

# Applied Regression Modeling: A Business Approach

## Computer software help: Data Desk

Data Desk is a data analysis and visualization program. It combines traditional statistical procedures with interactive exploratory tools to help users find patterns, relationships and exceptions in their data. Further information is available at [www.datadesk.com](http://www.datadesk.com). The following instructions are based on “Data Desk 6.1 for Windows.” Instructions for the Macintosh version should be broadly similar.

### Getting started and summarizing univariate data

- 1 If desired, change Data Desk’s default **options** by selecting Edit > Preferences.

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

You can also use File > Import to open text data files or use Data Desk/XL, an Excel add-in, to export Excel spreadsheets to Data Desk.

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

- 3 To **transform data** or compute a **new variable**, select the variable you want to transform (denoted  $Y$ ) and select Manip > Transform and the required transformation. Examples are Exponentials >  $\ln(y)$  for the natural logarithm of  $Y$  and Exponentials >  $y^2$  for  $Y^2$ .

Alternatively, select Manip > Transform > New Derived Variable, name the new derived variable, hit OK, and in the resulting text window type the equation for the new variable (this is particularly useful if the variable is a function of more than one of the existing variables).

The new variable should now appear in the same icon window as the original variable, and have an appropriate name, e.g.,  $LY$  for the natural logarithm of  $Y$  (check it looks correct by showing the numbers); it can now be used just like any other variable.

To create **indicator (dummy) variables** from a qualitative variable, select Manip > Transform > New Derived Variable.

Name the new derived variable, hit OK, and in the resulting text window write:

```
If TextOf('var') = "cat1" then 1 else 0
```

where *var* is the name of the qualitative variable and *cat1* is the category for which you want the indicator variable to have the value 1. Check that the correct indicator variable has been created by showing the numbers. Repeat for other indicator variables (if necessary).

- 4 Calculate **descriptive statistics** for quantitative variables by selecting the quantitative variable (denoted  $Y$ ) and selecting Calc > Summaries > Reports.

Use the HyperView menu (top-left triangle) to select the summaries, such as the Mean, that you would like.

- 5 Create **contingency tables** or **cross-tabulations** for qualitative variables by selecting the first qualitative variable (denoted  $Y$ , representing the row categories), shift-selecting the second qualitative variable (denoted  $X$ , representing the column categories), and selecting Calc > Contingency Tables.

Use the HyperView menu to calculate cell percentages (within rows, columns, or the whole table).

- 6 If you have quantitative variables and qualitative variables, you can calculate **descriptive statistics** for cases grouped in different categories by selecting the quantitative variable (denoted  $Y$ ), shift-selecting the qualitative variable (denoted  $X$ ), and selecting Calc > Summaries > Reports By Groups.

Use the HyperView menu to select the summaries, such as the Mean, that you would like.

If you want to group using two qualitative variables, first create a new variable consisting of all category combinations by selecting the two qualitative variables (one can be  $Y$ , the other  $X$ , it does not matter which) and selecting

Manip > Transform > Misc > Concatenate( $y,x$ ).

Then use the new variable as the qualitative variable ( $X$ ) in the previous instructions.

- 7 Data Desk 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 the quantitative variable (denoted  $Y$ ) and select Plot > Histograms.

- 8 To make a **scatterplot** with two quantitative variables, select the vertical axis quantitative variable (denoted  $Y$ ), shift-select the horizontal-axis quantitative variable (denoted  $X$ ), and select Plot > Scatterplots.

All possible scatterplots for more than two variables can be drawn simultaneously (called a **scatterplot matrix**) by selecting the variables you want plotted (it does not matter which are denoted  $Y$  or  $X$ ) and selecting Plot > Plot Matrix.

- 9 You can **mark or label cases** in a scatterplot with different colors/symbols according to the categories in a qualitative variable by selecting the qualitative variable and selecting Modify > Colors > Add > by Group Or Modify > Symbols > Add > by Group.

You can also **identify individual cases** in a scatterplot using labels by opening the variable window containing the labels and selecting the Query tool from the tools palette (fourth one down in the right column). You can then click on a point in the scatterplot and the label for that point will be displayed.

- 10 To make a **bar chart** for cases in different categories, select the qualitative variable that represents the different categories and select Plot > Bar Charts.

This will produce a frequency bar chart of the qualitative variable. For frequency bar charts of two qualitative variables use a newly created qualitative variable consisting of all category combinations (as in computer help #6).

Data Desk does not appear to offer an automatic way to have the bars represent summary functions for a quantitative variable, such as the mean.

