

R: Intermediate

Datafiles

For these exercises, download the files:

“Business Analytics – Week 8 Instructions.doc”

“Business Analytics – Week 8 oj.xls”

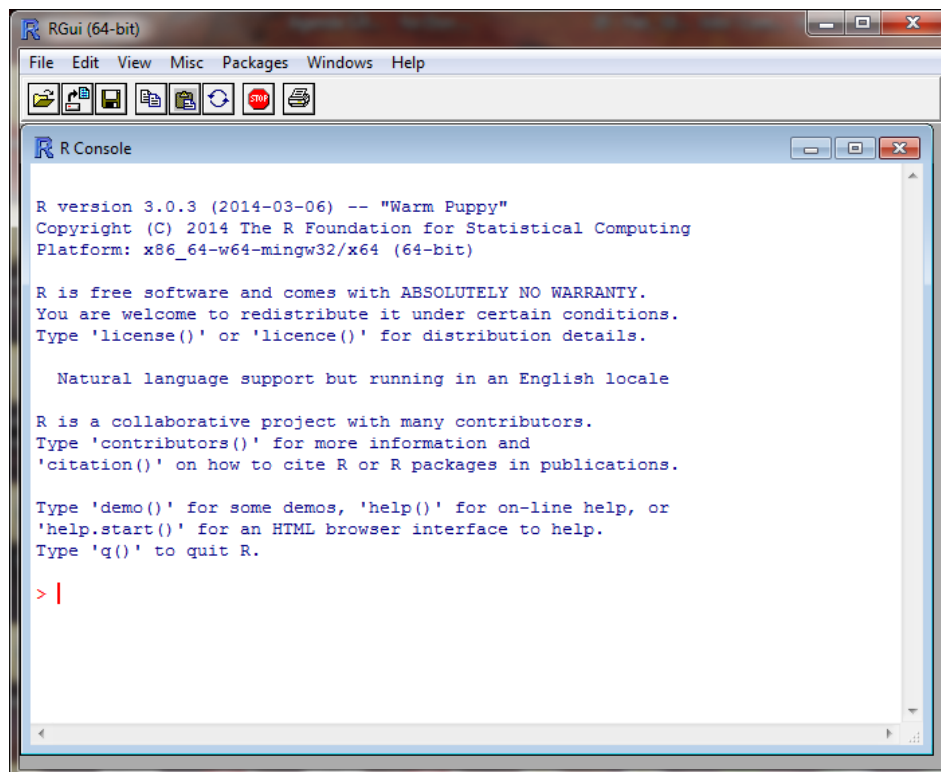
Review: Installation of R

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: <http://cran.r-project.org/bin/windows/base/>
2. For a Mac, go to <http://cran.r-project.org/bin/macosx/>
3. Click on Download R 3.0.3 for Windows
4. Click on Run, and follow the install instructions

Starting R

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 R

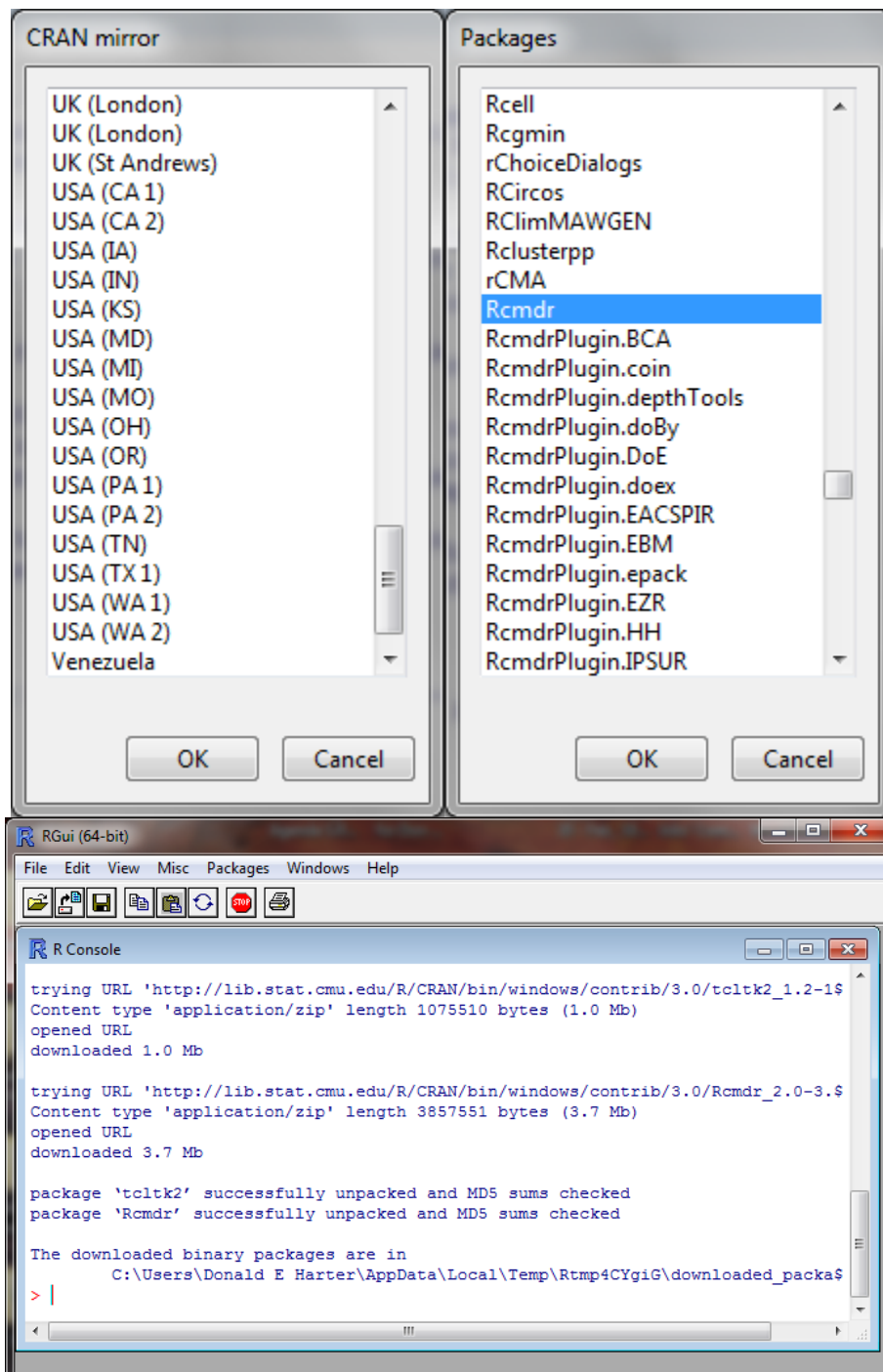


This is the command line screen. You can enter commands, but need to know the syntax. There is a simpler approach to running R, called Rcmdr (R Commander). If you are running a Whitman computer, Rcmdr is already installed. If not, you need to install it.

Installing R Commander

Follow these steps only if you don't already have Rcmdr installed.

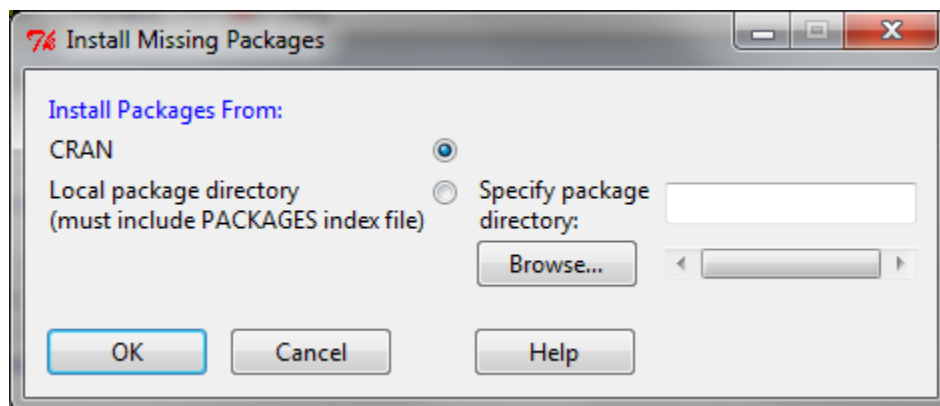
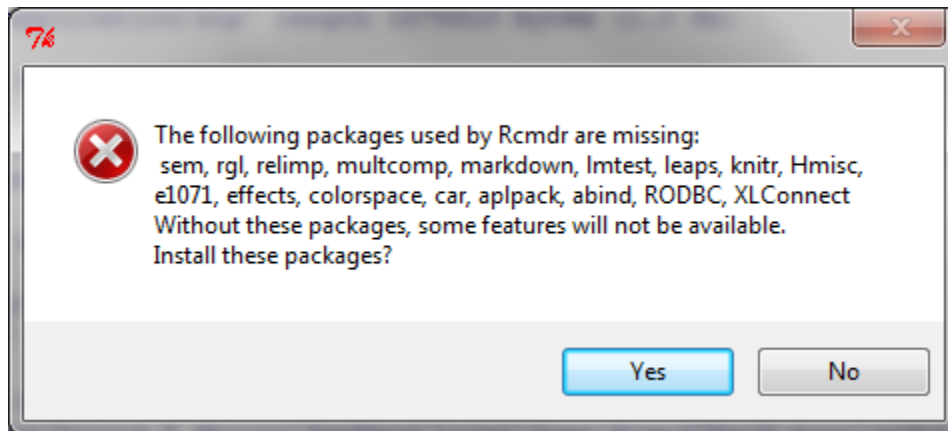
1. At the top of the screen, click on Packages
2. In the drop down menu, click on Install Package(s)
3. In the CRAN mirror, select the location closest to you; use USA (PA 1), then click OK
4. In the Packages screen, click on Rcmdr, then OK
5. When prompted to create a personal library, click Yes

