**R: Statistical Analysis**

**7.2.2 Overview of R Demo**

The following instructions for installing R and RStudio are for your own computer. You can use remote.whitman.syr.edu to access the server version of R and RStudio.

**Installation of R**

R is a free downloadable package capable of performing sophisticated statistical analysis and data mining. The software is already installed on the classroom laptops. To install on your own personal computer:

1. Go to the website: <https://cran.r-project.org/>.
2. Click on the Download link for your operating system.
   1. For Mac Users, click on R for (Mac) OS X. Note that if you have X11, you must install XQuartz, since it is no longer part of OS X. See the directions for on the page displayed after clicking on (Mac) OS X.
   2. For Windows users, click on R for Windows, then click on “install R for the first time.”
3. Click on Run, and follow the install instructions.

**Installation of RStudio**

RStudio is free downloadable cross-platform development tool for R. It assists in writing and running code in R. RStudio requires a 64-bit operating system and the 64-bit version of R.

1. Go to the website <https://rstudio.com/>.
2. Click on Download RStudio.
3. Under Choose Your Version, click on the Download button below RStudio Desktop Free.
4. Follow the instructions to download RStudio to your computer.
   1. If you have a Mac, click on macOS 10.12+ (64-bit).
   2. If you have a Windows PC, click on Windows 10/8/7 (64-bit).
   3. If you have a 32-bit computer, follow the instructions at the top of the page and click on “older version of RStudio.”

**Remote Access**

To run R and RStudio on the server:

1. Go to rds.syr.edu.
2. In the upper right corner click on Settings (picture of gear), then Download the rdp file.
3. Click on Whitman Remote Desktop in the upper left corner; this will download the Whitman Remote Desktop.rdp file.
4. Click on the Whitman Remote Desktop.rdp file in the lower left corner.
5. When the Whitman Remote Desktop Connection window opens, click Connect.
6. The login screen will appear; enter your Syracuse University email address and Syracuse University password.
7. After several seconds to a minute, your rds desktop will appear.

**R versus RStudio**

R is a command line system where you can enter R commands. RStudio is a programmer development environment for the R language. Not all programs used in this course work in RStudio, so please use R.

**Starting R (Use This If You Are Not Familiar with RStudio)**

1. Click on the Start button in the lower left corner of Windows.
2. Click on All Programs, then click on the R folder, then the latest version of R.
3. This is the command line screen.

**Starting RStudio (Alternative to Using R) (Not Everything in This Course Works in RStudio)**

1. Click on the Start button in the lower left corner of Windows.
2. Click on All Programs, then click on the RStudio folder, then RStudio.
3. This is the command line screen. You can enter commands but need to know the syntax.

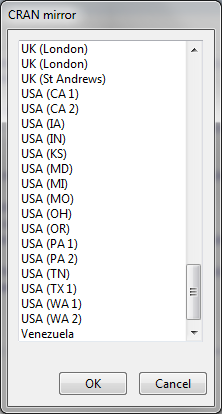
**Installing R Commander**

Follow these steps only if you do not already have Rcmdr installed.

1. In R, type the command:

install.packages("Rcmdr", dependencies = TRUE)

1. In the CRAN mirror, select the location closest to you; use a USA location near you, then click OK.
2. If prompted to create a personal library, click Yes.
3. If prompted to add missing packages, click Yes.



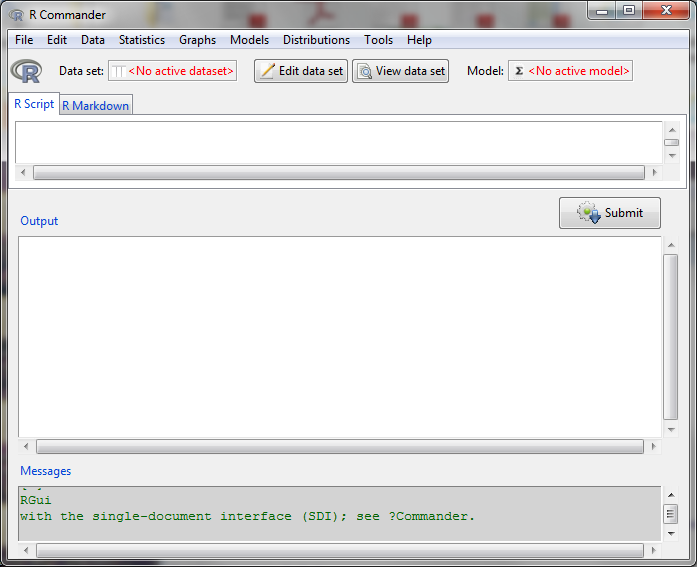
**Launch Rcmdr (R Commander)**

Rcmdr is a graphical user interface (GUI) that is easier to use than the command line. To launch Rcmdr:

1. Type in the command:

library(Rcmdr)

1. If you receive a warning message that some packages are missing, it will ask if you want them installed. Click Yes.
2. On the Install Missing Packages screen, click OK.
3. R will install the necessary software.
4. The R Commander screen will appear.



**7.2.3 R: Loading and Viewing Data**

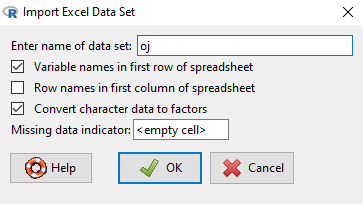
To access this and other excellent data sets used in the book *Data Mining and Business Analytics* with R, by Johannes Ledolter:

1. Download from the course platform the file “Week7-data.xlsx.”

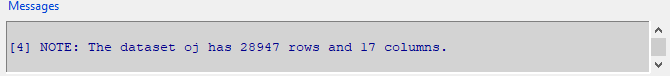
**Loading Data**

To load text data or .csv files (comma separated values files) into R:

1. Click on Data at the top of the Rcmdr screen
2. Click on Import Data > From Excel file …
3. Enter the name that you would like to use for this data set; type in oj (stands for orange juice), then click OK



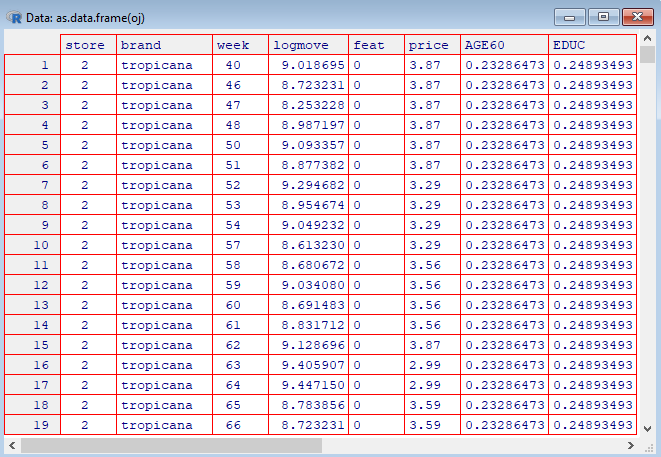
1. Click on the Week7-data file, then Open
2. Note that the dataset oj has 28,947 rows and 17 columns



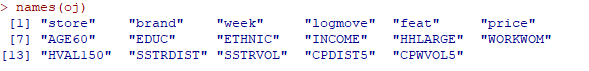
**Viewing Data Fields**

This data set lists weekly sales over 83 stores for three brands of products.

