 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 Stat > Regression > Regression.

Move the response variable into the Response box and the predictor variable into the Predictors box. Before hitting OK, click the Options button and move the predictor variable into the Prediction intervals for new observations box in the subsequent Regression: Options dialog box. Check Prediction limits and type the value of the confidence level that you want in the Confidence level box (the default is 95%). Click OK to return to the main 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 in the Session Window and also as two columns headed PLIM1 and PLIM2 in the Worksheet Window. Each time you ask Minitab 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 Y-value the end points will be called PLIM3 and PLIM4.

You can also obtain a prediction interval for an individual Y-value at an X-value that is not in the dataset by typing the value into the Prediction intervals for new observations box in the Regression: Options dialog box. In this case, the prediction interval for an individual Y-value at this X-value is displayed only in the Session Window (and not in the Worksheet Window).

This applies more generally to multiple linear regression also.

Multiple linear regression

20 To fit a **multiple linear regression model**, select Stat > Regression > Regression.

Move the response variable into the Response box and the predictor variables into the Predictors box.

21 To add a **quadratic regression line** to a scatterplot, select Editor > Add > Regression Fit, and Quadratic for the Model Order.

You can create a scatterplot with a quadratic regression line superimposed by selecting Graph > Scatterplot.

Choose With Regression and move the vertical axis variable into the first row of the Y variables column and the horizontal axis variable into the first row of the X variables column. Before hitting OK, click the Data View button, click the Regression tab in the subsequent Scatterplot - Data View dialog box, and change the Model Order from Linear to Quadratic. Hit OK to return to the Scatterplot - With Regression dialog box, and OK again to create the graph.

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. To display a **regression line for each subset** in a scatterplot, select Graph > Scatterplot and choose With Regression and Groups. After moving the vertical axis variable into the first row of the Y variables column and the horizontal axis variable into the first row of the X variables column, move the grouping variable into the Categorical variables for grouping box. Hit OK to create the graph.

23 Minitab 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 Minitab regression output and appropriate percentiles from a F-distribution.

24 To save **studentized residuals** in a multiple linear regression model, select Stat > Regression > Regression.

Move the response variable into the Response box and the predictor variables into the Predictors box. Before hitting OK, click the Storage button and check Standardized residuals under Diagnostic Measures in the subsequent Regression: Storage dialog box. Click OK to return to the main Regression dialog box, and then hit OK. The studentized residuals are saved as a variable called SRES1 in the Worksheet Window; they can now be used just like any other variable, for example, to construct residual plots. Each time you ask Minitab 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 SRES2.

25 To add a **loess fitted line** to a scatterplot (useful for checking the zero mean regression assumption in a residual plot), select Editor > Add > Smoother.

The default value of 0.5 for Degree of smoothing tends to be a little on the low side: I would change it to 0.75.

You can create a scatterplot with a loess fitted line superimposed by selecting Graph > Scatterplot.

Choose With Regression and move the vertical axis variable into the first row of the Y variables column and the horizontal axis variable into the first row of the X variables column. Before hitting OK, click the Data View button, click the Smoother tab in the subsequent Scatterplot - Data View dialog box, and change the Smoother from None to Lowess. Hit OK to return to the Scatterplot - With Regression dialog box, and OK again to create the graph.

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

Move the response variable into the Response box and the predictor variables into the Predictors box. Before hitting OK, click the Storage button and check Hi (leverages) under Diagnostic Measures in the subsequent Regression: Storage dialog box. Click OK to return to the main Regression dialog box, and then hit OK. The leverages are saved as a variable called HI1 in the Worksheet Window; they can now be used just like any other variable, for example, to construct scatterplots. Each time you ask Minitab to save leverages 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 leverages they will be called HI2.

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

Move the response variable into the Response box and the predictor variables into the Predictors box. Before hitting OK, click the Storage button and check Cook's distance under Diagnostic Measures in the subsequent Regression: Storage dialog box. Click OK to return to the main Regression dialog box, and then hit OK. Cook's distances are saved as a variable called COOK1 in the Worksheet Window; they can now be used just like any other variable, for example, to construct scatterplots. Each time you ask Minitab to save Cook's distances 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 Cooks' distances they will be called COOK2.

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

Move the response variable into the Response box and the predictor variables into the Predictors box. Before hitting OK, click the Graphs button and select Standardized under Residuals for Plots under Standardized Residual Plots in the subsequent Regression - Graphs dialog box. Check Residuals versus fits under Individual plots to create a scatterplot of the studentized residuals on the vertical axis versus the standardized predicted values on the horizontal axis. You could also move individual predictor variables into the Residuals versus the variables box to create residual plots with each predictor variable on the horizontal axis. Click OK to return to the main 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 Stat > Basic Statistics > Correlation.

Move the variables into the Variables box and hit OK.

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

Move the response variable into the Response box and the predictor variables into the Predictors box. Before hitting OK, click the Options button and check Variance inflation factors under Display in the subsequent Regression - Options dialog box. Click OK to return to the main Regression dialog box, and then hit OK. The variance inflation factors are in the last column of the main regression 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 Graph > Scatterplot.

Choose With Connect and Groups and move the "X1effect" variable into the first row of the Y variables column and X1 into the first row of the X variables column.

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 Categorical variables for grouping box before hitting OK. See Section 5.4 for an example.