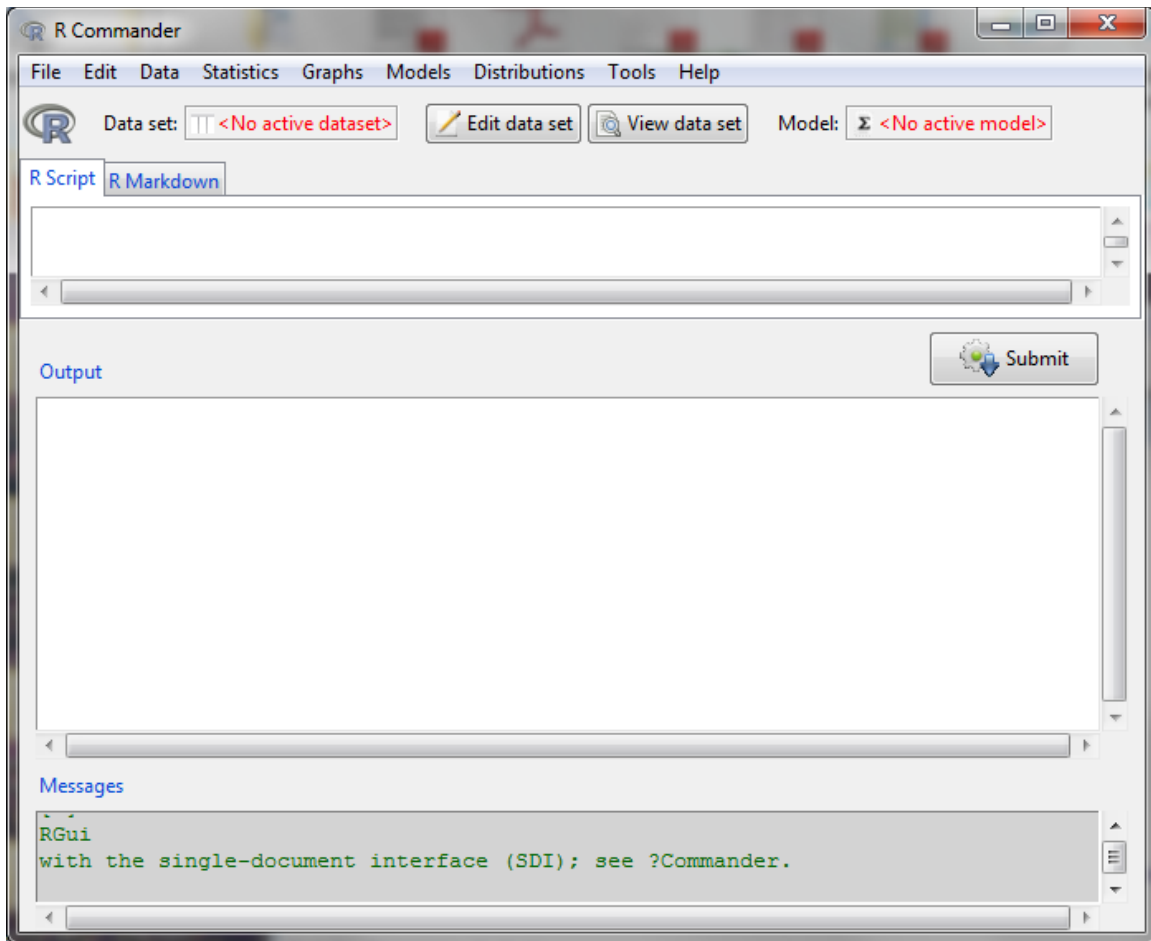


Launch Rcmdr (R Commander)

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

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



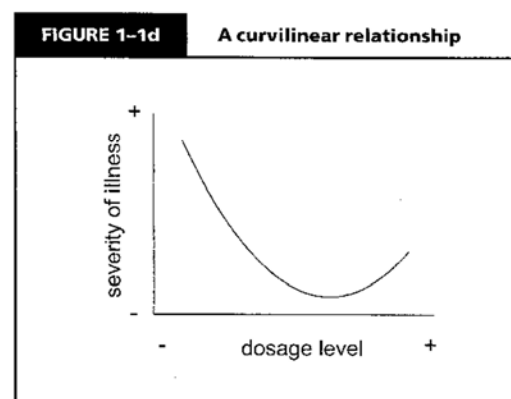
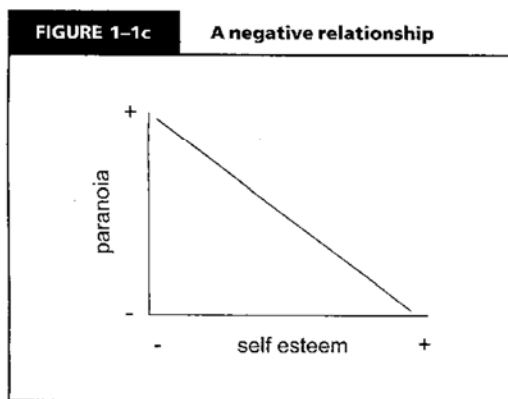
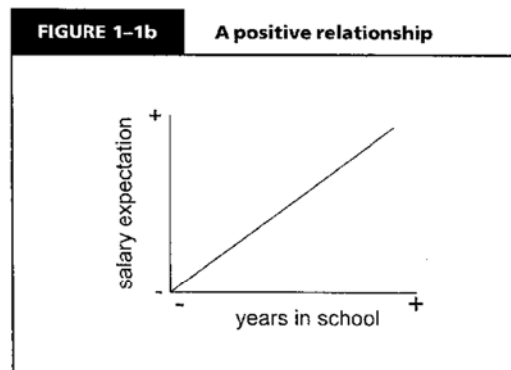
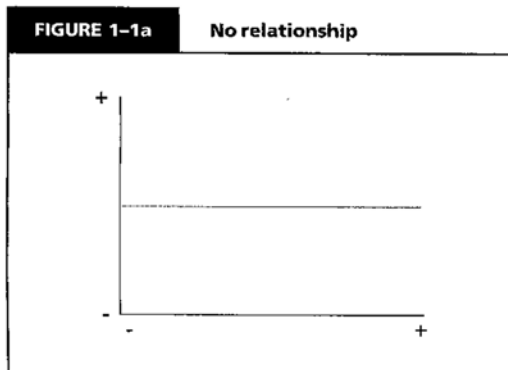


Modeling Background (Correlation & Regression)

Identifying data relationships is key to modeling behavior of customer, student, and corporate data. First, let's consider two variables and the relationships between them. When comparing two data variables, you can have:

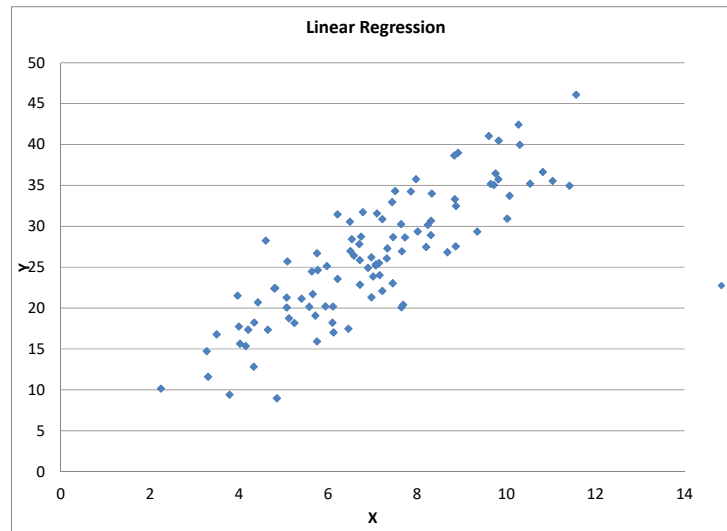
1. No relationship between the variables
2. A positive relationship (when one variable goes up, the other goes up)
3. A negative relationship (when one variable goes up, the other goes down)
4. A curvilinear relationship (a non-linear relationship)

Examples of these, from The Research Methods Knowledge Base by Trochim & Donnelly (2007):

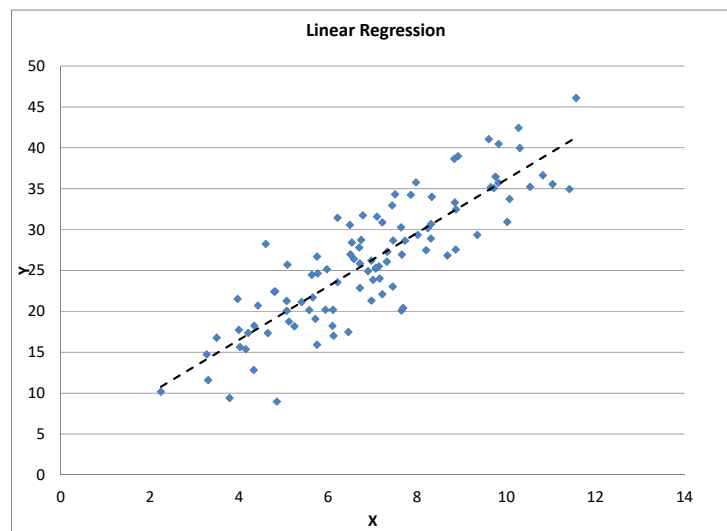


Regression

Linear regression is a technique that calculates the relationship between a dependent variable Y and one or more independent variables, or X's. Assume that you have data similar to the picture below.



You can calculate a regression trend line based on the data. This dashed line represents \hat{Y} which is the estimate of the Y equation.



The vertical distance between the line and the data point is called the residual or error term.