1. Let’s view the data. The variable logmove is the logarithm of product moved (the number of products sold in a week). Click on the button View data set in Rcmdr.



1. To view a list of the variables in R, click on Data, Active Data Set, Variables in Active Data Set.



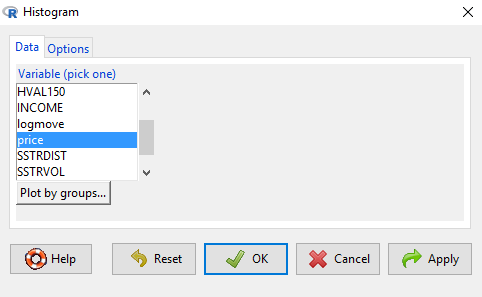
1. Notice that R generates the command names(oj). This is the command line version.

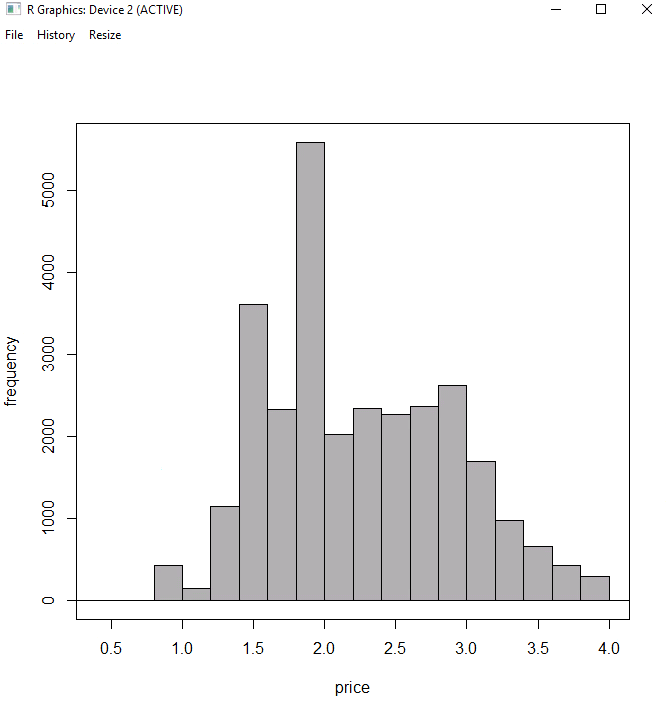
**7.3.2 R: Histograms, Box Plots, Scatterplots, Mean Plots, XY Plots Demo**

**Histograms**

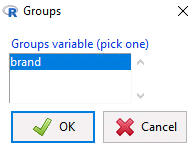
To create a histogram:

1. Click on Graphs, Histogram.
2. Click on the variable price, OK.

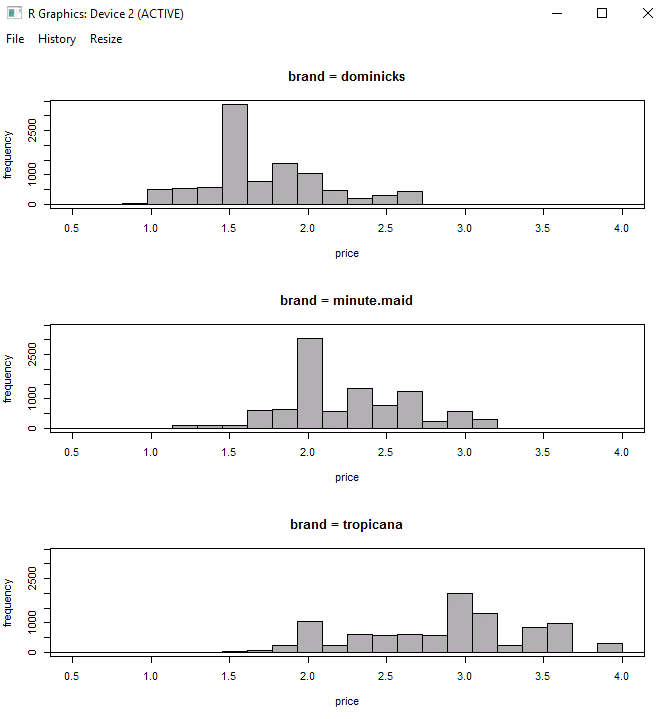




1. Next, plot by groups. Click on Graphs, Histogram, Plot by groups.



1. Click on brand, then OK.
2. Click on OK to generate the histograms.

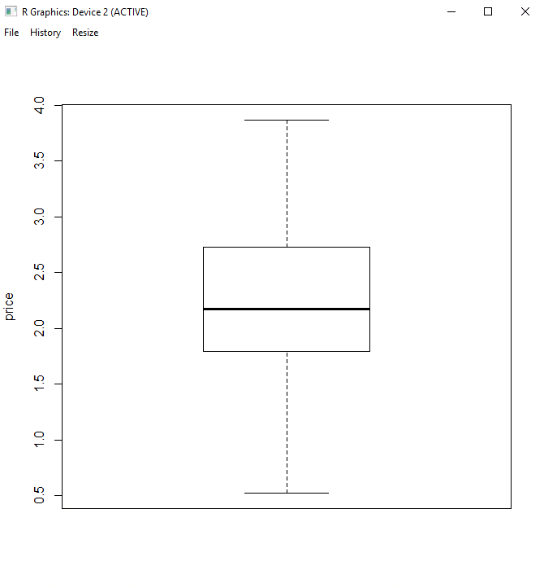


1. Which brand is the premium brand?

**Box Plots**

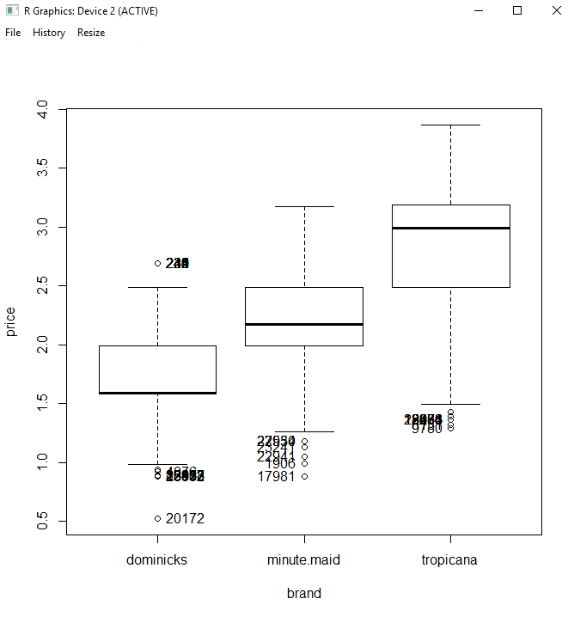
To create a box plot:

1. Click on Graphs, Boxplot, price, OK.
   1. The upper line is the maximum, up to 1.5 times the interquartile range (box size: 75%-ile minus 25%-ile).
   2. The lower line is the minimum, up to 1.5 times the interquartile range (box size: 75%-ile minus 25%-ile).
   3. The middle line is the median (50%-ile).
   4. The top of the box is the 75%-ile.
   5. The bottom of the box is the 25%-ile.
   6. The interquartile distance is the distance between the 25%-ile and 75%-ile.