- 11 To make **boxplots** for cases in different categories, select the quantitative variable (denoted  $Y$ ), shift-select the qualitative variable (denoted  $X$ ), and select Plot > Boxplot  $y$  by  $x$ .  
For two qualitative variables, use a newly created qualitative variable consisting of all category combinations (as in computer help #6).
- 12 To make a **QQ-plot** (also known as a **normal probability plot**) for a quantitative variable, select the quantitative variable (denoted  $Y$ ) and select Plot > Normal Prob Plot.
- 13 To compute a **confidence interval** for a univariate population mean, select the quantitative variable (denoted  $Y$ ) and select Calc > Estimate.  
In the resulting window, select t-Interval for Individual  $\mu$ 's, select Individual (rather than Total), specify the confidence level for the interval, and hit Show Results.
- 14 To do a **hypothesis test** for a univariate population mean, select the quantitative variable (denoted  $Y$ ) and select Calc > Test.  
In the resulting window, select t-Test of Individual  $\mu$ 's, select Individual (rather than Total), specify the significance level (Alpha level) for the test, type the (null) hypothesized value into the " $H_0:\mu =$ " box, select the alternative hypothesis ( $H_a$ ) to be lower-tail (" $\mu <$ "), two-tail (" $\mu \neq$ "), or upper-tail (" $\mu >$ "), and hit Show Results.

## Simple linear regression

- 15 To fit a **simple linear regression model** (i.e., find a least squares line), select the response variable (denoted  $Y$ ), shift-select the predictor variable (denoted  $X$ ), and select Calc > Regression.  
Some of the items in the HyperView menu are addressed below.
- 16 To add a **regression line** or **least squares line** to a scatterplot, select Add Regression Line from the scatterplot's HyperView menu.
- 17 Data Desk does not appear to offer an automatic way to find **95% confidence intervals for the regression parameters** in a simple or multiple linear regression model.  
It is possible to calculate these intervals by hand using Data Desk regression output and appropriate percentiles from a t-distribution.
- 18 Data Desk 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 simple linear regression model.  
It is possible to calculate such an interval by hand using Data Desk regression output and an appropriate percentile from a t-distribution.  
**This applies more generally to multiple linear regression also.**
- 19 Data Desk does not appear to offer an automatic way to find a **prediction interval** for an individual  $Y$ -value at a particular  $X$ -value in a simple linear regression model.  
It is possible to calculate such an interval by hand using Data Desk regression output and an appropriate percentile from a t-distribution.  
**This applies more generally to multiple linear regression also.**

## Multiple linear regression

- 20 To fit a **multiple linear regression model**, select the response variable (denoted  $Y$ ), shift-select the predictor variables (denoted  $X$ ), and select Calc > Regression.

Some of the items in the HyperView menu are addressed below.

- 21 Data Desk does not appear to offer an automatic way to to add a **quadratic regression line** to a scatterplot.
- 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, select Add Color Regression Lines from the HyperView menu.
- 23 Data Desk does not appear to offer an automatic way to 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 Data Desk regression output and appropriate percentiles from a F-distribution.

- 24 To save **studentized residuals** in a multiple linear regression model, select Compute > IStudRes from the regression's HyperView menu.

The studentized residuals are saved as a variable called IStudRes(\*), where the star abbreviates the response variable name; they can now be used just like any other variable, for example, to construct residual plots. (Data Desk can also calculate "deleted studentized residuals," which it calls EStudRes.)

- 25 To add a **loess fitted line** to a scatterplot (useful for checking the zero mean regression assumption in a residual plot), select Smoothing > Add Lowess Smooth from the scatterplot's HyperView menu.

Select Smoothing > Smoothing Options to change the value of the Lowess Span %; you can experiment to find a value that captures the major trends in the scatterplot without being overly "wiggly."

- 26 To save **leverages** in a multiple linear regression model, select Compute > Leverages from the regression's HyperView menu.

The leverages are saved as a variable called leverages(\*), where the star abbreviates 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, select Compute > Cook from the regression's HyperView menu.

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

- 28 To create a **residual plot** automatically in a multiple linear regression model, select Scatterplot studentized residual vs predicted from the regression's HyperView menu.
- This will create a scatterplot of the externally studentized 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 the variables (it does not matter which are denoted  $Y$  or  $X$ ) and select Calc > Correlations > Pearson Product-Moment.
- 30 Data Desk does not appear to offer an automatic way to find **variance inflation factors** in multiple linear regression.
- 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 the "X1 effect" variable (denoted  $Y$ ), shift-select the X1 variable (denoted  $X$ ), and select Plot > Scatterplots.
- If the "X1effect" variable just involves X1 (e.g.,  $1+3X1+4X1^2$ ), the resulting plot should be fine, albeit the effect will be represented by points rather than a line (as in Section 5.4). If you would prefer a line, select Add Regression Line from the scatterplot's HyperView menu (as in computer help #16).
- Otherwise, if the "X1effect" variable also involves a qualitative variable (e.g.,  $1-2X1+3D2X1$ , where D2 is an indicator variable), you should then select the qualitative variable and select Modify > Colors > Add > by Group (as in computer help #9) and finally select Add Color Regression Lines from the scatterplot's HyperView menu (as in computer help #22). See Section 5.4 for an example.