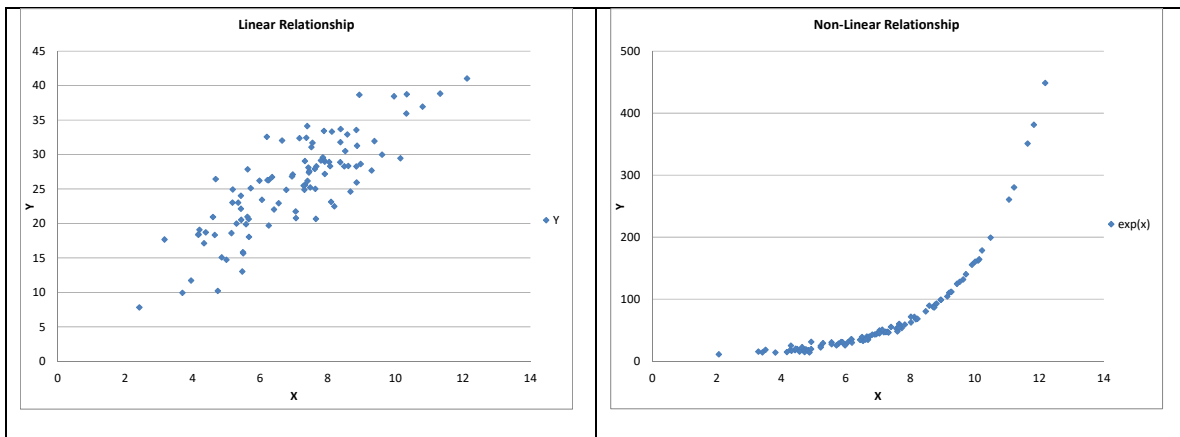
Regression Diagnostics

There are several assumptions of linear regression:

1. The relationships are linear
2. The X variables (explanatory variables) are not correlated
3. Distribution of residuals
 - a. The error terms have constant variance
 - b. The errors terms are not correlated
 - c. There are no outliers

Assumption #1: the relationship is linear

Let's examine each of these assumptions. In the pictures below, the left picture has data with a linear relationship, the right picture had non-linear data. Linear regression can only be used on data with a linear relationship. Transformations can be used to transform non-linear data into linear data. For example, exponential data like the data on the right can be converted into a linear relationship by taking the logarithm of both the Y and X variables.



Test for Linearity

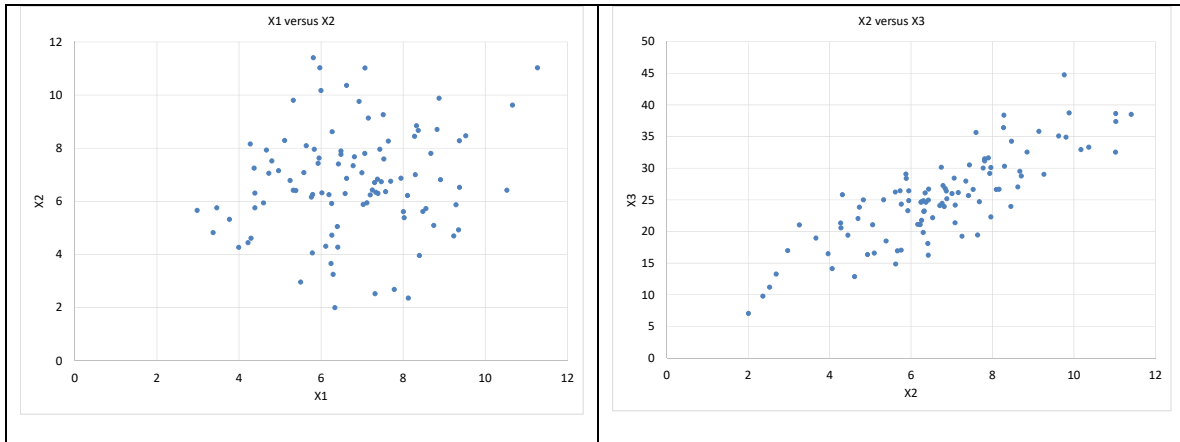
The Ramsey Regression Equation Specification Error Test (RESET) (1969) to test for linearity

Solution to non-linearity

The best solution for non-linear data is to transform the data using logarithms, squares, square roots, or inverses ($1/\text{variable}$). There are more advanced techniques which can assist in determining the correct transformation (Box-Cox for the Y variable; Box-Tidwell for the X variables).

Assumption #2: The X variables are not correlated (no multi-collinearity)

When including more than one explanatory or independent variable (i.e., X variable) in an analysis, you must ensure that they are not related to each other. If you plot the X variables, you should see no pattern, such as the picture on the left between variables X1 and X2. If you see a relationship, such as on the right between X2 and X3, then multi-collinearity exists.



Test for Multi-collinearity

The Variance Inflation Factor test of correlated explanatory variables

Solution to Multi-collinearity

If two or more variables are collinear (highly correlated), there are three solutions:

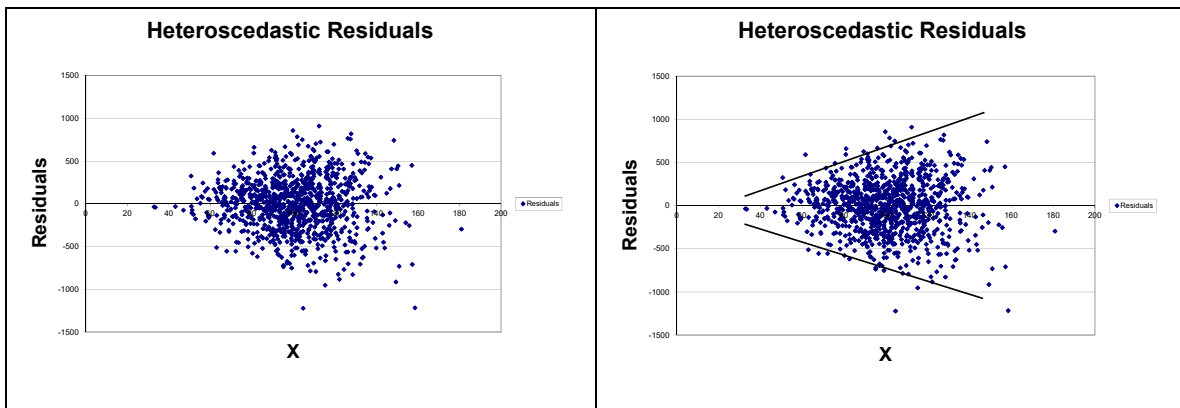
1. Combine the variables, for example, take an average of the variables
2. Drop one of the variables
3. Use factor analysis to combine variables

Assumption #3a: The error terms do not have constant variance (Heteroscedasticity)

The residuals (error terms) of a regression must have constant variance over a range of X values. If the size of the error terms depends on an X value, this is called heteroscedasticity.

Heteroscedasticity is often caused by performing a linear regression on non-linear data. In the charts below, there is no relationship between the X variable and the error term. On the right, the residuals or errors are heteroscedastic; the size of the error is dependent on the X value.

The picture below shows heteroscedastic residuals. Notice that the variability of the errors or residuals tends to grow larger for larger values of X. The picture on the right has lines added indicating the general growth in variability.



Test for Heteroscedasticity

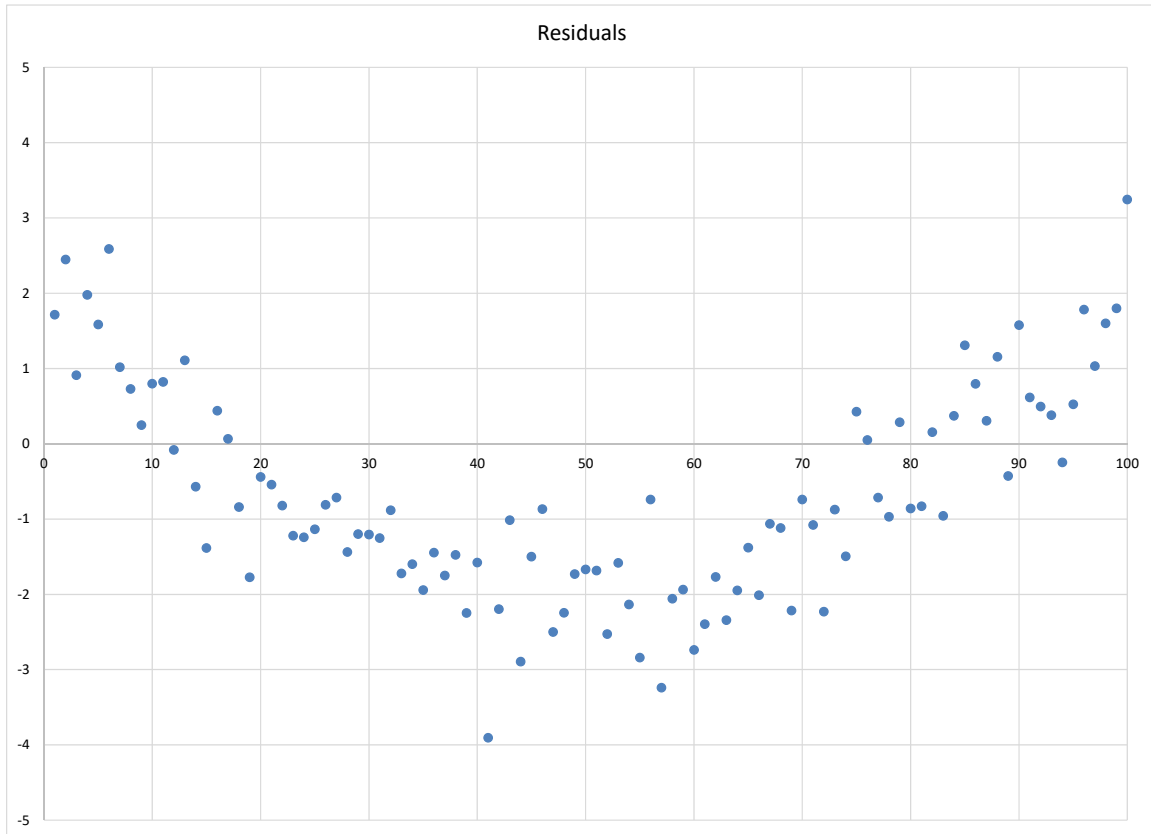
Breusch-Pagan test of heteroscedasticity

Solution to Heteroscedasticity

Heteroscedasticity is often caused by performing linear regression on non-linear data. Generally, solving non-linearity problems with transformations reduces or eliminates heteroscedasticity. If the problem is not completely resolved with a transformation, additional advanced techniques including Huber regression can correct lingering issues.

Assumption #3b: The error terms are not correlated (Serial Correlation)

When dealing with data over time, it's possible for the error terms from one time period to be highly correlated with the previous time period. This is called serial correlation. The error terms or residuals will have a pattern that is not random, such as in the picture below.



Test for Serial Correlation

Durbin-Watson test of serial correlation

Solution to Serial Correlation

To correct for serial correlation there are a number of techniques in time series, including rho differencing and ARCH.

Assumption #3c: There are no outliers

An outlier is a data point that is significantly different from other data points. Outliers are often the result of unusual circumstances or data entry errors. The data below has an outlier.