To create a box plot by brand:

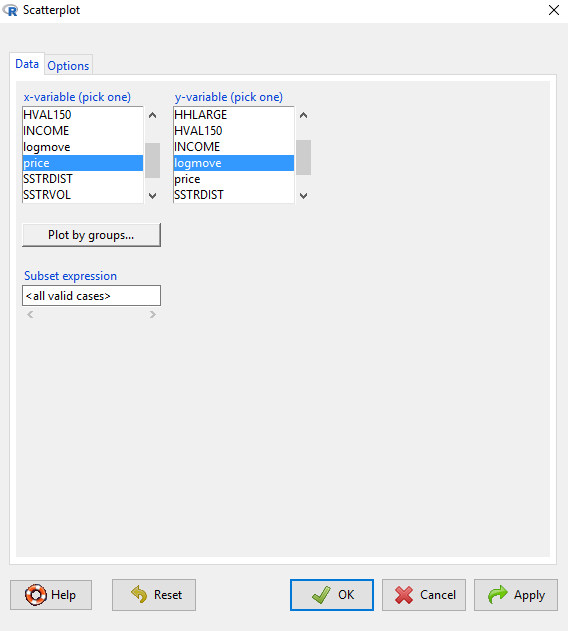
1. Click on Graphs, Boxplot, price.
2. Click on Plot by groups, select brand, OK, then click OK again.
3. Data points beyond the whiskers are outliers.
4. Outliers for boxplots are points that are more than 1.5 times the interquartile distance from the 25%-ile or 75%-ile.



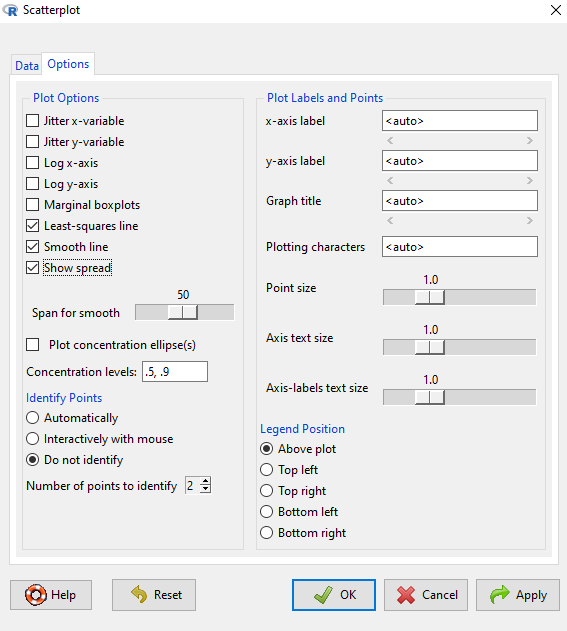
**Scatterplots**

To generate a scatterplot:

1. Click on Graphs, Scatterplot.
2. Select price as the x-variable.
3. Select logmove as the y-variable.

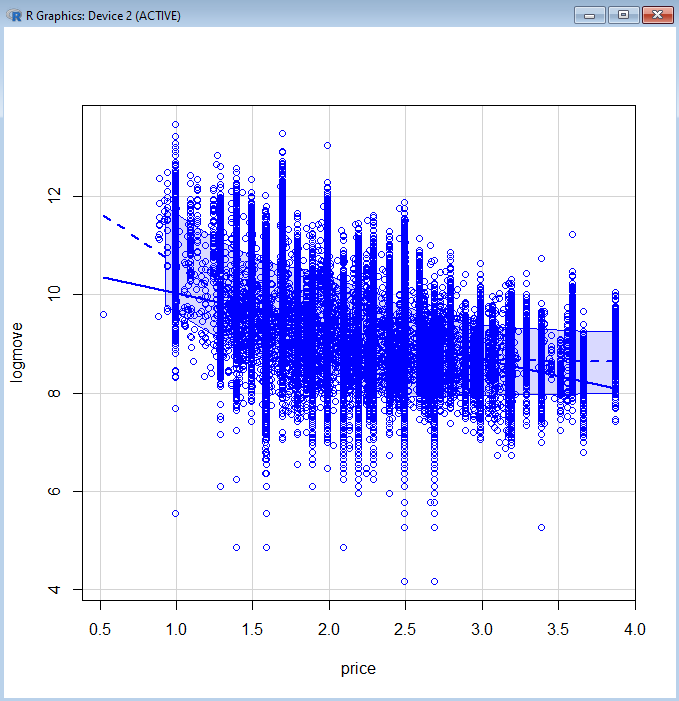


1. Click on the Options tab, select:
   1. Least-squares line
   2. Smooth line
   3. Show spread
2. Click on OK.



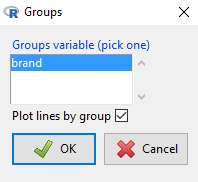
To interpret the chart:

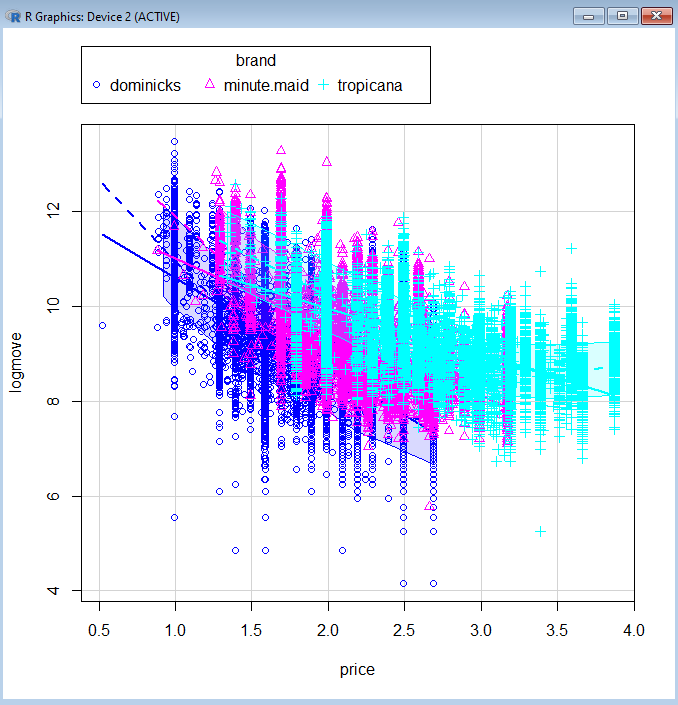
1. The blue dots are the price versus log(sales) for each time period, store, and brand.
2. The blue solid line is the linear regression line through the data.
3. The dashed line is the average of logmove as price changes.
4. The dash-dot lines are one standard deviation above and below the average line.



Now generate a scatterplot by brand:

1. Click on Graphs, Scatterplot.
2. Select price as the x-variable.
3. Select logmove as the y-variable.
4. Click on Plot by Groups, select brand, then OK.
5. Click on OK.