Test for Outliers

Bonferroni outlier test

Solution to Outliers

If the data point is clearly an outlier, you can drop the bad data point, but mention in your analysis that you dropped outliers.

Download Datasets

To access some excellent data sets used in the book “Data Mining and Business Analytics with R,” by Johannes Ledolter:

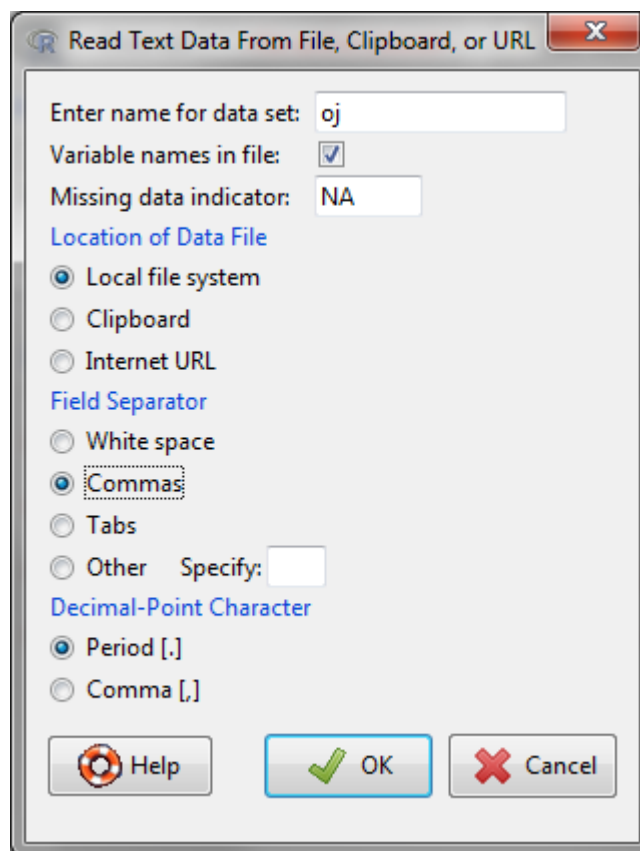
1. Go to the website:
<http://www.biz.uiowa.edu/faculty/jledolter/DataMining>
2. Click on Data Text
3. Right click on oj.csv, then save on your computer
4. Remember where you saved the file

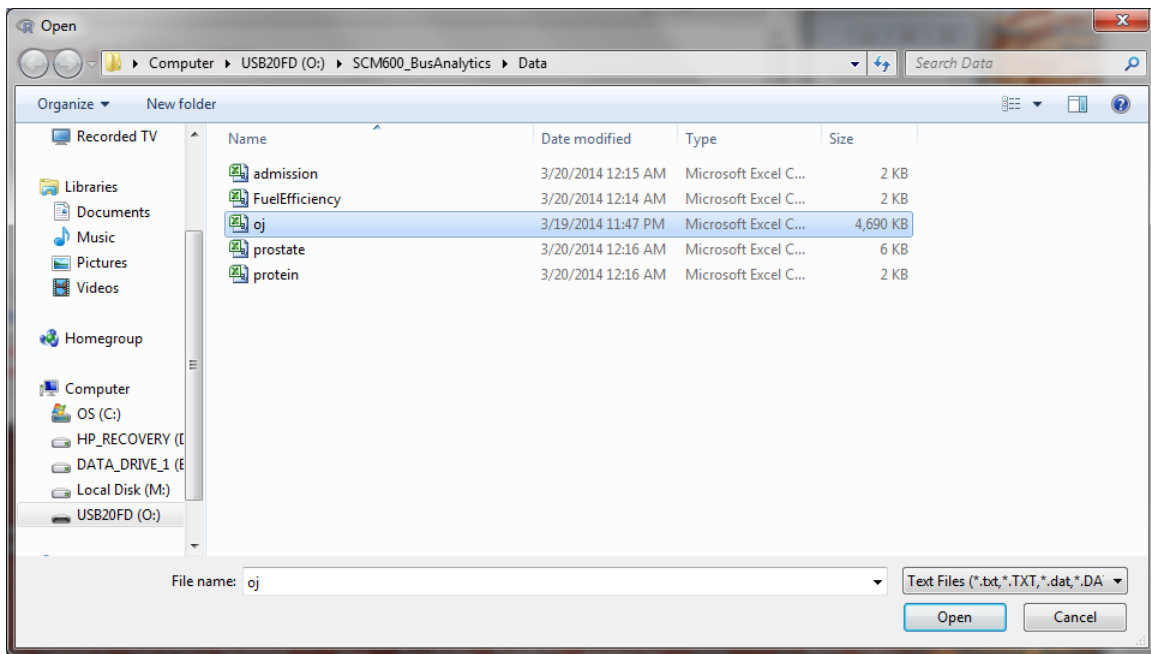
The Business Analytics - Week 8 oj.csv (orange juice) file can be downloaded from the course website.

Loading Data

To load data into R:

1. Click on Data at the top of the screen
2. Click on Import Data > From text file ...
3. Enter the name that you would like to use for this data set; type in oj
4. Change Field Separator to Commas, then OK
5. Click on the oj file, then Open

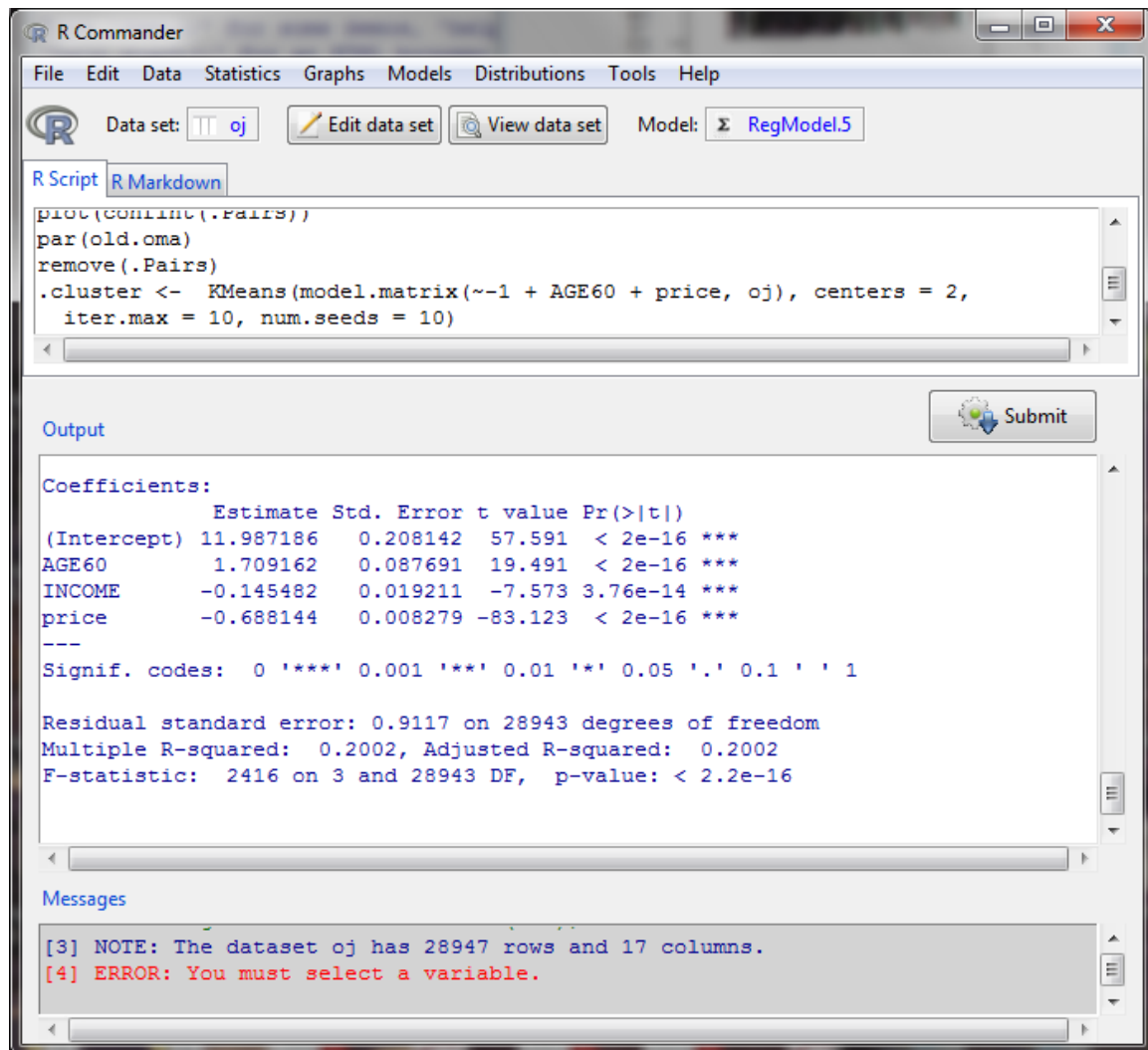




Session 8.4: Regression and the RESET test for linearity

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
4. Click OK



The screenshot shows the R Commander window with the following components:

- Menu Bar:** File, Edit, Data, Statistics, Graphs, Models, Distributions, Tools, Help.
- Toolbar:** Data set: oj, Edit data set, View data set, Model: RegModel.5.
- R Script Tab:** Contains the following R code:

```
plot(ConInt(.Pairs))
par(old.oma)
remove(.Pairs)
.cluster <- KMeans(model.matrix(~-1 + AGE60 + price, oj), centers = 2,
  iter.max = 10, num.seeds = 10)
```
- Output Tab:** Displays the regression results:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.987186   0.208142  57.591  < 2e-16 ***
AGE60         1.709162   0.087691  19.491  < 2e-16 ***
INCOME       -0.145482   0.019211  -7.573  3.76e-14 ***
price        -0.688144   0.008279 -83.123  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

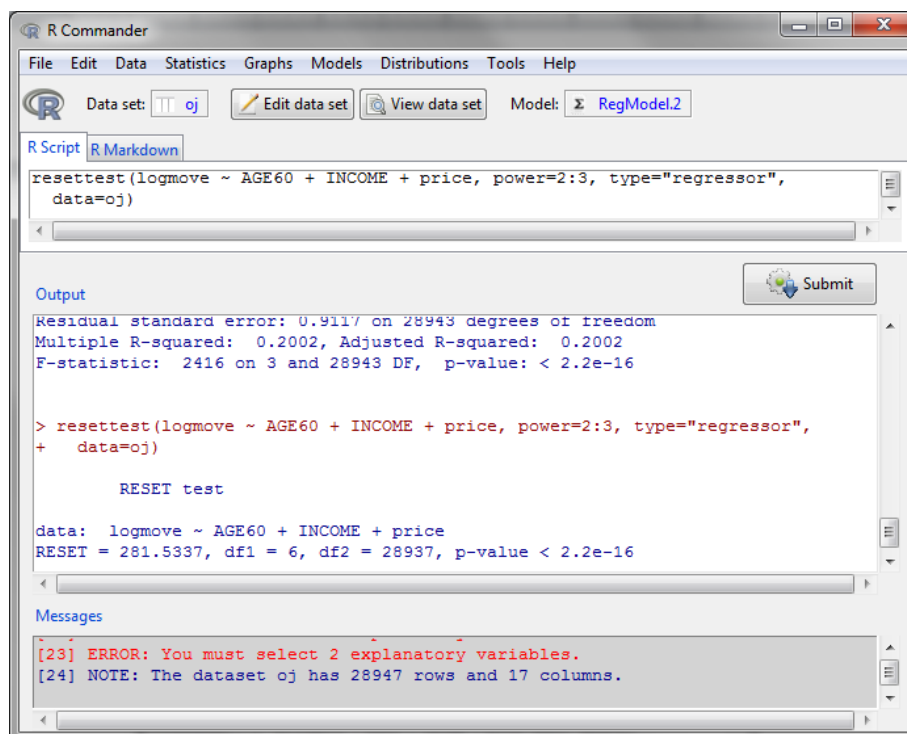
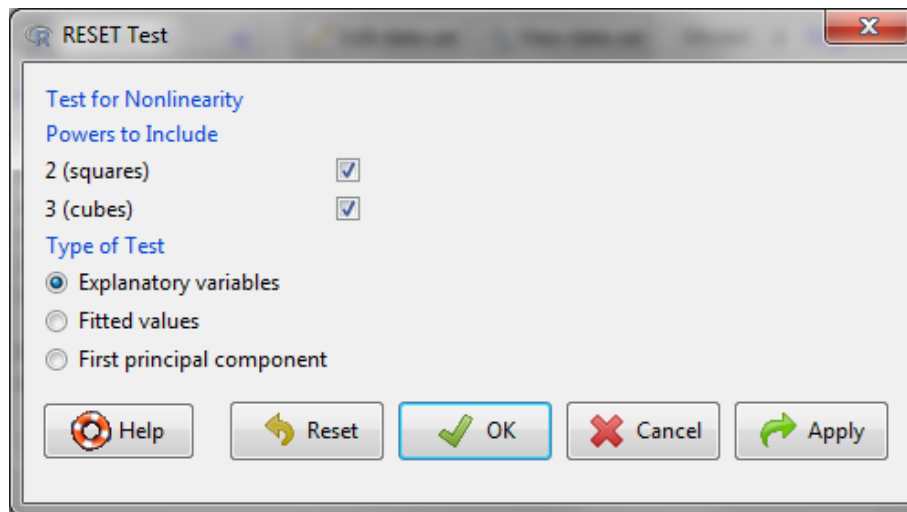
Residual standard error: 0.9117 on 28943 degrees of freedom
Multiple R-squared:  0.2002, Adjusted R-squared:  0.2002
F-statistic: 2416 on 3 and 28943 DF, p-value: < 2.2e-16
```
- Messages Tab:** Shows the following messages:

```
[3] NOTE: The dataset oj has 28947 rows and 17 columns.
[4] ERROR: You must select a variable.
```

Ramsey Regression Equation Specification Error Test (RESET) (1969) to test for linearity

To test if your equation is linear:

1. Click on Models, Numerical Diagnostics, RESET test for Non-linearity
2. Click OK

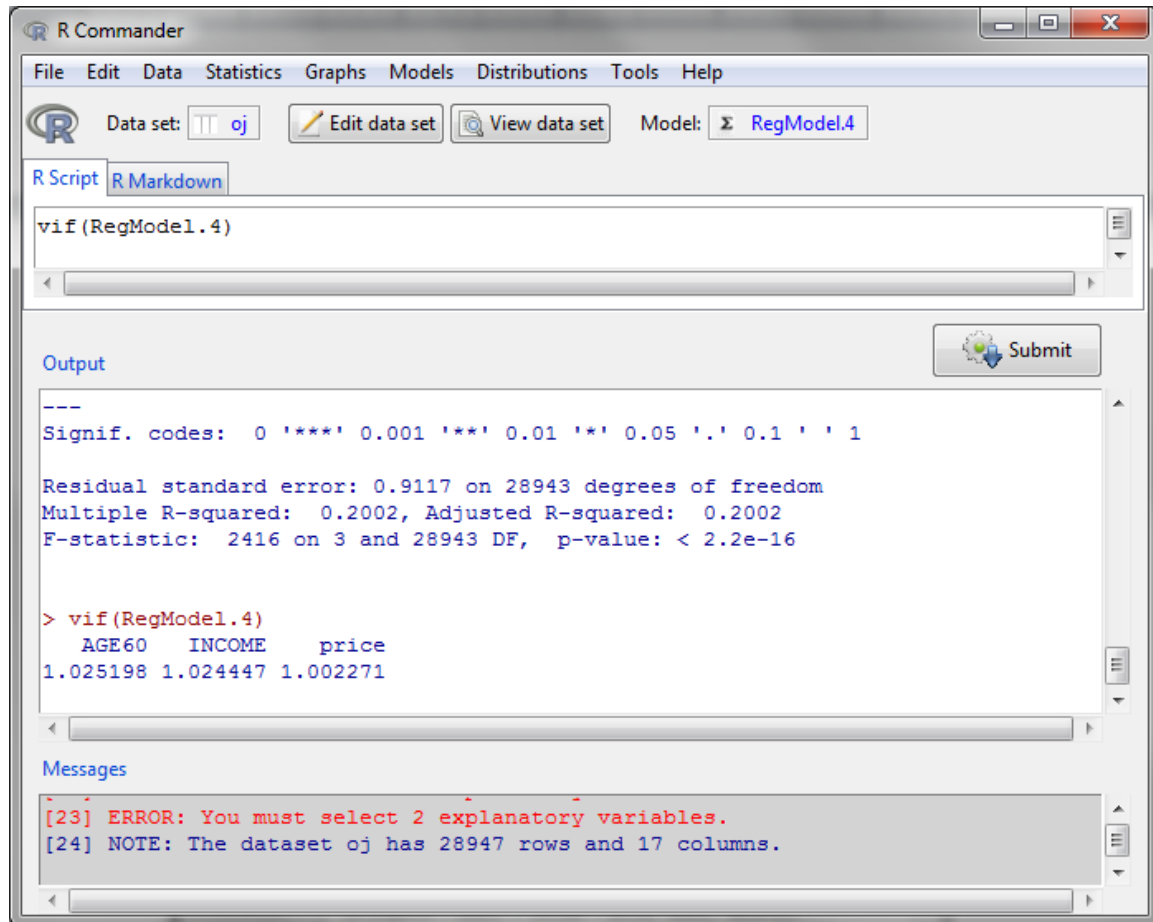


If the p-value is less than 0.05, then there is a non-linearity problem.

Session 8.5: Variance Inflation Factor test of correlated explanatory variables

To calculate the Variance Inflation Factor:

1. Click on Models, Numerical Diagnostics, Variance Inflation Factor



The screenshot shows the R Commander window. The 'Data set' is 'oj' and the 'Model' is 'RegModel.4'. The 'R Script' pane contains the command `vif(RegModel.4)`. The 'Output' pane displays the following information:

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9117 on 28943 degrees of freedom
Multiple R-squared:  0.2002, Adjusted R-squared:  0.2002
F-statistic: 2416 on 3 and 28943 DF, p-value: < 2.2e-16

> vif(RegModel.4)
    AGE60    INCOME    price 
1.025198 1.024447 1.002271
```

The 'Messages' pane shows the following messages:

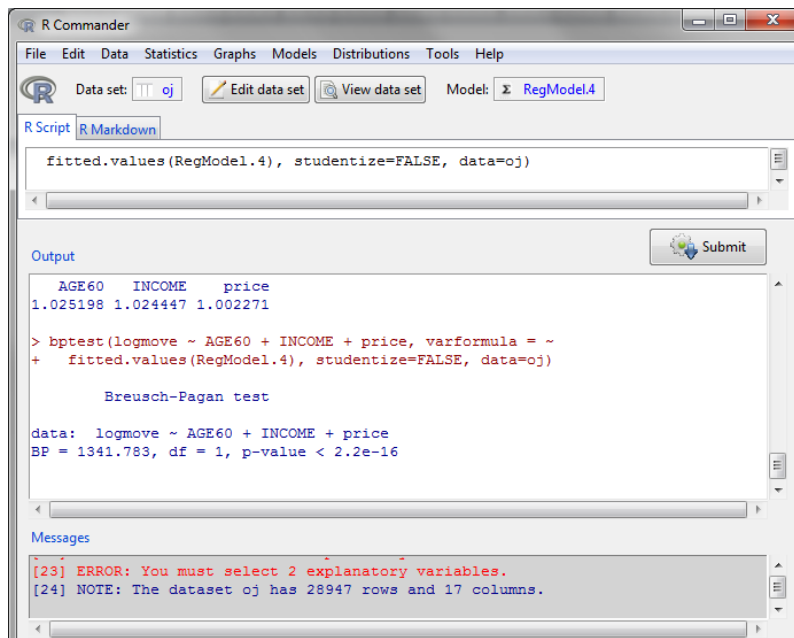
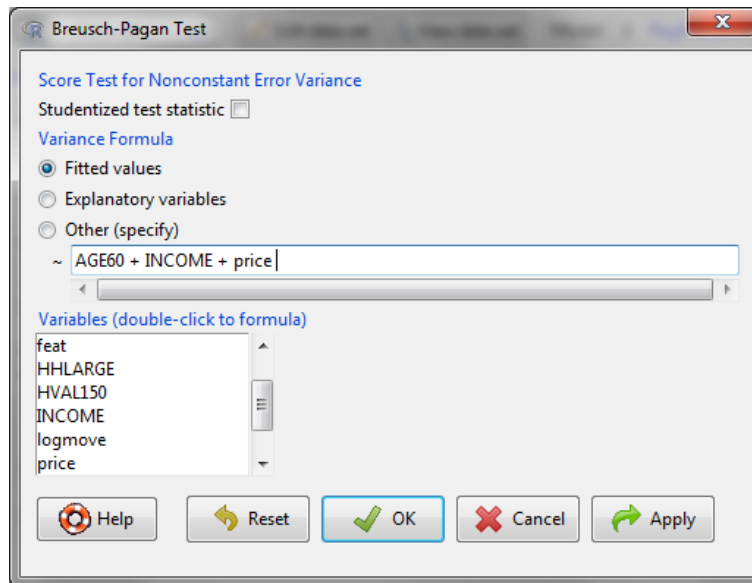
```
[23] ERROR: You must select 2 explanatory variables.
[24] NOTE: The dataset oj has 28947 rows and 17 columns.
```

If the variance inflation factors are less than 10, then there is no multi-collinearity. If multi-collinearity exists, then drop variables or combine variables. Factor analysis is one technique for combining variables.