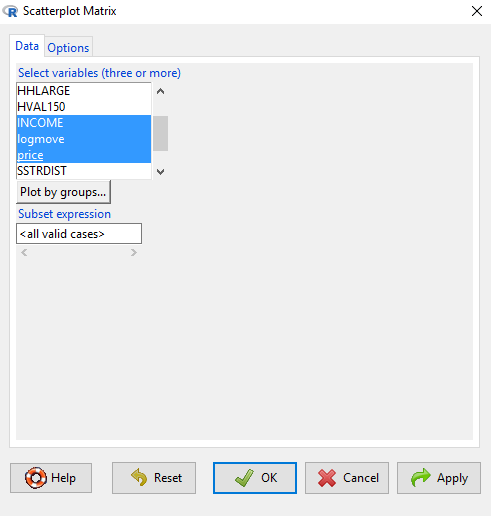


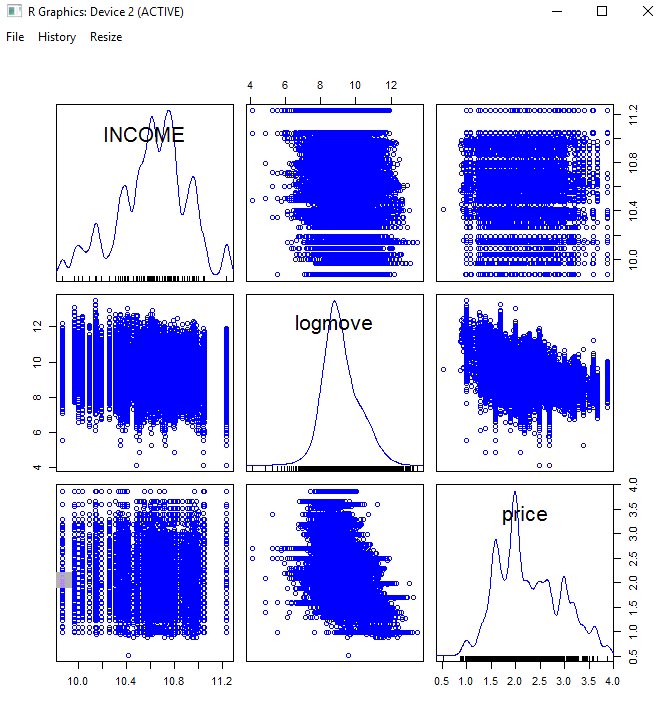
1. Interpretation:
   1. Brand Dominicks is in dark blue.
   2. Brand Minute Maid is in purple/pink.
   3. Brand Tropicana is in light blue.
2. Which is the premium brand?

**Plotting Pairwise Scatterplots with More Than Two Variables**

To create a matrix of scatterplots with more than two variables:

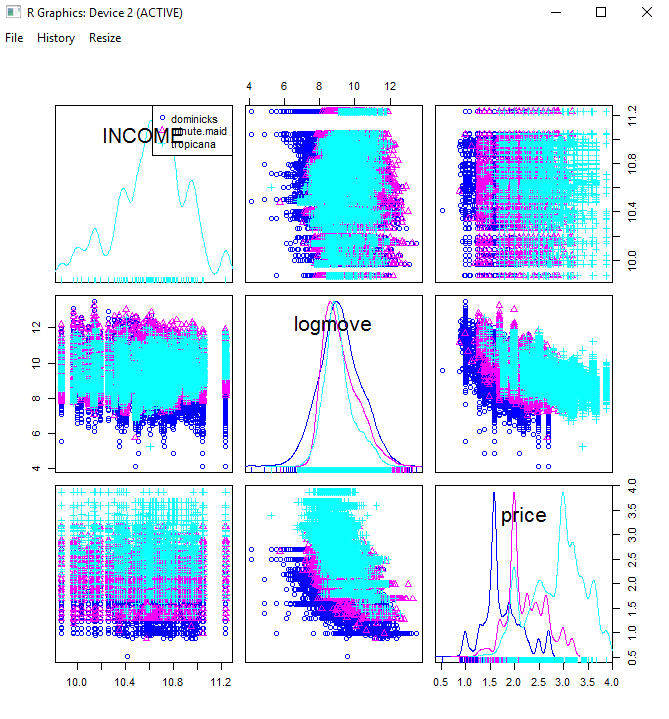
1. Click on Graphs.
2. Click on Scatterplot Matrix.
3. To select multiple variables, hold down the control key, then select INCOME, logmove, and price.
4. Click OK.





The diagonal is the distribution of data points (density function). Off-diagonal are the scatterplots for the pair of variables listed to the side and above/below the scatterplot.

1. Now perform a Scatterplot Matrix by Groups (brand).
2. Click on Graphs, Scatterplot Matrix.
3. Check that INCOME, logmove and price are still highlighted.
4. Click Plot by Groups.
5. Click on brand, then OK.
6. Click OK again.

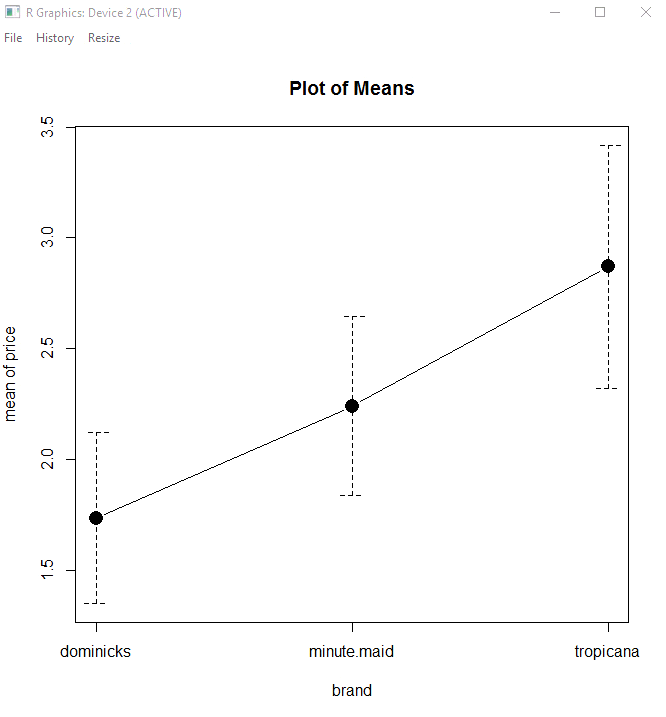


1. Once again, the brands are color coded.

**Plot of Means**

To determine if the different brands have different prices, on average, plot the means:

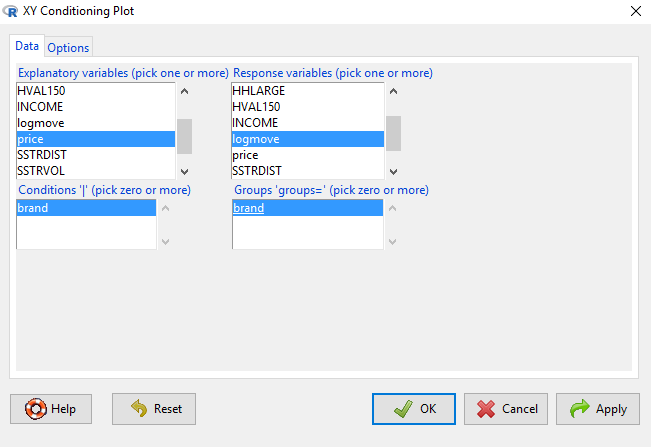
1. Click on Graphs, Plot of Means.
2. Select price.
3. In the Options tab, click on standard deviations, then OK.