Session 8.6: Breusch-Pagan test of heteroscedasticity

Heteroscedasticity means that the error terms are vary depending on values of the explanatory variables. To test for heteroscedasticity:

1. Click on Models, Numerical Diagnostics, Breusch-Pagan test for heteroscedasticity
2. Double click on AGE60, INCOME, price
3. Click on OK

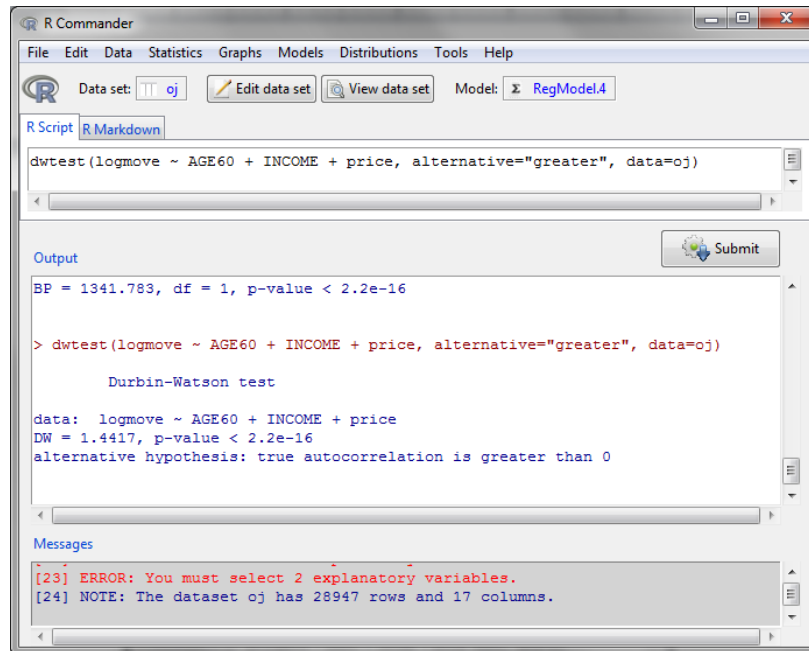


If the p-value is less than 0.05, then there is a problem with heteroscedasticity. Generally, this is a sign that the equation is non-linear.

Session 8.7: Durbin-Watson test of serial correlation

Serial correlation occurs when the errors terms are correlated. To test this,

1. Click on Models, Numerical Diagnostics, Durbin-Watson test for autocorrelation

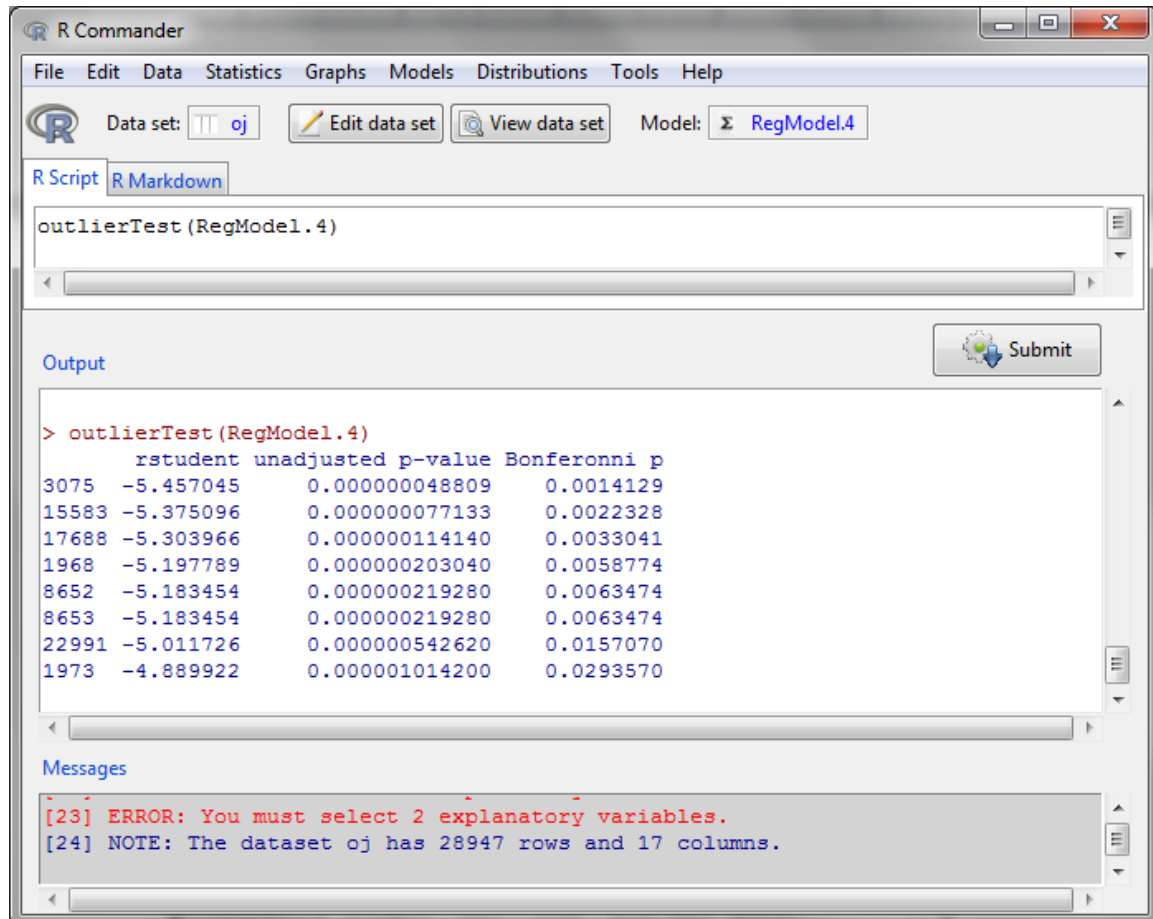


If the p-value is less than 0.05, there is a problem with serial correlation.

Session 8.8: Bonferroni outlier test

Outliers are extreme data points that can influence the results and lead to incorrect coefficients. To identify outliers,

1. Click on Models, Numerical Diagnostics, Bonferroni outlier test



The screenshot shows the R Commander window with the 'Models' menu open and 'Numerical Diagnostics' selected. The 'Bonferroni outlier test' option is highlighted. The output window displays the results of the test for the model 'RegModel.4'.

```
outlierTest(RegModel.4)
```

	rsstudent	unadjusted	p-value	Bonferonni	p
3075	-5.457045	0.000000048809	0.0014129		
15583	-5.375096	0.000000077133	0.0022328		
17688	-5.303966	0.000000114140	0.0033041		
1968	-5.197789	0.000000203040	0.0058774		
8652	-5.183454	0.000000219280	0.0063474		
8653	-5.183454	0.000000219280	0.0063474		
22991	-5.011726	0.000000542620	0.0157070		
1973	-4.889922	0.000001014200	0.0293570		

Messages:

```
[23] ERROR: You must select 2 explanatory variables.  
[24] NOTE: The dataset oj has 28947 rows and 17 columns.
```

In this example, data points numbered 3075, 15583, 17688, etc., are outliers. It's usually best to remove these data points from your data.

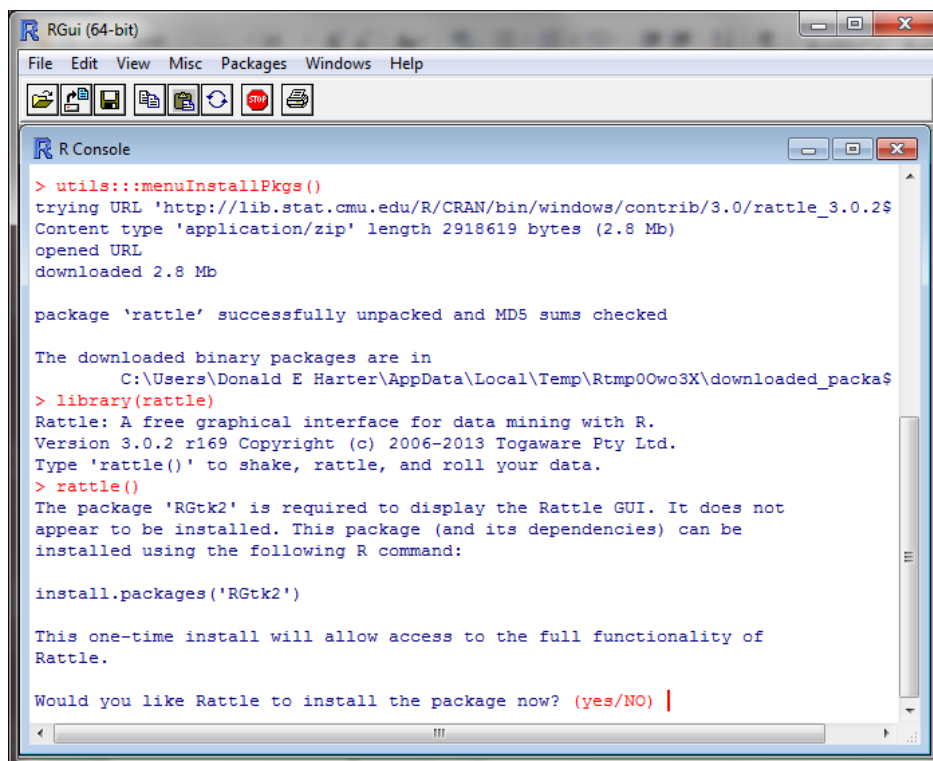
Session 8.9: Data Mining

Data mining tools allow you to explore in more detail groupings of data and more sophisticated analysis. Rattle is an add-in to R that facilitates data mining.

Installing Rattle

Follow these steps only if you don't already have Rattle installed.

1. At the top of the screen, click on Packages
2. In the drop down menu, click on Install Package(s)
3. In the CRAN mirror, select the location closest to you; use USA (PA 1), then click OK
4. In the Packages screen, click on rattle, then OK
5. Type `library(rattle)`
6. Type `rattle()`
7. When it asks "Would you like Rattle to install ...", type yes
8. If you receive an error message about GTK+, then install GTK+ by clicking OK
9. If you receive an error message about XML, click Yes to install
10. Similarly for cairoDevice



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

> utils::menuInstallPkgs()
trying URL 'http://lib.stat.cmu.edu/R/CRAN/bin/windows/contrib/3.0/rattle_3.0.2$
Content type 'application/zip' length 2918619 bytes (2.8 Mb)
opened URL
downloaded 2.8 Mb

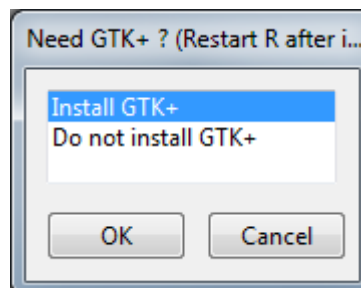
package 'rattle' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\Donald E Harter\AppData\Local\Temp\Rtmp0Owo3X\downloaded_packa$
> library(rattle)
Rattle: A free graphical interface for data mining with R.
Version 3.0.2 r169 Copyright (c) 2006-2013 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> rattle()
The package 'RGtk2' is required to display the Rattle GUI. It does not
appear to be installed. This package (and its dependencies) can be
installed using the following R command:

install.packages('RGtk2')

This one-time install will allow access to the full functionality of
Rattle.