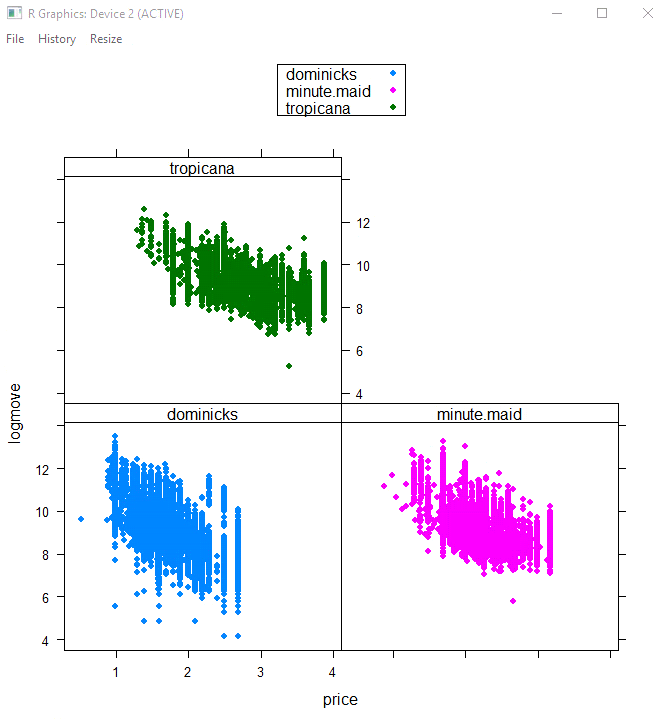


**XY Plots**

Now let’s generate XY plots by brand:

1. Click on Graphs, XY Conditioning Plot.
2. Select price for the explanatory variable.
3. Select logmove for the response variable.
4. Click on brand for each.
5. Click OK.

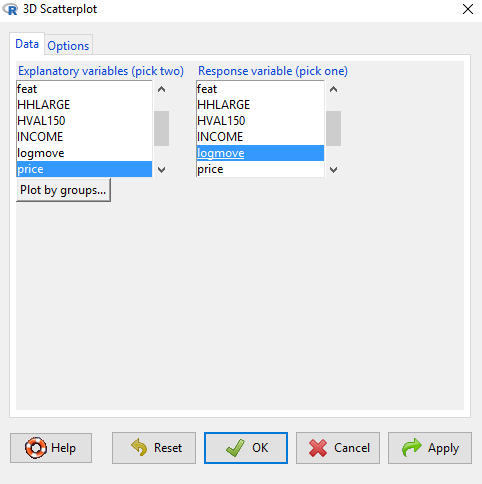
****

****

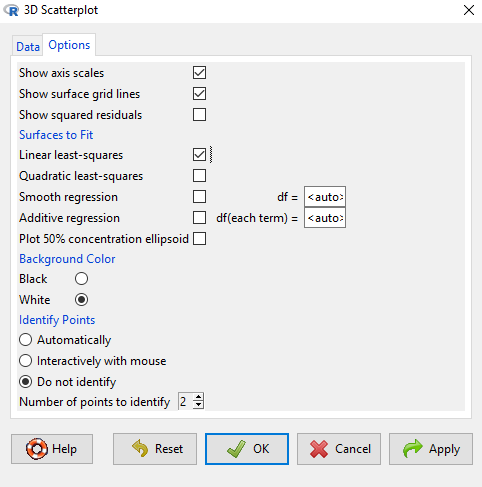
**7.4.2 R: 3D Graphs Demo**

To generate 3D graphs:

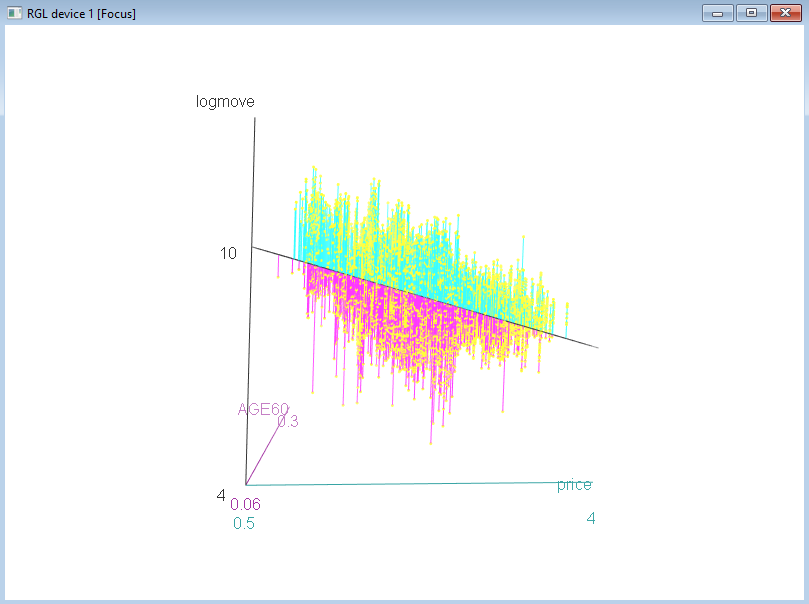
1. Graphs, 3D Graph, 3D Scatterplot.
2. Select AGE60 and price as explanatory variables by holding down the control key, then clicking on AGE60 and price.
3. Select logmove (log of sales) as the response variable.



1. Click on the Options tab, select Linear least-squares.



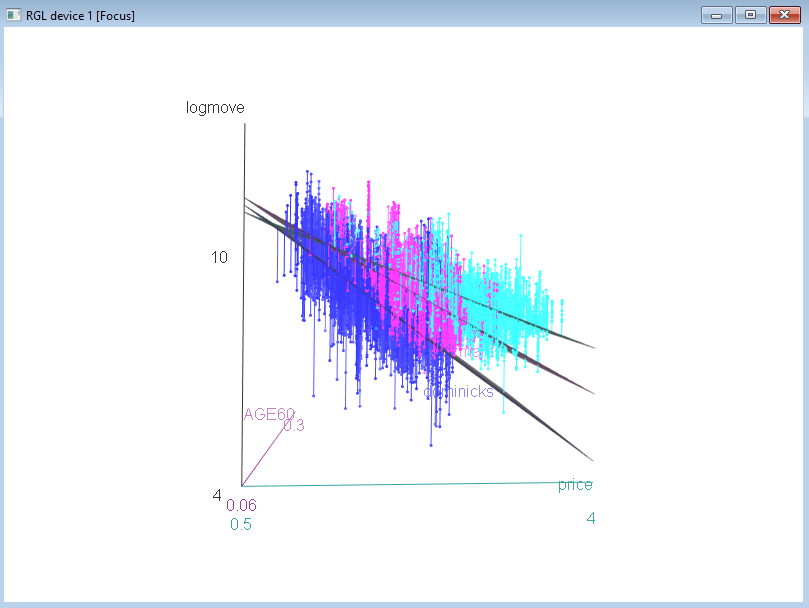
1. Click OK.
2. Note: The graph might be behind one of your screens. Expand the window by clicking on the box in the upper right corner of your graph.
3. Rotate the graph by clicking on the graph with your mouse, hold the mouse button down, and move.



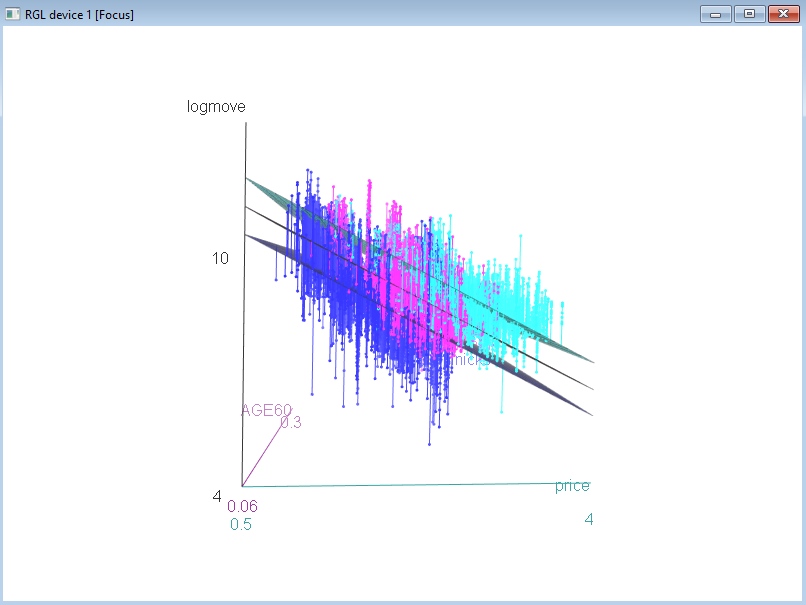
1. The plane is the regression plane that shows how price and age affect the logarithm of sales (logmove).
2. The dots are the data points.
3. The lines from the plane to the dots are error terms, called residuals.
4. Does price affect sales?
5. Does age affect sales?

To generate 3D graphs by brand:

1. Graphs, 3D Graph, 3D Scatterplot.
2. Select AGE60 and price as explanatory variables by holding down the control key, then clicking on AGE60 and price.
3. Select logmove (log of sales) as the response variable.
4. Click Plot by groups, select brand, then OK.
5. Click OK.
6. Expand the window by clicking on the box in the upper right corner of your graph.
7. Rotate the graph by clicking on the graph with your mouse, hold the mouse button down, and move.



1. Which brand is more sensitive to price?
2. This model has a different slope for each brand.
3. Different slopes reflect different elasticities of demand.
4. Now rerun the 3D graph by clicking on Graphd, 3D Graphs, 3D Scatterplot.
5. Click on Plot by: brand, then check the box Parallel regression surfaces, then click OK, and OK again.
6. This model has a different intercept for each brand, but the same slope.
7. Different intercepts for each brand represent brand premium.



1. The technical term for different slopes and different intercepts are:
   1. Dummy variables produce different intercepts
   2. Moderating effects (interactions) produce different slopes

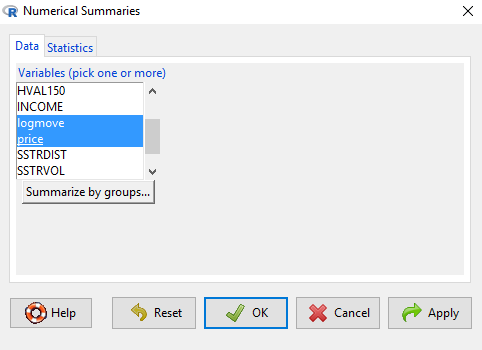
**Saving Graphs**

You can save graphs by clicking Graphs, Save Graph to File, then select the type of file.

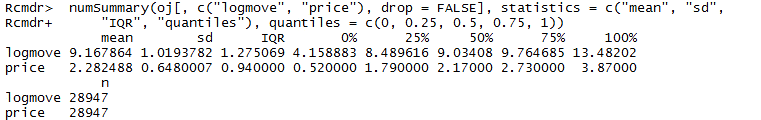
**7.5.2 R: Statistical Summaries Demo**

The mean, standard deviation and quartiles can be found by:

1. Click on Statistics, Summaries, Numerical Summaries.
2. Select logmove, price by holding down the control key.

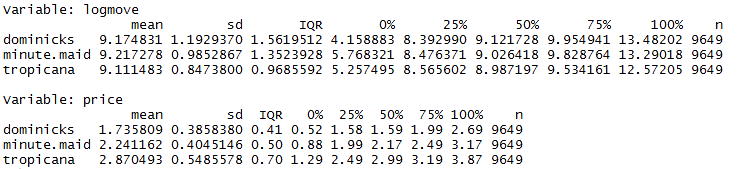


1. Click OK.



To categorize by brand, the mean, standard deviation and quartiles can be found by:

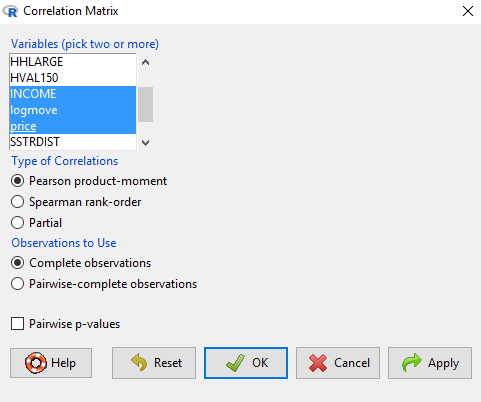
1. Click on Statistics, Summaries, Numerical Summaries.
2. Select logmove, price by holding down the control key.
3. Click Summarize by Groups, click brand, then OK.
4. Click OK.



**7.6.2 R: Correlations Demo**

To generate a correlation matrix:

1. Click on Statistics, Summaries, Correlation Matrix.
2. Hold down the control key and select INCOME, logmove, price.

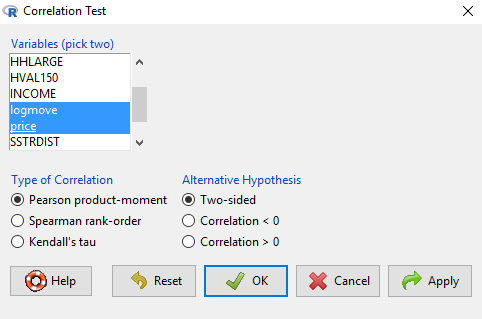


1. Click OK.

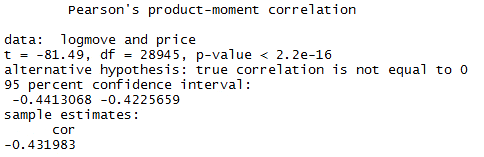
Correlation output

The matrix shows the correlation, but not the statistical significance. To calculate significance:

1. Click on Statistics, Summaries, Correlation Test.
2. Select both logmove and price.



1. Click OK.

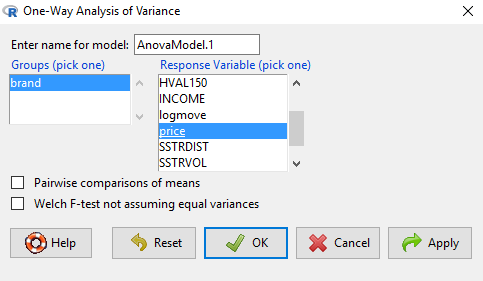


1. The *p*-value is less than 0.05, so the correlation is statistically significant.

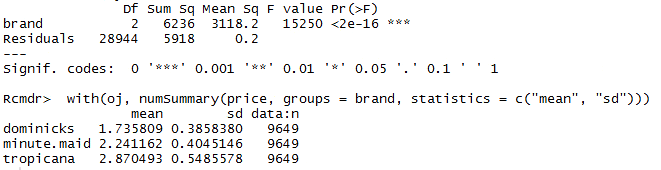
**7.7.2 R: ANOVA Demo**

ANOVA stands for analysis of variance. It compares the means of several groups to determine if the groups are different. Let’s see if prices are different across brands.

1. Click on Statistics, Means, One-way ANOVA.
2. For the response variable, click on price.



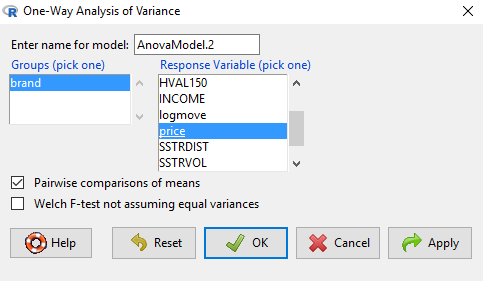
1. Click OK.



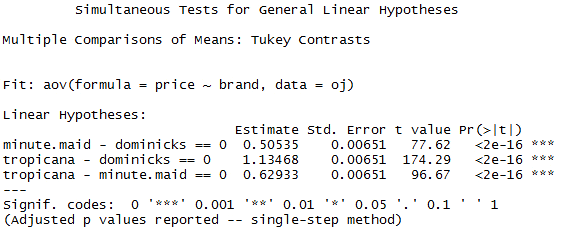
1. The F-statistic *p*-value [Pr(>F)] is less than 0.05. That means that one of the brands has a price that is statistically different from the others.

To determine which products are different, we need to perform a pairwise comparison.

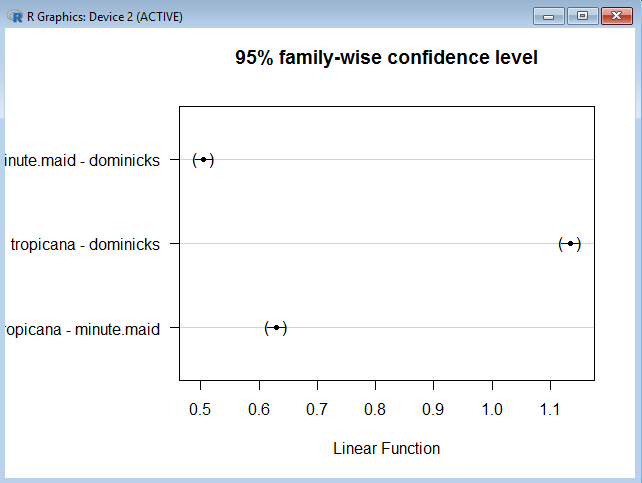
1. Click on Statistics, Means, One-way ANOVA.
2. For the response variable, click on price.
3. Check the box Pairwise comparison of means.



1. Click OK.
2. The pairwise comparison estimates the price and confidence interval for each brand (lower and upper interval values). Do they overlap?



1. The graph portrays the estimate of price difference between brands and the confidence interval. If the difference is not zero, then we can conclude that the product prices compared are different.



**7.8.1 R: Regression and Modeling**

So far, we have been performing regressions on a dependent variable Y against an independent variable X. For example, we can examine how education (X) affects income (Y). Pictorially, this would appear as:

Education

Income

+

The line and arrow identify a relationship between education and income. The plus sign above the line indicates that the relationship is positive, i.e., if education increases, then income increases.

This relationship can be written as:

Income = f(Education)

Which means that income is a function of education. One formulation of this could be the linear relationship:

Income = β0 + β1 \* Education

Where β0 is the intercept and β1 is the coefficient for education.

Now consider a third variable: technology. Technology has the potential for increasing the value of educated employees. Technology itself does not generate income for an employee but affects the value of education. This is called a moderating variable and is shown as:

+

Education

Income

Technology

+

This new model means that as education increases, income increases. The moderating effect of technology on education implies that technology further increases the value of education. This is modeled as an interaction term:

Income = β0 + β1 \* Education + β2 \* Education \* Technology

Therefore, the effect of education on income is influenced by the level of technology that an employee has.

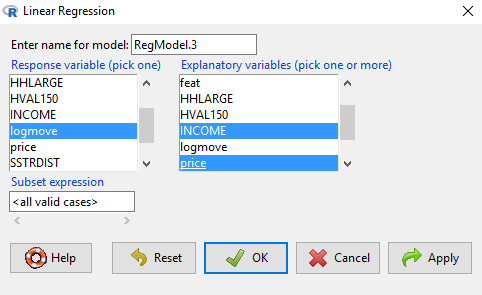
**Summary**

A dummy variable changes the intercept. A moderating effect (interaction) changes the slope.

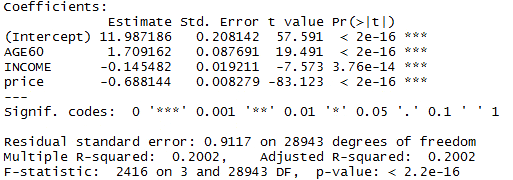
**7.8.2 R: Regression Demo**

Linear regression of the log of sales against age, income, and price can be performed by:

1. Click on Statistics, Fit Models, Linear Regression.
2. For response variable, click on logmove.
3. For explanatory variables, hold down the control key and click on AGE60, INCOME, price.



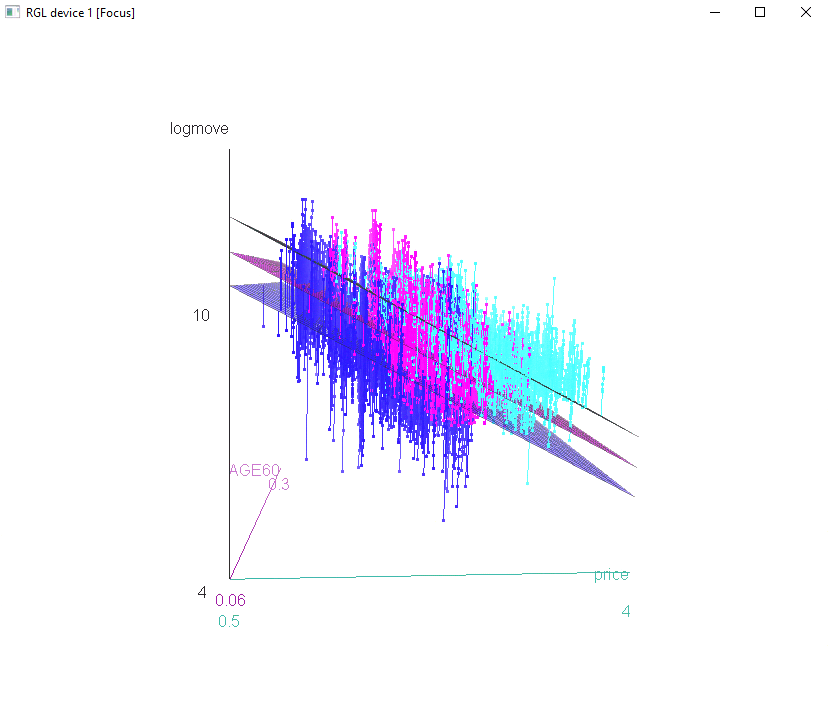
1. Click OK.



1. Is the equation statistically significant? If the *p*-value of the F-statistic is < 0.05, then the equation is statistically significant.
2. How much of the variability in the log of sales is explained by the explanatory variables? The R2 measures explanatory power.
3. Which explanatory variables are statistically significant?
4. How does each explanatory variable affect sales? Which affect it positively and which negatively? How do you interpret this (what does it really mean)?

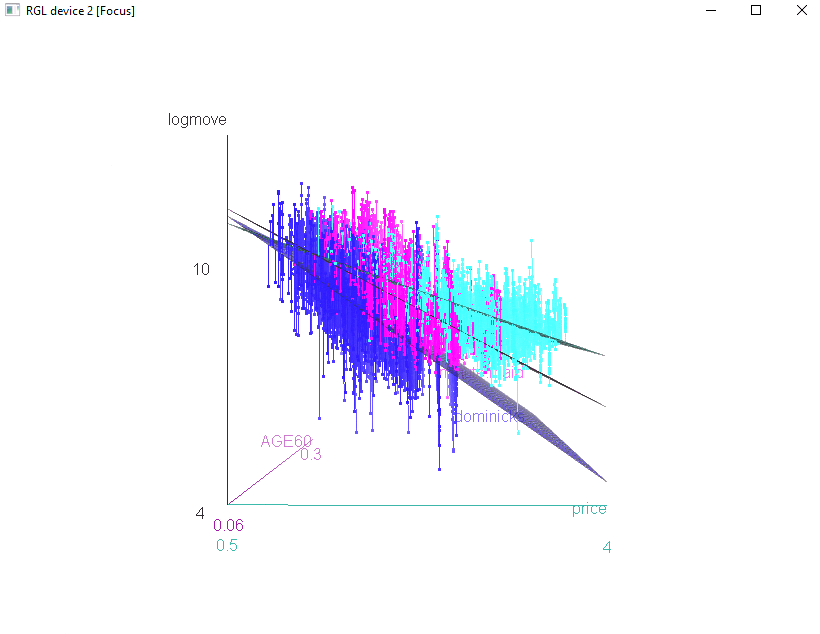
**7.9.2 R: Regression with Dummy Variables Demo**

1. Dummy variables are variables that take on the value or zero or one. For example, a dummy variable for homeowner would be 1 when the person is a homeowner and zero when the person is not a homeowner.
2. Dummy variables change the intercept in a regression equation.
3. Any categorical variable can be coded as a dummy variable. For example, education status can be coded as 1 for student, 0 for nonstudent. Similarly, employment status can be coded as 1 for employed, 0 for unemployed.
4. When a category has two possibilities (student, nonstudent), you only need one variable to represent the two categories.
5. When a category has three or more possibilities (Tropicana, Minute Maid, Dominicks), then you need n-1 dummy variables, where n is the number of categories. In the orange juice example, we would have a dummy variable for Minute Maid and Tropicana. The intercept for Dominicks would be the regular intercept (called base case). The intercept for Minute Maid would be the base intercept plus the Minute Maid dummy coefficient. The intercept for Tropicana would be the base intercept plus the Tropicana dummy coefficient.
6. The following picture is an example of the use of a dummy variable.



**7.10.2 R: Regression with Moderating Effects Demo**

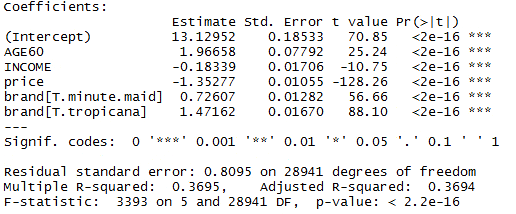
1. A moderating effect occurs when one variable magnifies the effect of another variable.
2. A moderating effect changes the slope in a regression equation.
3. A moderating effect is modeled by multiplying two variables together.
4. The following picture is an example of a moderating effect (interaction) where the brand interacts with price.
5. The effect of price on logmove for each brand is reflected in the different slopes.



**Regression with Dummy Variables**

A dummy variable in a regression can assist in determining if the intercept changes when the brand changes. To perform this more sophisticated regression:

1. Click on Statistics, Fit Models, Linear Model.
2. Click Reset.
3. Double click on logmove to make it the Y or dependent variable.
4. Double click on AGE60 to make it an X or explanatory variable.
5. Double click on INCOME to make it an X or explanatory variable.
6. Double click on price to make it an X or explanatory variable.
7. Double click on brand (notice it says it’s a factor) to create a dummy variable for the brands. A dummy variable takes on the value of zero or one; zero if it is not that brand, one if it is that brand.
8. Click on OK.



1. The equation becomes:

Logmove = 13.13 + 1.97 \* AGE60 – 0.18 \* INCOME – 1.35 \* price

+ 0.73 \* brand(minute.maid) dummy

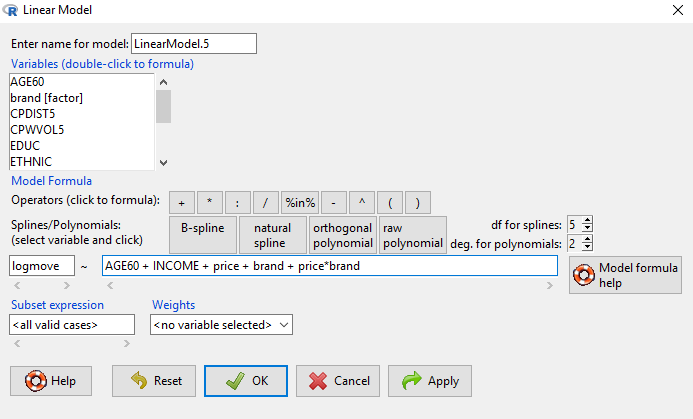
+ 1.47 \* brand(tropicana) dummy

1. Notice that there is a dummy for Minute Maid and Tropicana, but not Dominicks. Dominicks is the base, so the intercept represents Dominicks’ intercept.
2. In the Output section, the intercept is 13.12952. This is the intercept for Dominicks. The intercept for Minute Maid is 13.12952 + 0.72607. The intercept for Tropicana is 13.12952 + 1.47162.
3. This means, all else being equal, the log of sales is highest for Tropicana.

**Moderating Effects (Interactions of Price and Brand)**

In the previous example, we examined if the intercept is different for each brand. It’s possible that the slope of the relationship between price and sales varies by brand. To test this, we create what is called an interaction term. An interaction is two variables multiplied together.

1. Click on Statistics, Fit Models, Linear Model.
2. Click Reset.
3. Double click on logmove.
4. Double click on AGE60.
5. Double click on INCOME.
6. Double click on price.
7. Double click on brand (notice it says it’s a factor).
8. Double click on price (again).
9. Click on the multiplication sign (\*).
10. Click on brand.
11. Click on OK.



1. The coefficient on price is -1.94480. That means as price increases, the log of sales declines. But since we included an interaction term, this only applies to Dominicks.
2. We need to include the price\*brand effect for Minute Maid and Tropicana. For Minute Maid, the coefficient on price is -1.94480 + 0.47545 = -1.46935.
3. For Tropicana, the coefficient is ‑1.94480 + 0.94817 = ‑0.99663.
4. Which brand is more sensitive to price?