Would you like Rattle to install the package now? (yes/NO) |
```



Launch Rattle

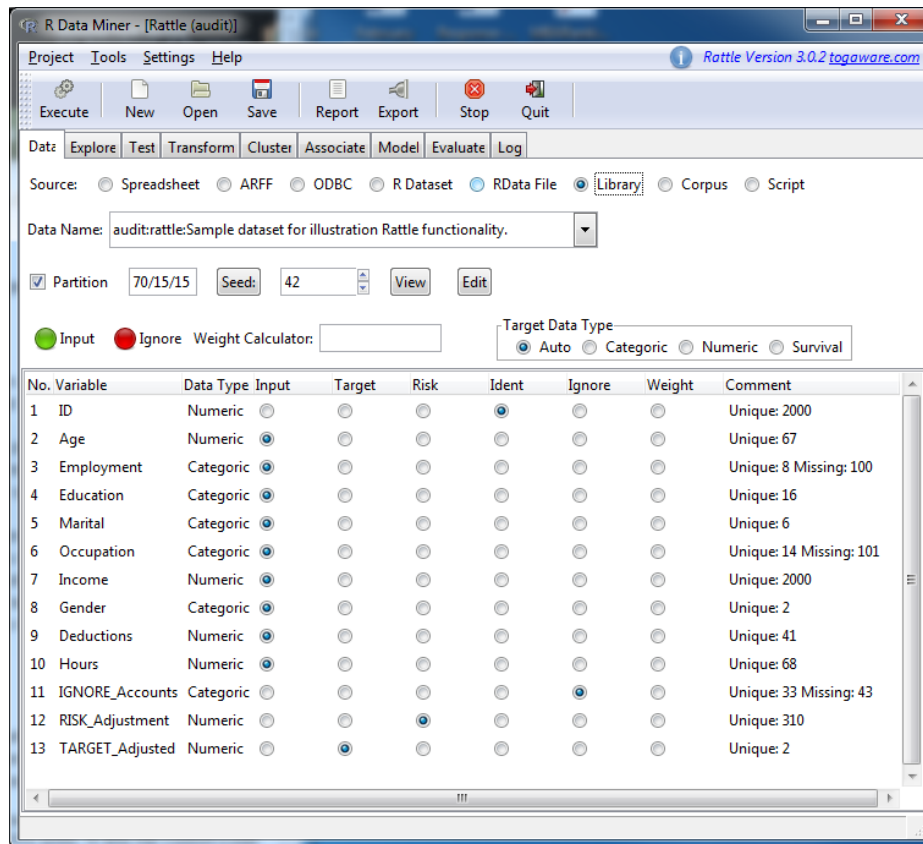
After Rattle has already been installed, you can always launch Rattle by:

1. Type library(rattle)
2. Type rattle(), press enter, then type yes, press enter

Loading Data

The package R has some built in data sets. To load data into R:

1. In the R Data Miner [Rattle] window, click on the Data tab
2. Check that Source indicates Library
3. Next to Data Name, use the drop down menu to select audit: rattle: Sample dataset
4. Click on Execute
5. This data set represents income tax audit data



Summary Statistics

1. Click on the tab Explore
2. Check the radio button Summary
3. Click on Execute

R Data Miner - [Rattle (audit)]

Project Tools Settings Help

Rattle Version 3.0.2 togaware.com

Execute New Open Save Report Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ Summary ☐ Distributions ☐ Correlation ☐ Principal Components ☐ Interactive

☒ Summary ☐ Describe ☐ Basics ☐ Kurtosis ☐ Skewness ☐ Show Missing ☐ Cross Tab

Age		Employment		Education		Marital	
Min.	:17.00	Private	:980	HSgrad	:441	Absent	:464
1st Qu.	:27.00	Consultant	:99	College	:314	Divorced	:179
Median	:37.00	PSLocal	:92	Bachelor	:239	Married	:656
Mean	:38.53	SelfEmp	:58	Master	:64	Married-spouse-absent	:16
3rd Qu.	:48.00	PSState	:51	Vocational	:63	Unmarried	:50
Max.	:83.00	(Other)	:51	Yr11	:55	Widowed	:35
		NA's	:69	(Other)	:224		

Occupation		Income		Gender		Deductions	
Executive	:207	Min.	:1599	Female	:428	Min.	:0.0
Professional	:175	1st Qu.	:33688	Male	:972	1st Qu.	:0.0
Clerical	:166	Median	:59371			Median	:0.0
Repair	:160	Mean	:83900			Mean	:64.3
Service	:148	3rd Qu.	:112494			3rd Qu.	:0.0
(Other)	:474	Max.	:481260			Max.	:2824.0
NA's	:70						

Hours		RISK_Adjustment		TARGET_Adjusted	
Min.	:1.00	Min.	: -1453	Min.	:0.000
1st Qu.	:40.00	1st Qu.	:0	1st Qu.	:0.000
Median	:40.00	Median	:0	Median	:0.000
Mean	:40.43	Mean	:1995	Mean	:0.235
3rd Qu.	:45.00	3rd Qu.	:0	3rd Qu.	:0.000
Max.	:99.00	Max.	:99999	Max.	:1.000

Find: Find Next

Data summary generated.

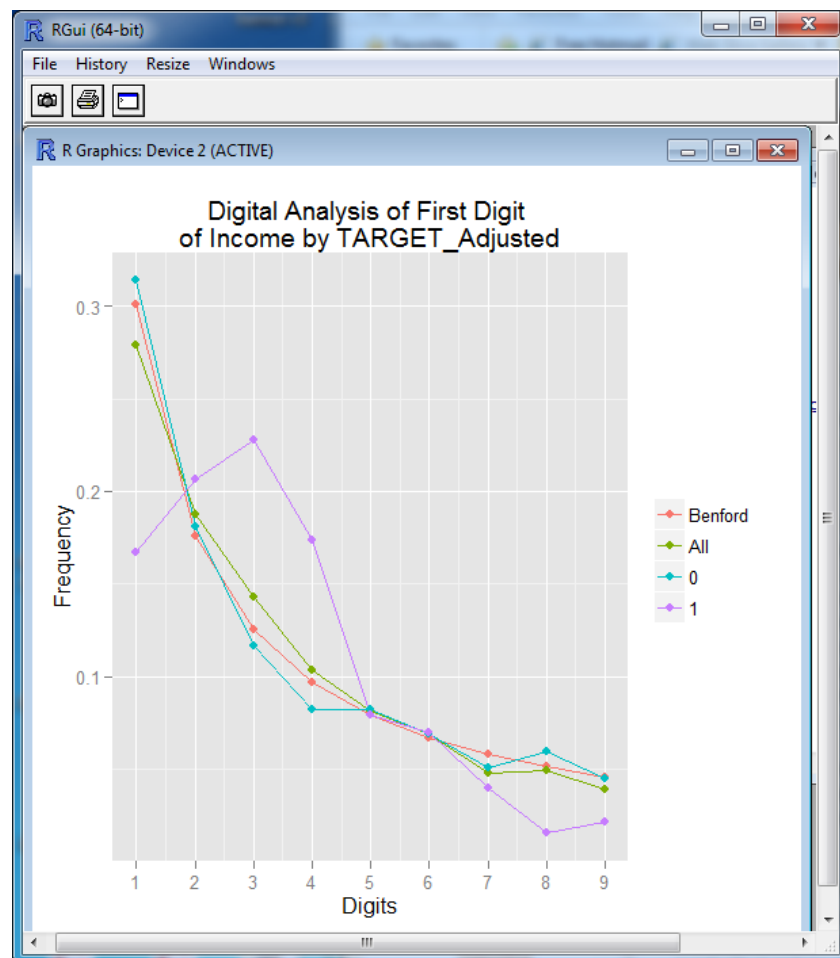
Session 8.10: Benford's Law – Detecting Fraud with Data Mining

In auditing (accounting, financial audits, tax audits), there is a rule called Benford's Law that specifies the frequency of the first digit in almost any financial number. For example, approximately 30% of financial numbers start with the digit 1.

The data set that was loaded describes 2000 income tax audits. To compare the result of income tax audits to Benford's Law:

1. Click on the tab Explore
2. Check the radio button Distributions
3. In the line for Income, check the box under Benford
4. Click Execute

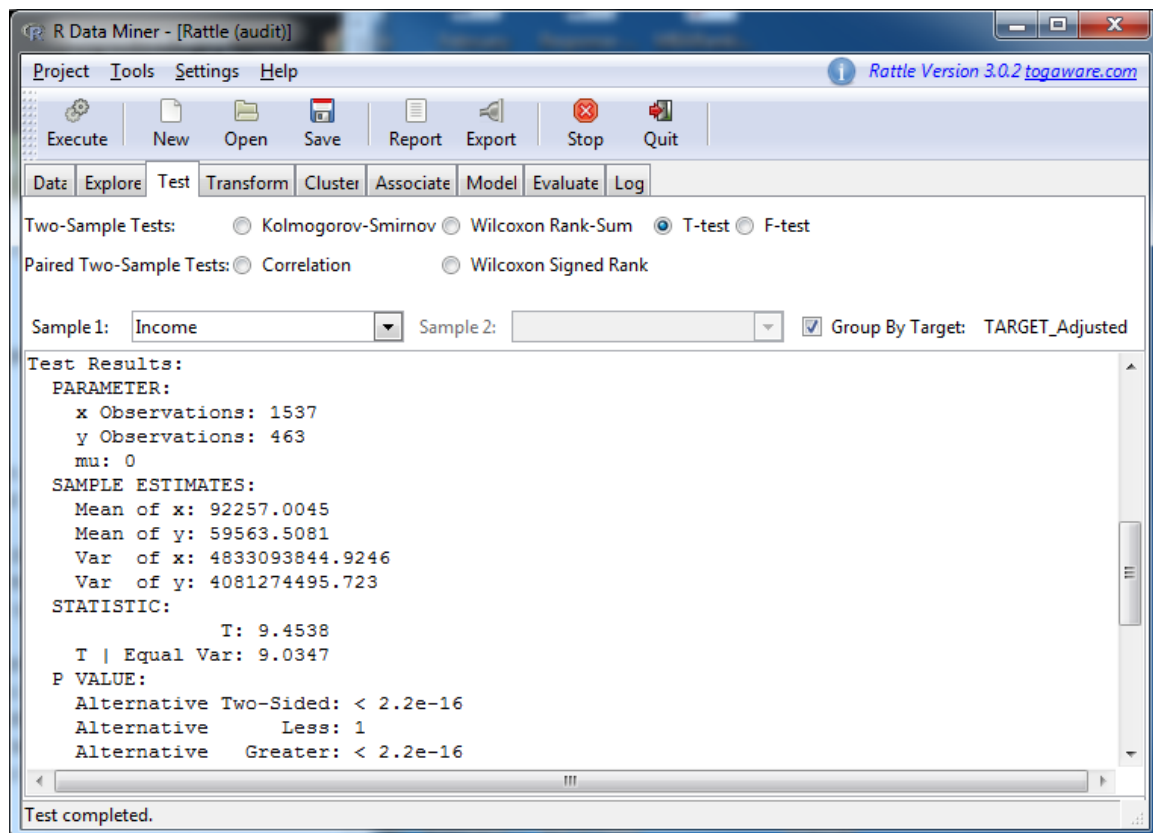
The Benford line is the expected frequency of the first digit of tax payers' income. The All line is the frequency of the first digit of all tax returns filed for 2000 people. The 1 line is the frequency of first digits of income for tax payers who were asked to fix their tax returns. The 0 line is those tax payers were not asked to fix tax returns. The 1 line departs significantly from the Benford line.



Statistical Tests

It appeared from the Benford curve that the distribution of tax violators (coded as 1) were different from non-violators. Let's test the means of the two distributions to see if they are different.

1. Click on the tab Test
2. For Two-Sample Tests, click on the radio button T-test
3. For Sample 1, use the drop down arrow to select Income
4. Click Execute



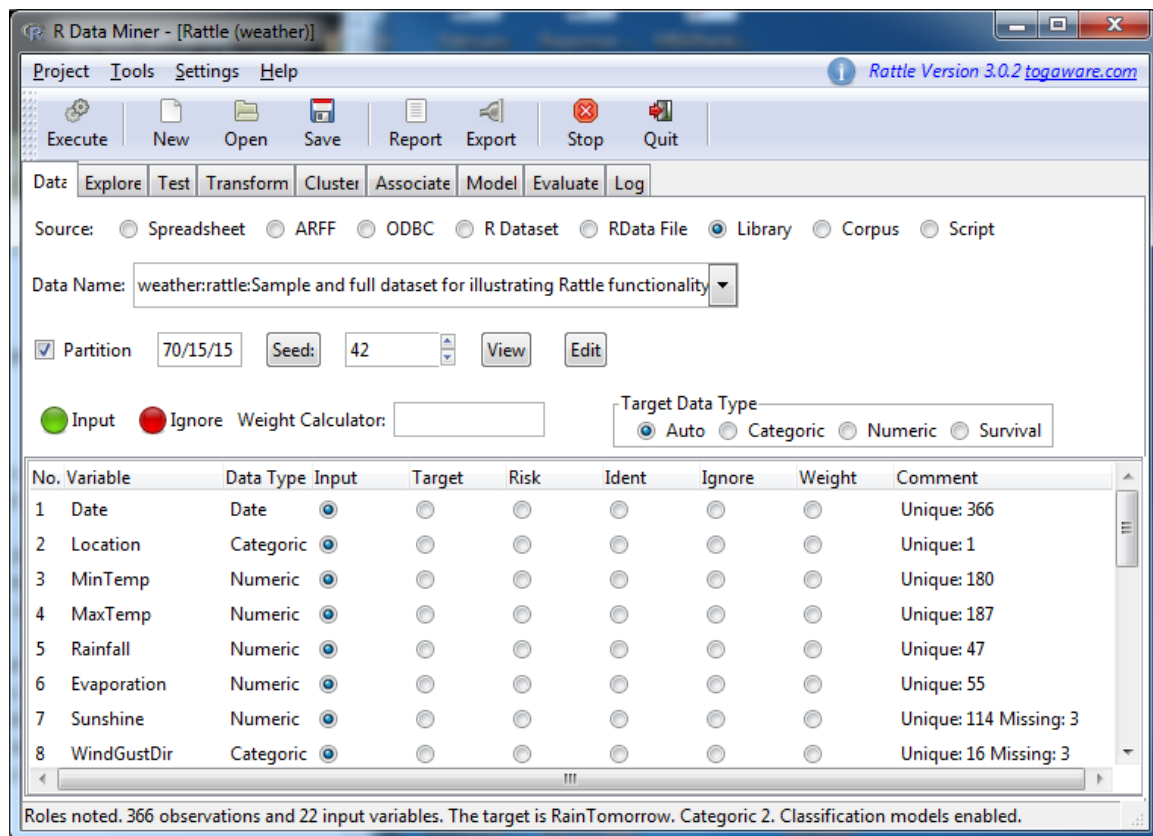
The X observation is for those coded as 0; Y observation for those coded 1. Look at the p-value for the test. Are the two groups different? What is the average income for each group? Which group appears to be misreporting their income more frequently? The higher or lower income group?

Session 8:11: Decision Trees

Loading Data

To load data into R:

1. In the R Data Miner [Rattle] window, click on the Data tab
2. Check that Source indicates Library
3. Next to Data Name, use the drop down menu to select weather: rattle: Sample dataset
4. Click Execute
5. This data set represents historical weather data

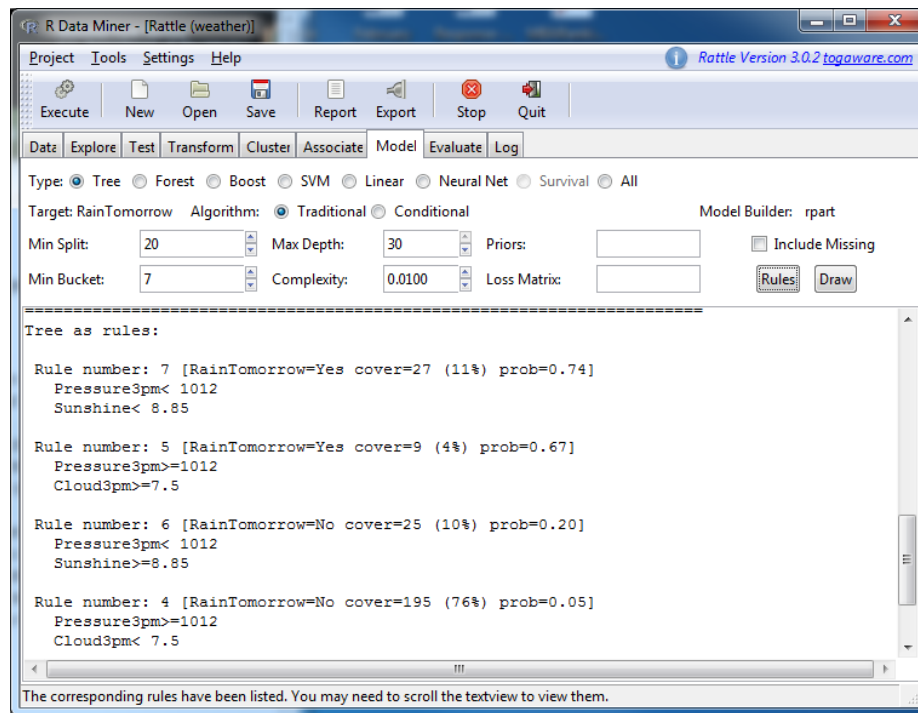


Generating Decision Tree

Create a decision tree to determine how to predict tomorrow's weather.

1. Click on the Model tab
2. Check the radio button Tree
3. Click Execute
4. Click on the button Draw

The resulting decision rules to predict rain tomorrow are:



The decision tree is:

