Correction for Regression Assumption Violations

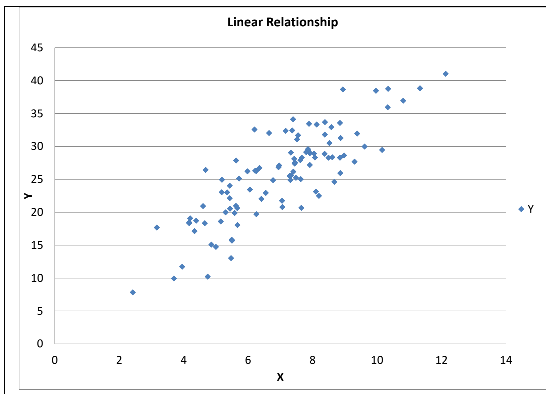
Regression Diagnostics

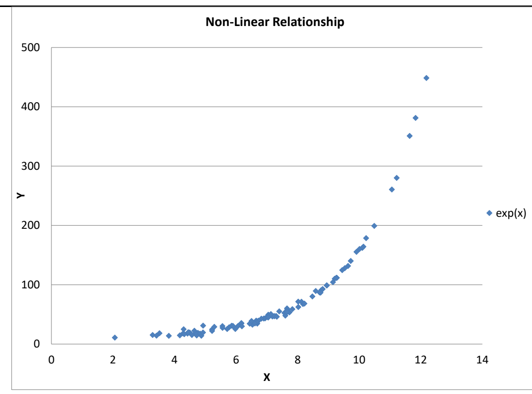
There are several assumptions of linear regression:

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 (violation: non-linearity)

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.





Effects of non-linearity

If the data is not linear, and you use a linear regression, the regression will generate biased (incorrect) coefficients.

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/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).

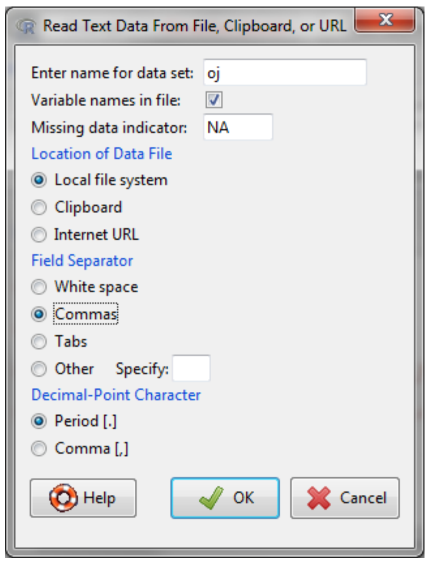
Download Datasets

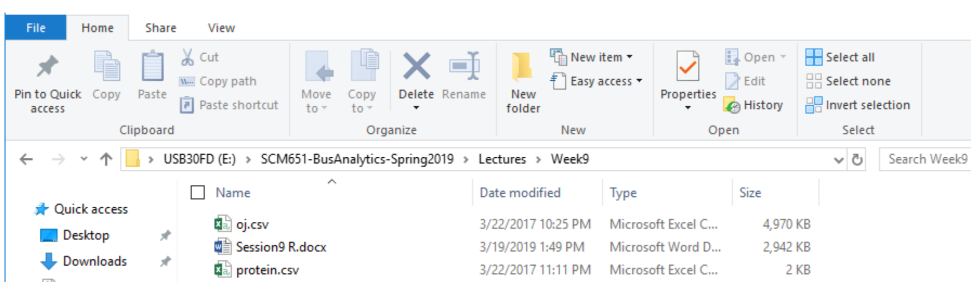
Use the updated oj dataset which includes a column labeled “move”, representing sales.

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

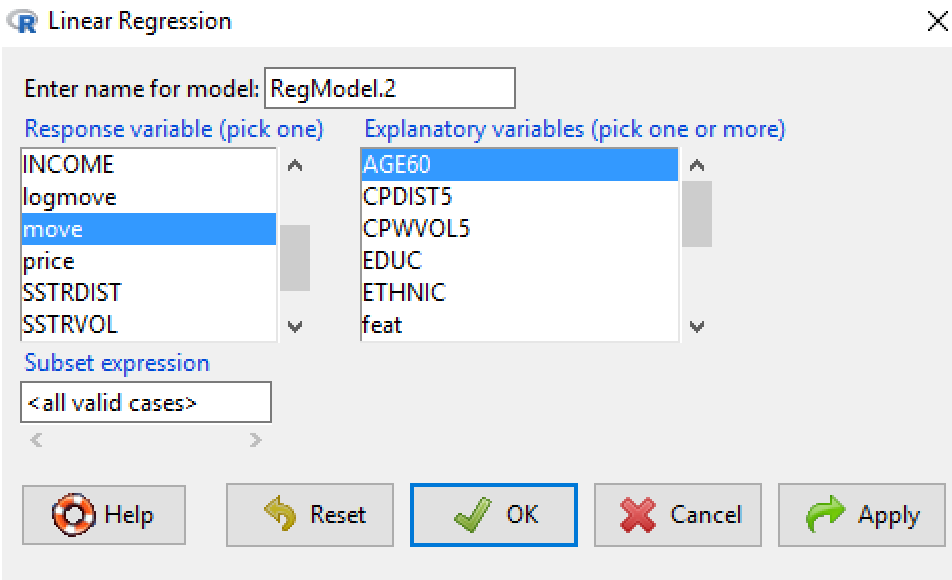


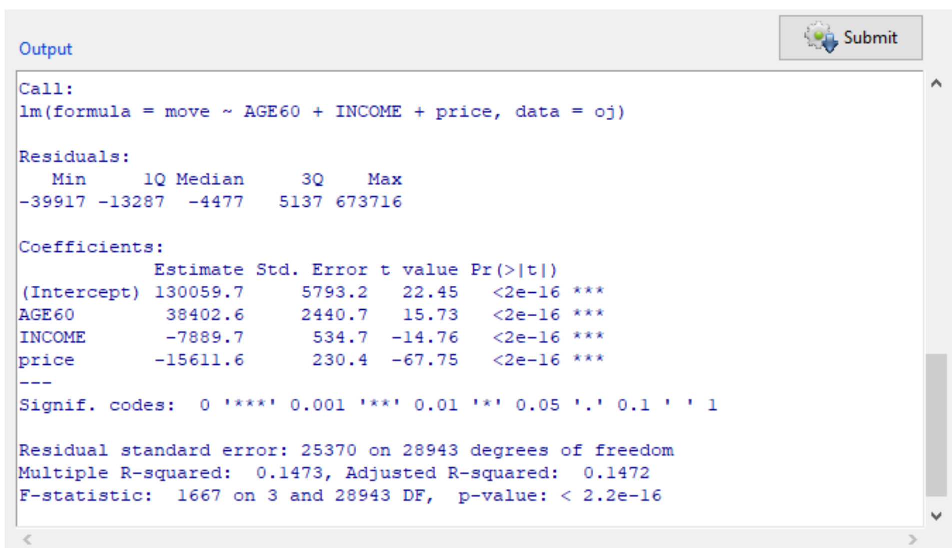


Linear Regression

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 move (which is the volume of products moved or sold)
3. For explanatory variables, hold down the control key and click on AGE60, INCOME, price
4. Click OK





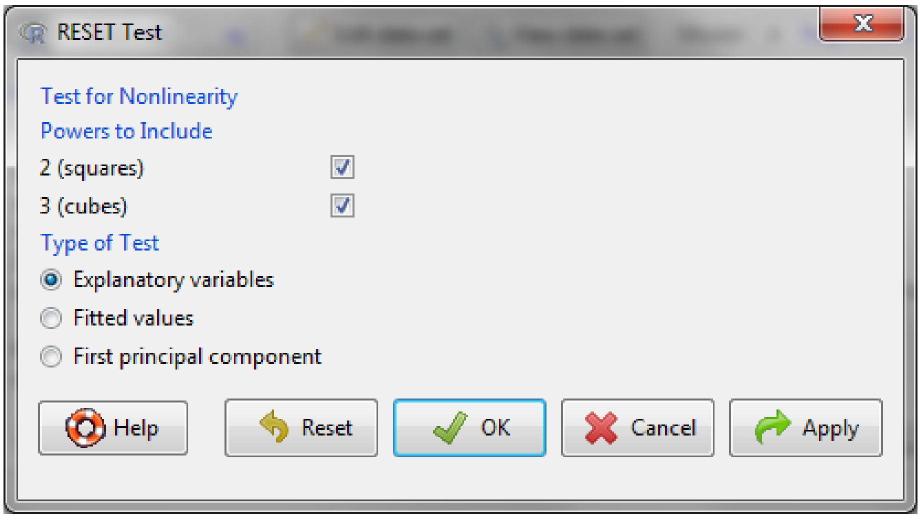

```
Call:
lm(formula = move ~ AGE60 + INCOME + price, data = oj)

Residuals:
   Min     1Q Median     3Q    Max 
-39917 -13287  -4477   5137 673716 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 130059.7     5793.2   22.45   <2e-16 ***
AGE60        38402.6     2440.7   15.73   <2e-16 ***
INCOME       -7889.7      534.7  -14.76   <2e-16 ***
price       -15611.6      230.4  -67.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25370 on 28943 degrees of freedom
Multiple R-squared:  0.1473, Adjusted R-squared:  0.1472 
F-statistic:  1667 on 3 and 28943 DF,  p-value: < 2.2e-16
```

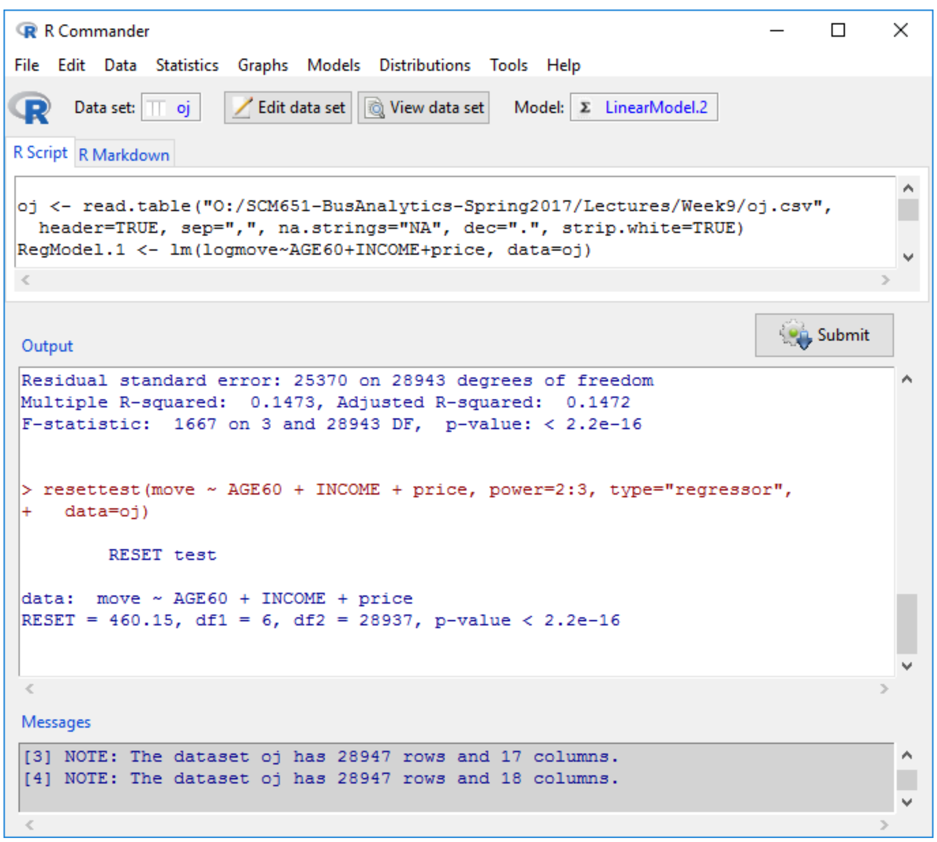
Assumption #1: Linearity

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





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

Solution to Non-Linearity

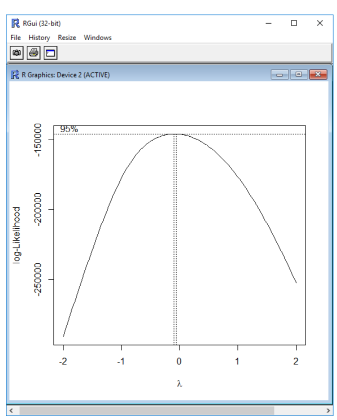
Non-linearity can result from a non-linear dependent (Y) variable or a non-linear independent (X) variable. The Box-Cox technique corrects for non-linearity in Y; the Box-Tidwell technique corrects for non-linearity in X.

Box-Cox correction for the Y-variable

When the non-linearity test indicates that your data is non-linear, first use the Box-Cox technique (George Box & D.R. Cox, 1964) to determine if the Y variable (response variable) is the problem and identify the solution. The solution is usually a transformation.

Install the Box-Cox tools set:

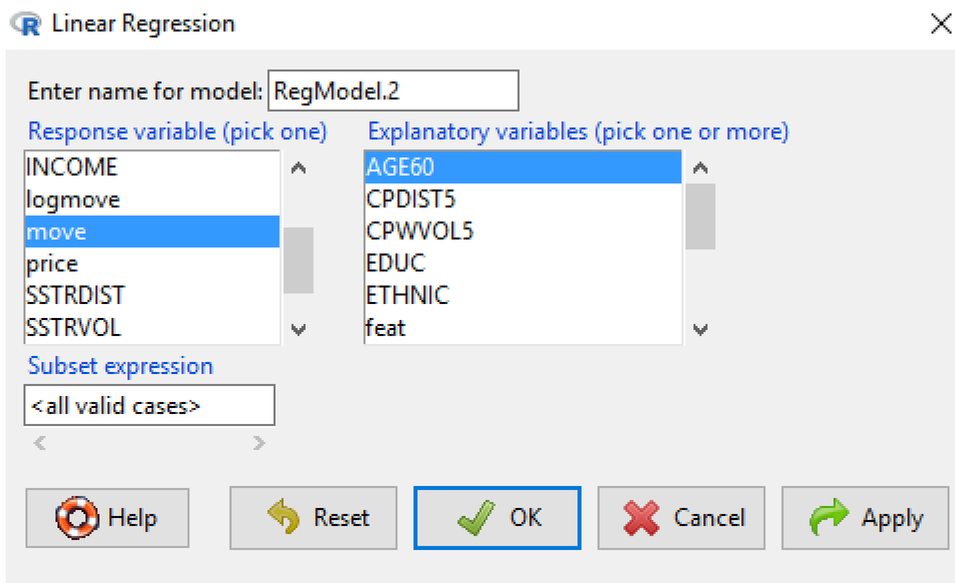
1. In the RGui screen, type:
   install.packages("MASS", dependencies=TRUE)
2. Type library(MASS)
3. Type the following command
   boxcox(lm(move~AGE60+INCOME+price,data=oj),lambda=seq(-2,2,by=.1))
4. The following components are necessary for boxcox
   a. boxcox – name of command
   b. lm – linear model
   c. move~AGE60+INCOME+price – model formulation
   d. data=oj – source of data
   e. lambda=seq(-2,2,by.1) – range of lambda and increment
5. Look on the chart for where lambda peaks; this is the maximum likelihood
6. In this example, it peaks around a lambda value of zero
7. Interpretation: lambda, in general, is the power of Y
   a. 3 means that you should raise Y to the 3 power (Y^3)
   b. 2 means that you should raise Y to the 2 power (Y^2)
   c. 1 means that you should raise Y to the 1 power (Y)
   d. ½ means that you should raise Y to the ½ power ($Y^{1/2}$) or sqrt(Y)
   e. 0 means that you should transform Y by taking the logarithm (log(Y))
   f. -½ means that you should raise Y to the -½ power ($Y^{-1/2}$) or 1/sqrt(Y)
   g. -1 means that you should transform Y by raising it to the -1 power (1/Y)
   h. -2 means that you should transform Y by raising it to the -2 power ($1/Y^2$)
   i. -3 means that you should transform Y by raising it to the -3 power ($1/Y^3$)
8. What should the transformation of our variable “move” be?

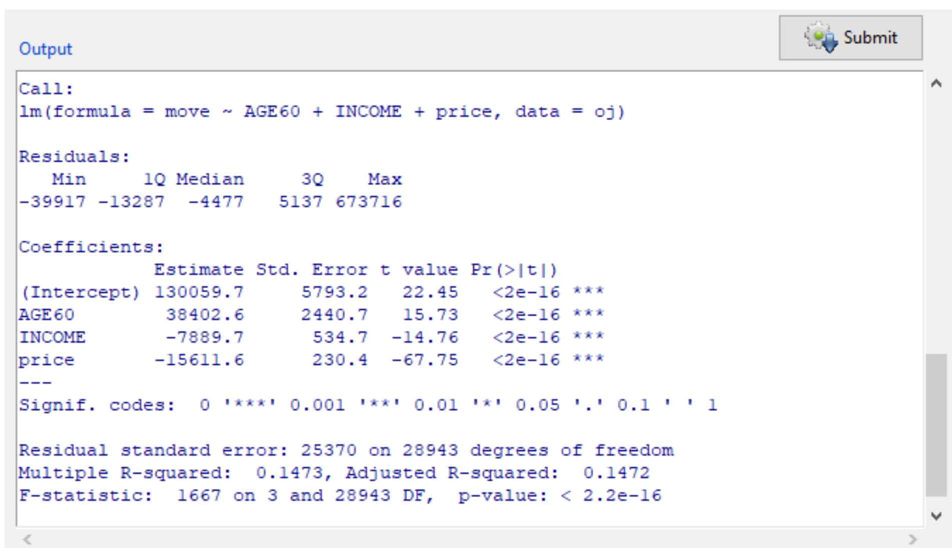


Testing the equation after correction for non-linearity in Y

Let's compare the regression results, before and after the Box-Cox correction for non-linearity in the Y-variable.

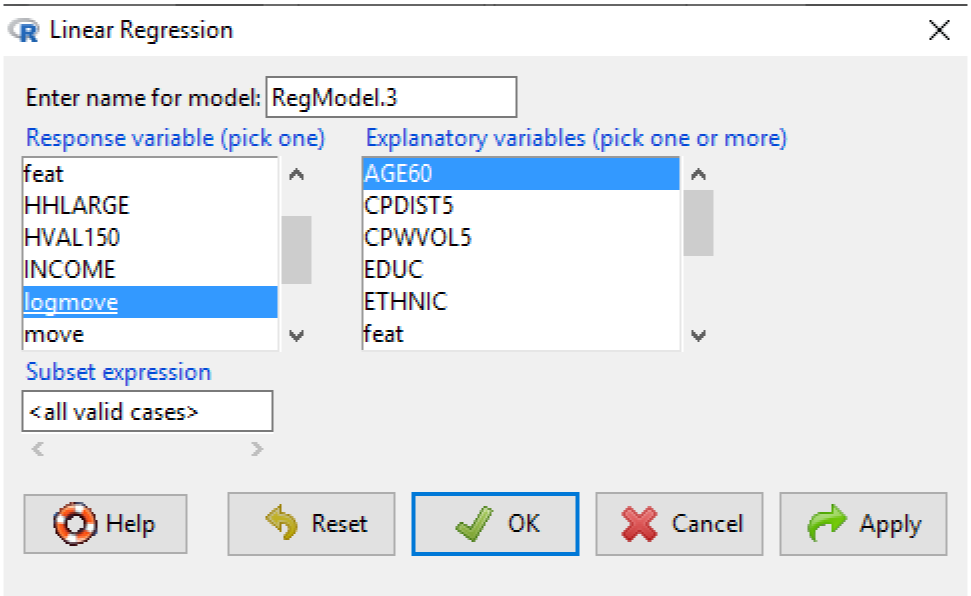
1. Click on Statistics, Fit Models, Linear Regression
2. For response variable, click on move (which is the volume of products moved or sold)
3. For explanatory variables, hold down the control key and click on AGE60, INCOME, price
4. Click OK

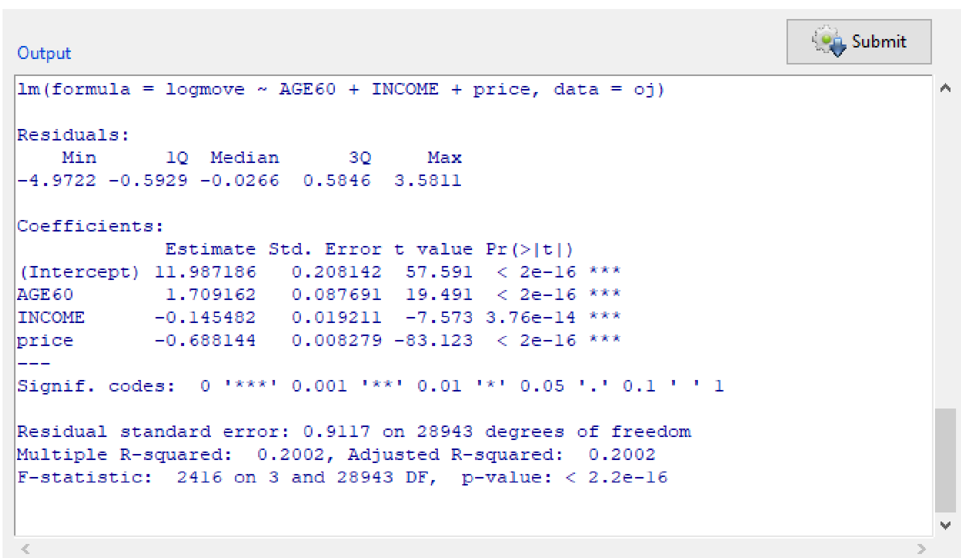





```
Call:
lm(formula = move ~ AGE60 + INCOME + price, data = oj)

Residuals:
   Min     1Q Median     3Q    Max
-39917 -13287  -4477   5137 673716

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 130059.7     5793.2   22.45   <2e-16 ***
AGE60        38402.6     2440.7   15.73   <2e-16 ***
INCOME       -7889.7      534.7  -14.76   <2e-16 ***
price       -15611.6      230.4  -67.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25370 on 28943 degrees of freedom
Multiple R-squared:  0.1473, Adjusted R-squared:  0.1472
F-statistic:  1667 on 3 and 28943 DF,  p-value: < 2.2e-16
```

5. The R-squared for move~AGE60+INCOME+price is 0.1472
6. Next, run the linear regression for logmove instead of move
7. Click on Statistics, Fit Models, Linear Regression
8. For response variable, click on logmove
9. For explanatory variables, hold down the control key and click on AGE60, INCOME, price
10. Click OK





11. The R-squared for logmove~AGE60+INCOME+price is 0.2002

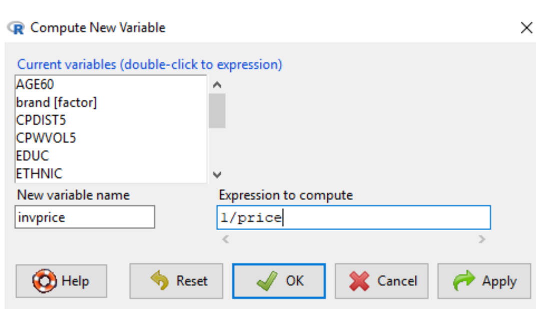
Box-Tidwell correction for the X-variable

After correcting for any non-linearity in the Y-variable, next correct for non-linearity in the X-variable. The Box-Tidwell technique (George Box and P.W. Tidwell (1962)) corrects for non-linear independent variables.

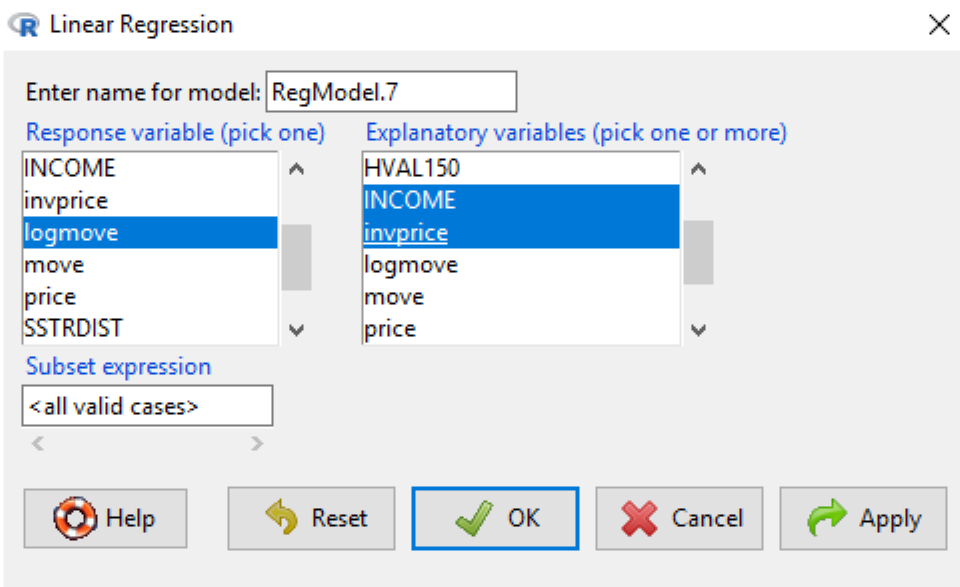
1. In the RGui screen, type:
   install.packages("car", dependencies=TRUE)
2. Type library(car)
3. Type the following command
   boxTidwell(logmove~price, data=oj, tol=0.001, max.iter=25)
4. The following components are necessary for boxcox
   a. boxTidwell – name of command
   b. logmove~AGE60+INCOME+price – model formulation
   c. data=oj – source of data
   d. tol – tolerance level, stopping threshold
   e. max.iter=25 – maximum number of iterations for the maximum likelihood

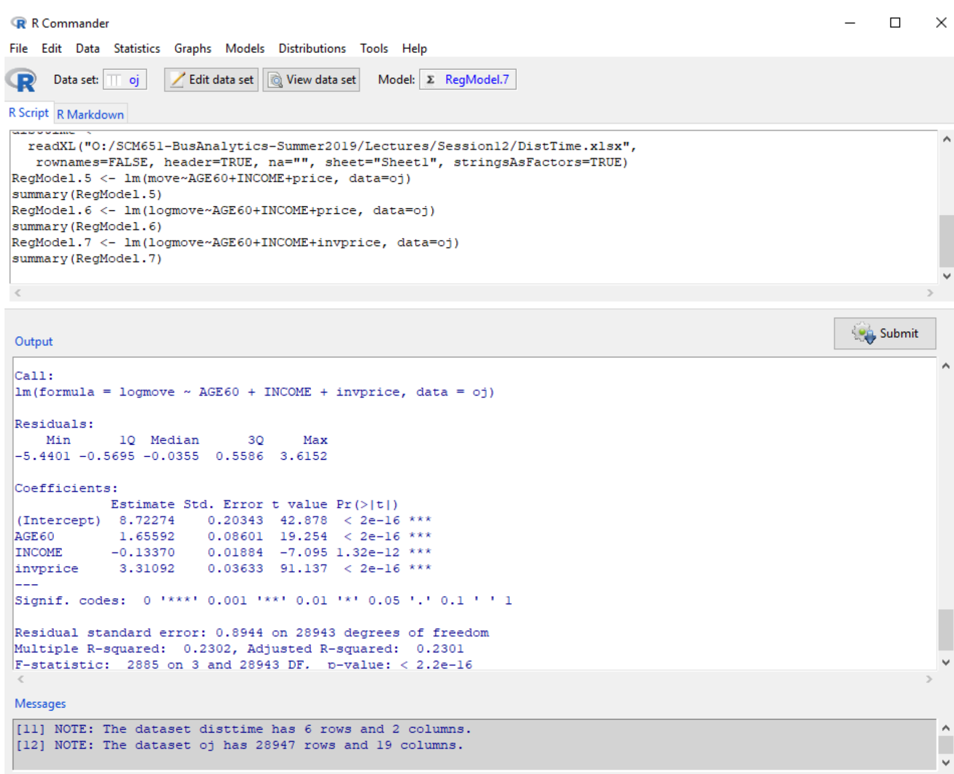
```
MLE of lambda      Score Statistic (z)      Pr(>|z|)
    -1.0341                34.858           < 2.2e-16 ***
---
Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1
iterations =  4
```

5. Interpretation
   a. 3 means that you should raise Y to the 3 power (Y^3)
   b. 2 means that you should raise Y to the 2 power (Y^2)
   c. 1 means that you should raise Y to the 1 power (Y)
   d. ½ means that you should raise Y to the ½ power ($Y^{1/2}$) or sqrt(Y)
   e. 0 means that you should transform Y by taking the logarithm (log(Y))
   f. -½ means that you should raise Y to the -½ power ($Y^{-1/2}$) or 1/sqrt(Y)
   g. -1 means that you should transform Y by raising it to the -1 power (1/Y)
   h. -2 means that you should transform Y by raising it to the -2 power ($1/Y^2$)
   i. -3 means that you should transform Y by raising it to the -3 power ($1/Y^3$)
6. What should the transformation of our variable “price” be?
7. We need to create a new variable 1/X
8. In Rcmdr, click on Data, Manage variables in active data set, Compute new variable
9. For New variable name, enter invprice (for inverse of price)
10. In Expression to compute, enter 1/price
11. Click OK



12. Next, run the linear regression for logmove~AGE60+INCOME+invprice
13. Click on Statistics, Fit Models, Linear Regression
14. For response variable, click on logmove
15. For explanatory variables, hold down the control key and click on AGE60, INCOME, invprice
16. Click OK



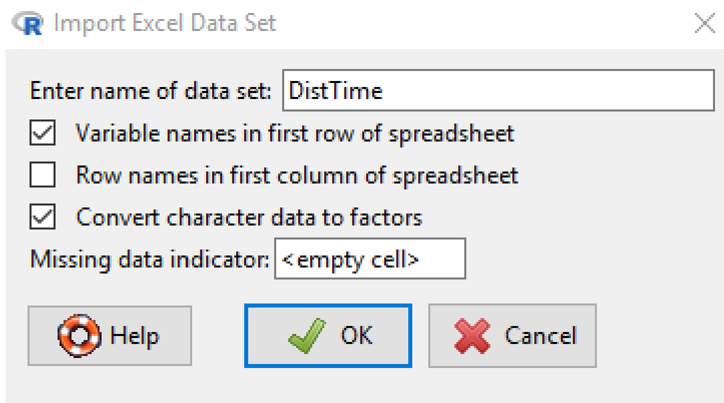


17. The R-squared for logmove~AGE60+INCOME+invprice is 0.2302

Scientific Example

Datasets do not need to be large to find interesting results. Load the following data with only six observations, perform a regression of distance on time, then use Box-Cox to find the form of the equation.

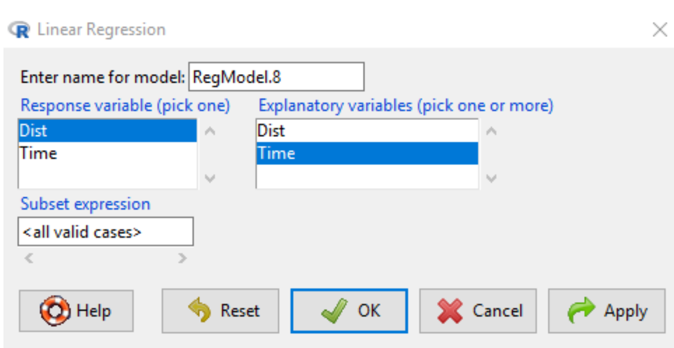
1. Click on Data at the top of the screen
2. Click on Import Data > From Excel file …
3. Enter the name that you would like to use for this data set; type in DistTime
4. Click OK

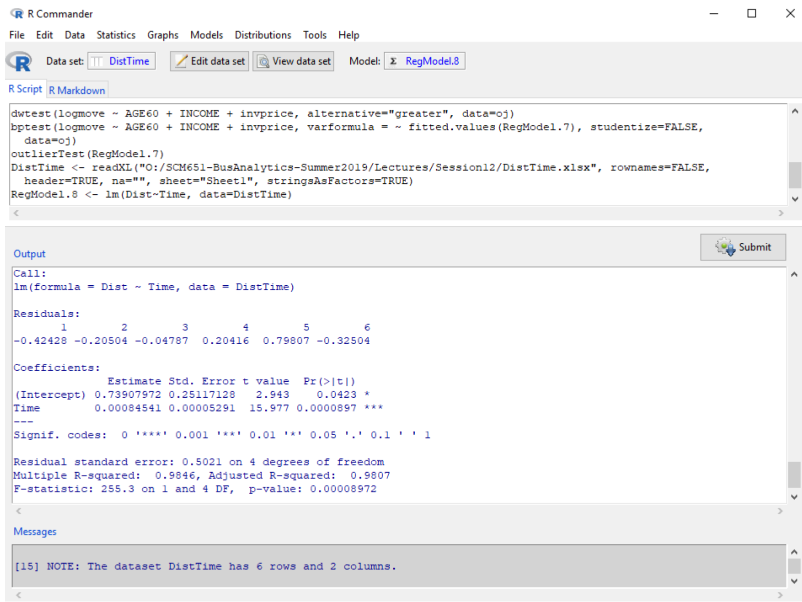


5. Click on the DistTime file, then Open
6. Click on View data set to view the six data observations

	Dist	Time
1	0.389	87.77
2	0.724	224.70
3	1.000	365.25
4	1.524	686.95
5	5.200	4332.62
6	9.510	10759.20

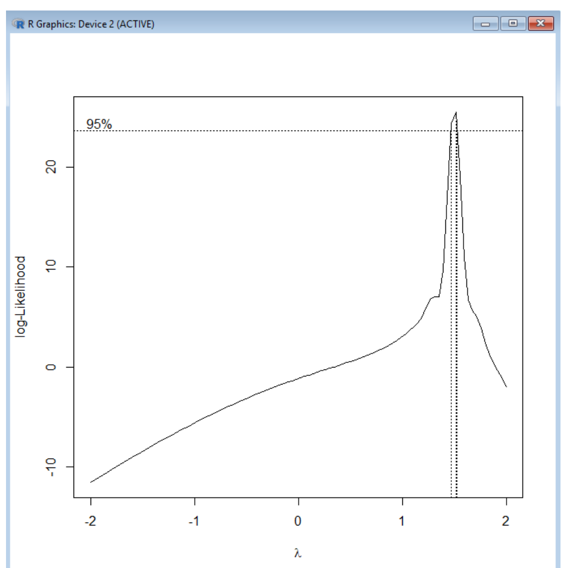
7. Run the regression by clicking on Statistics, Fit models, Linear regression
8. Click on Dist for the Response variable (Y) and Time for the Explanatory variable (X)
9. Click OK





10. Next run Box-Cox

    boxcox(lm(Dist~Time,data=DistTime),lambda=seq(-2,2,by=.1))



11. The lambda is 1.5, or written as a fraction, 3/2
12. The equation then is

    $Dist^{3/2} = \beta * Time$

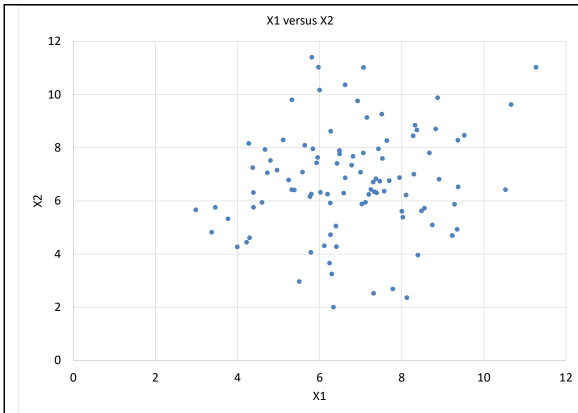
13. Taking the square of each side, we get

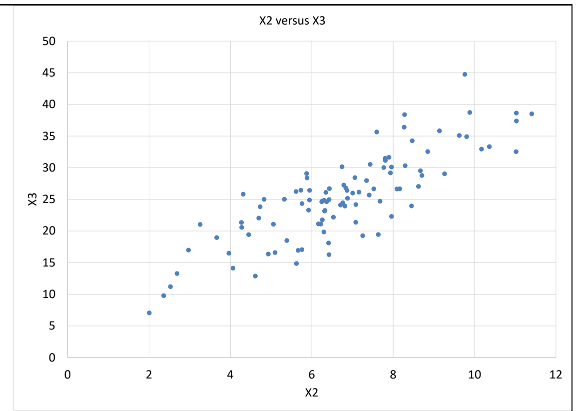
    $Dist^{3} = \beta' * Time^{2}$

14. This is Kepler's Third Law of Planetary Motion (Johannes Kepler 1619)

Assumption #2: The X variable are not correlated (violation: multicollinearity)

When including more than on 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.





Effects of multi-collinearity

If the independent variables (x-variables are correlated, the sign +/- will be reversed on one of the coefficients.

Test for Multi-collinearity

The Variance Inflation Factor test of correlated explanatory variables

Solution to Multi-collinearity

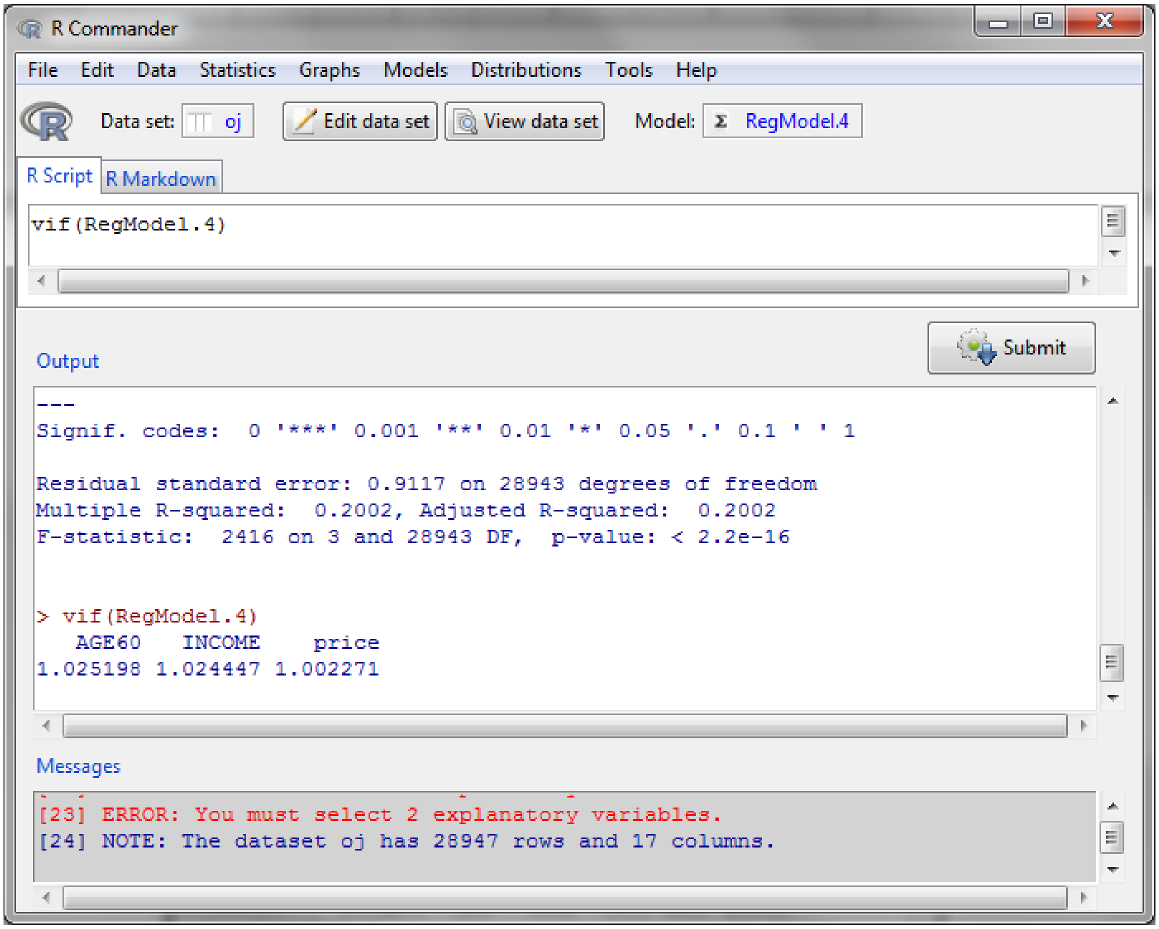
If two or more variables are collinear (highly correlated), there are three solutions:

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

Variance Inflation Factor test of correlated explanatory variables

To calculate the Variance Inflation Factor:

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



2. 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.

Correction for Multi-collinearity: Factor Analysis

Factor analysis identifies how many unique concepts are captured in the variables in your data.

Install

To install the modules, we need the psych library. Enter the following commands.

```
install.packages("psych",dependencies=TRUE)

library(psych)
```

Download Datasets

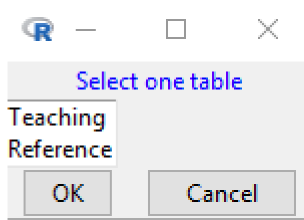
The teaching preference spreadsheet is on the G:drive in Session 13 and on BlackBoard. This data set is from Charles Zaiontz, from the website:

http://www.real-statistics.com/multivariate-statistics/factor-analysis/factor-analysis-example/

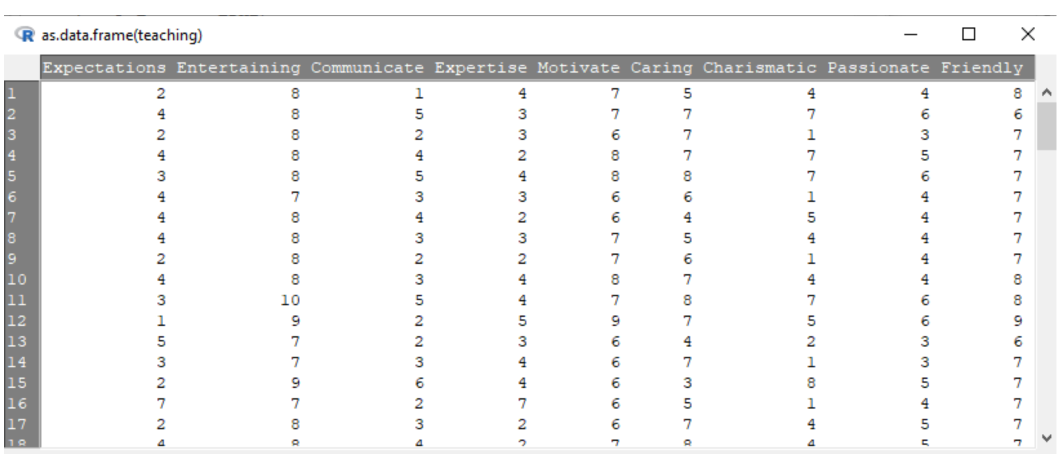
Loading Data

To load data into R:

1. Click on Data at the top of the 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 teaching, then OK
4. Click on the Teaching file, then Open
5. In this example, the Teaching spreadsheet has two worksheets, Teaching and Reference; click on Reference, then OK



6. In Rcmdr, click on View data



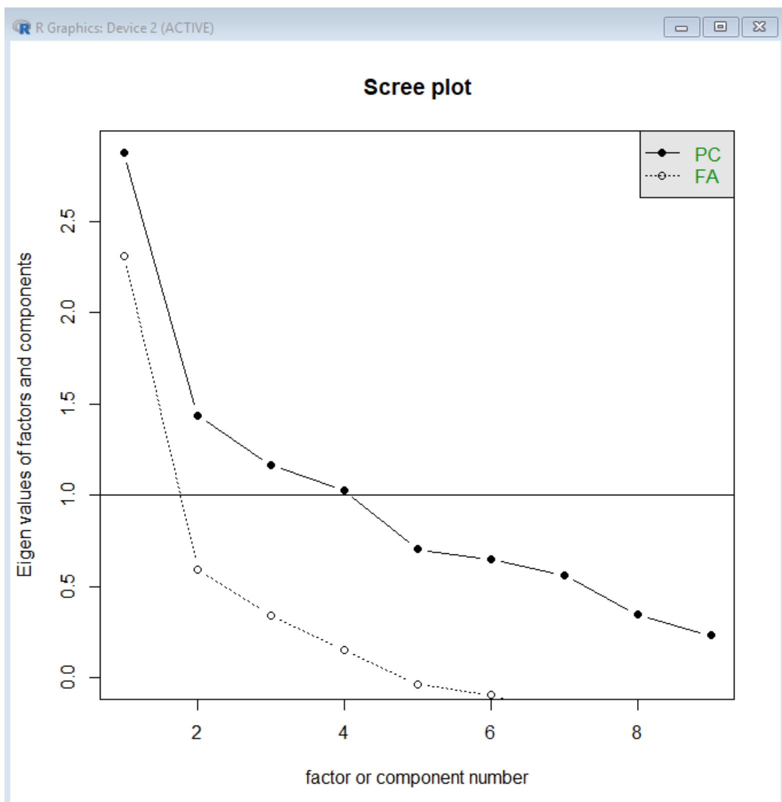
	Expectations	Entertaining	Communicate	Expertise	Motivate	Caring	Charismatic	Passionate	Friendly
1	2	8	1	4	7	5	4	4	8
2	4	8	5	3	7	7	7	6	6
3	2	8	2	3	6	7	1	3	7
4	4	8	4	2	8	7	7	5	7
5	3	8	5	4	8	8	7	6	7
6	4	7	3	3	6	6	1	4	7
7	4	8	4	2	6	4	5	4	7
8	4	8	3	3	7	5	4	4	7
9	2	8	2	2	7	6	1	4	7
10	4	8	3	4	8	7	4	4	8
11	3	10	5	4	7	8	7	6	8
12	1	9	2	5	9	7	5	6	9
13	5	7	2	3	6	4	2	3	6
14	3	7	3	4	6	7	1	3	7
15	2	9	6	4	6	3	8	5	7
16	7	7	2	7	6	5	1	4	7
17	2	8	3	2	6	7	4	5	7

7. This data represents what students feel are important characteristics for an instructor.
8. The characteristics are:

Expectations	Setting high expectations for the students
Entertaining	Entertaining
Communicate	Able to communicate effectively
Expertise	Having expertise in their subject
Motivate	Able to motivate
Caring	Caring
Charismatic	Charismatic
Passion	Having a passion for teaching
Friendly	Friendly and easy-going

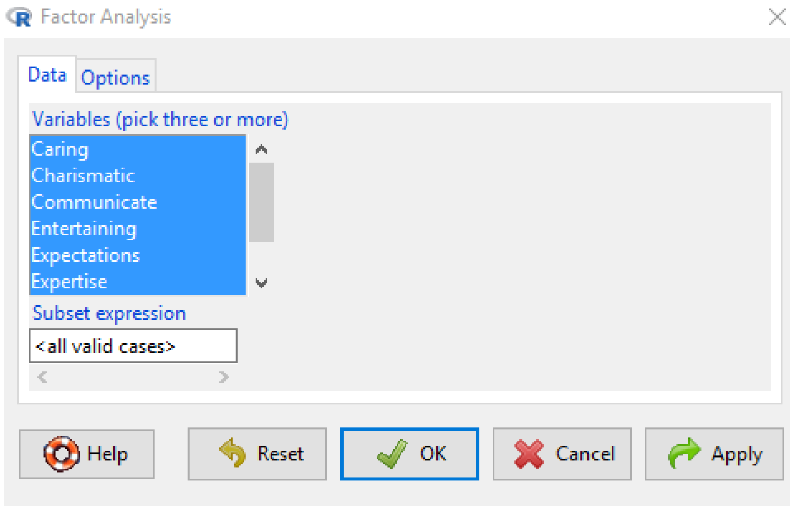
9. A screeplot will indicate how the measures above collapse into unique factors. Type the command:

scree(teaching)

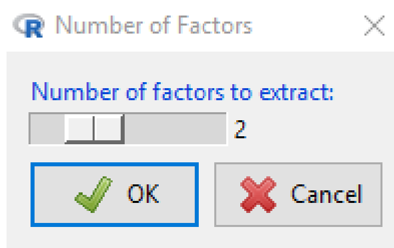


10. There are two techniques represented above, Principal Component Analysis (PC) and Factor Analysis (FA). The left side of the chart indicates Eigenvalues. The Kaiser criterion (Kaiser, 1960) recommends that the number of principal components or factors is the number of dots above the 1.0 line (eigenvalue > 1.0)
11. Now determine exactly how many factors we need.
12. Click on Statistics, Dimensional Analysis, Factor Analysis

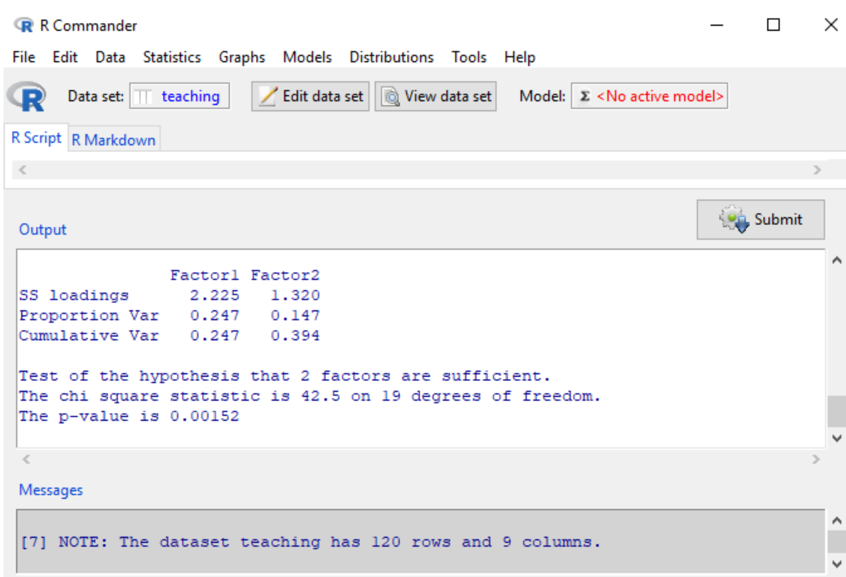
13. Highlight all the variables by holding down the control key and clicking each variable (or click on the first, hold the shift button down, then click on the last variable).



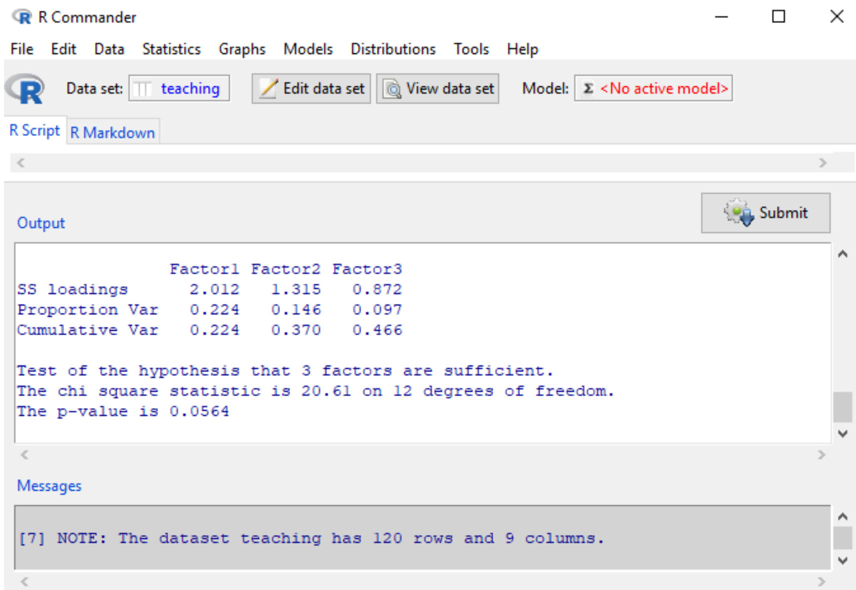
14. When asked for number of factors to extract, change to 2, then OK



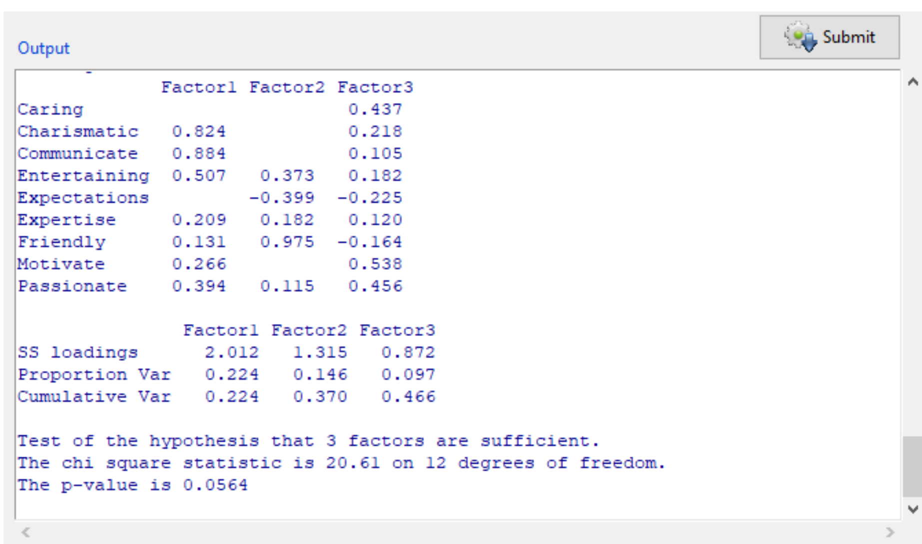
15. The hypothesis is that 2 factors are sufficient. If p<0.05, then 2 are not sufficient and we need to test 3 factors. In this case, p= 0.00152, so 2 is not sufficient



16. Click on Statistics, Dimensional Analysis, Factor Analysis, then OK
17. Change number of factors to 3, then OK



18. Now, p=0.0564. Therefore, 3 factors are sufficient. This means that the original variables can be collapsed into three concepts.
19. Click on Statistics, Dimensional Analysis, Factor Analysis
20. Click on the Options tab, and check the button for Regression method, then OK
21. Set the Number of factors to extract to 3, the OK



22. There are three factors. The numbers in the columns are loadings, which measure how much the original variable influences the factor. Which variables have a load of more than 0.500 for factor 1? Factor 2? Factor 3?
23. How would you interpret Factors 1, 2, 3?
24. In Rcmdr, click on View data; scroll to the right

25. The three new variables are F1, F2, F3, our new factors, calculated from the original variables
26. These are the variables that you would use in a regression
27. By selecting the Verimax rotation, the factors F1, F2, F3 will not be correlated, so multi-collinearity in regression will not be a problem

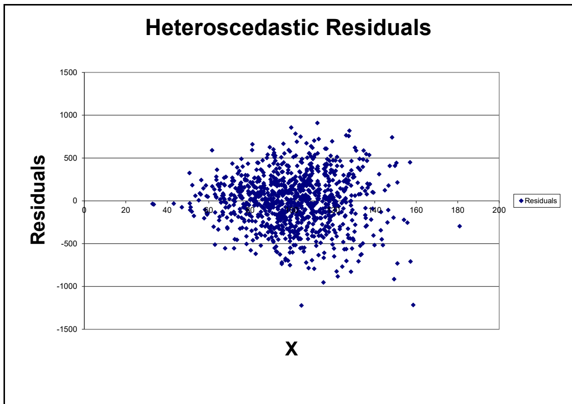
as.data.frame(teaching)

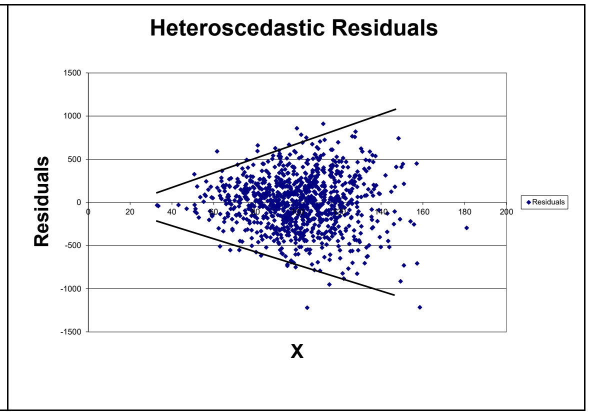
	ate	Caring	Charismatic	Passionate	Friendly	F1	F2	F3
1	7	5	4	4	8	-0.979186112	1.3314709	-0.0952331548
2	7	7	7	6	6	0.638199552	-2.1508267	1.1183800856
3	6	7	1	3	7	-1.211641842	-0.3836267	-0.1732023856
4	8	7	7	5	7	0.354300389	-0.4456160	0.7441874343
5	8	8	7	6	7	0.674046227	-0.3875633	1.3351334063
6	6	6	1	4	7	-0.840939374	-0.5277226	-0.6827543492
7	6	4	5	4	7	0.232887389	-0.7037285	-0.9052153492
8	7	5	4	4	7	-0.311156035	-0.5095170	-0.1858942397
9	7	6	1	4	7	-1.208444079	-0.3097434	0.2744069275
10	8	7	4	4	8	-0.262721865	1.2553476	0.0632830297
11	7	8	7	6	8	1.087519721	1.2134900	0.8288194046
12	9	7	5	6	9	-0.360678041	3.2300026	1.3726750114
13	6	4	2	3	6	-1.100657137	-2.2379099	-0.7741736478
14	6	7	1	3	7	-0.919323213	-0.5028850	-0.6086545203
15	6	3	8	5	7	1.460040232	-0.7688906	-0.3686660170
16	6	5	1	4	7	-0.982010664	-0.5508256	-0.9338108103
17	6	7	4	5	7	-0.438149022	-0.4046231	0.3277535562
18	7	8	4	5	7	-0.049297354	-0.4412491	0.4509694507
19	5	4	6	4	9	-0.140566651	2.6806387	-1.6971096907
20	9	9	8	5	7	1.092871265	-0.4217942	1.4765586667
21	6	3	3	6	8	0.004066874	1.0970168	-0.7124576402
22	6	4	4	3	7	0.505367626	-0.8368609	-1.4700973383
23	7	7	7	5	7	0.522726036	-0.4815022	0.6252276056
24	6	5	7	4	8	0.262896501	1.0171167	-0.9518546698
25	6	8	4	5	8	-0.662558349	1.3495923	0.2735379225
26	8	7	9	6	7	1.392385482	-0.5428580	0.9883872577
27	7	5	5	5	7	-0.256082157	-0.4091707	0.4363734323
28	6	8	7	4	8	-0.635187538	1.3360781	0.2063587559
29	7	7	1	5	7	-0.706798020	-0.3794901	0.2869615162
30	6	5	5	5	8	-0.442950963	1.2280982	-0.2809939752

Assumption #3a: The error terms do not have constant variance (violation: 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.





Effects of heteroscedasticity

If the residuals are heteroscedastic, the standard errors and p-values will be incorrect.

Test for Heteroscedasticity

Breusch-Pagan test of heteroscedasticity

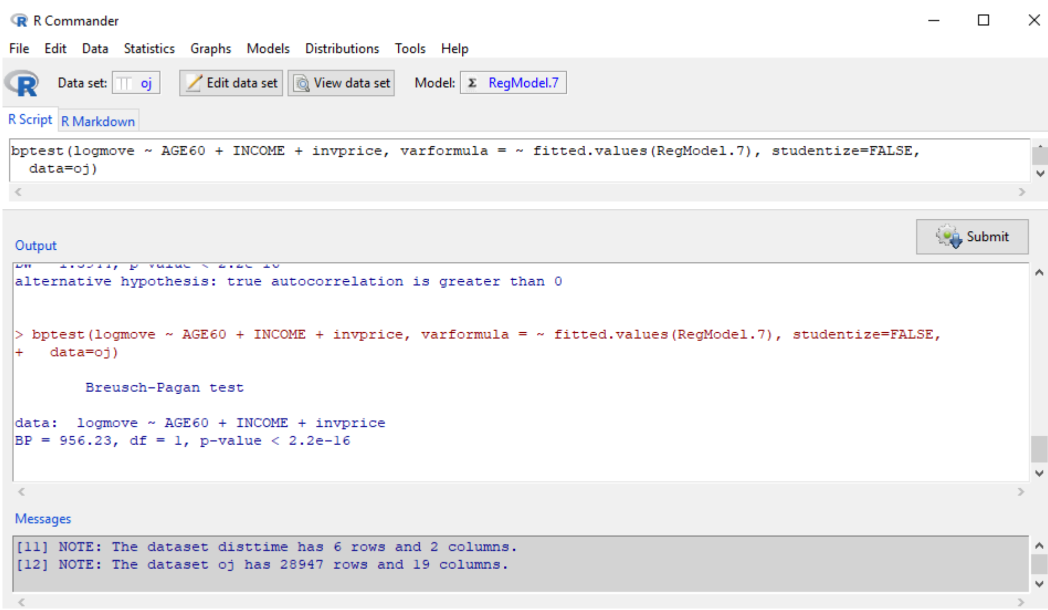
Solution to Heteroscedasticity

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.

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, invprice
3. Click on OK

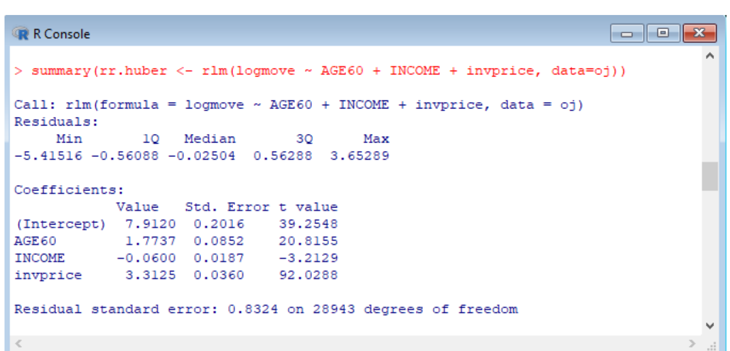


4. If the p-value is less than 0.05, then there is a problem with heteroscedasticity. Generally, this is a sign that the equation is non-linear.
5. If you have already corrected for non-linearity, then more sophisticated techniques (robust Huber regression for heteroscedasticity) must be used. Install MASS if not already installed.

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

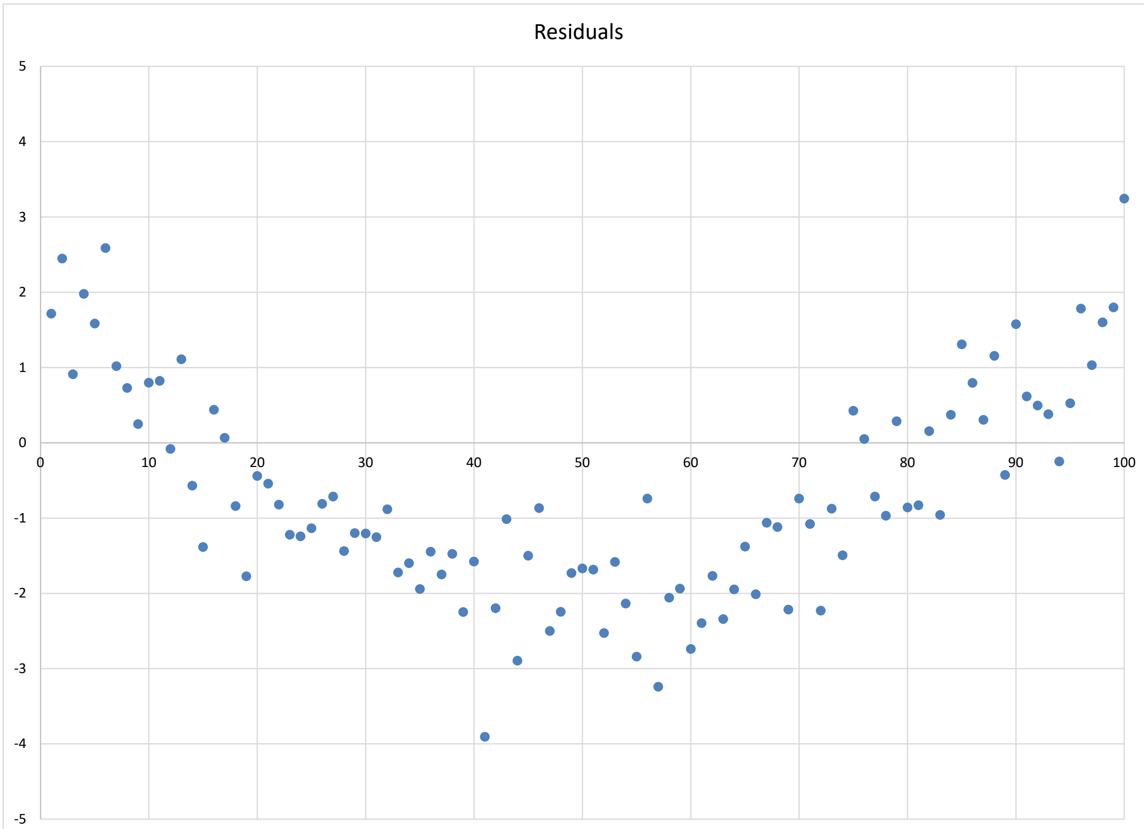
   library(MASS)

   summary(rr.huber <- rlm(logmove ~ AGE60 + INCOME + invprice, data=oj))



Assumption #3b: The residuals are not correlated (violation: 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.



Effects of serial correlation

If the residuals have serial correlation, the standard errors will be underestimated and the p-values will be incorrect

Test for Serial Correlation

Durbin-Watson test of serial correlation

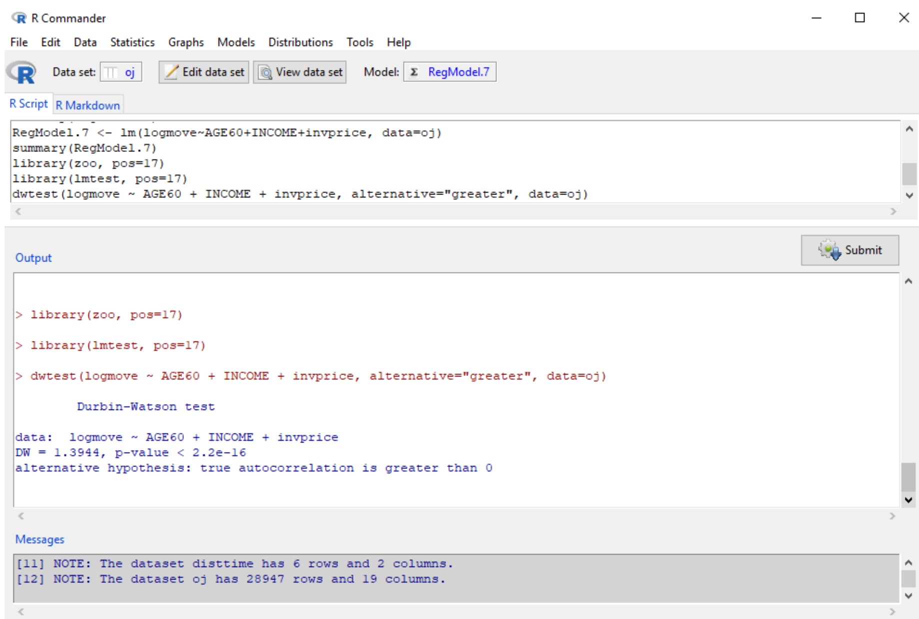
Solution to Serial Correlation

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

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
2. Select rho > 0, then OK

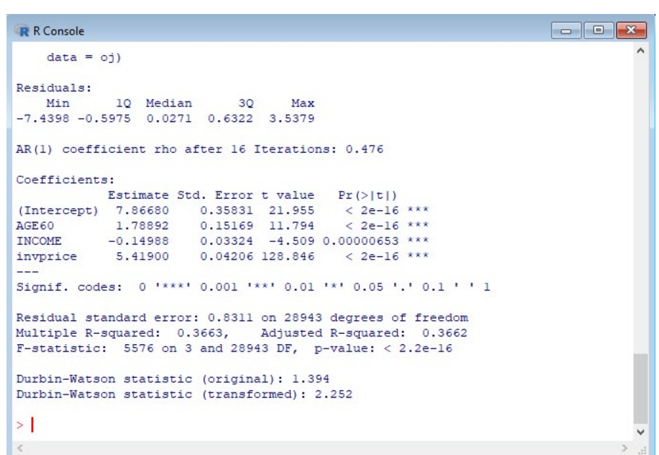


3. If the p-value is less than 0.05, there is a problem with serial correlation.

Correction for Serial Correlation

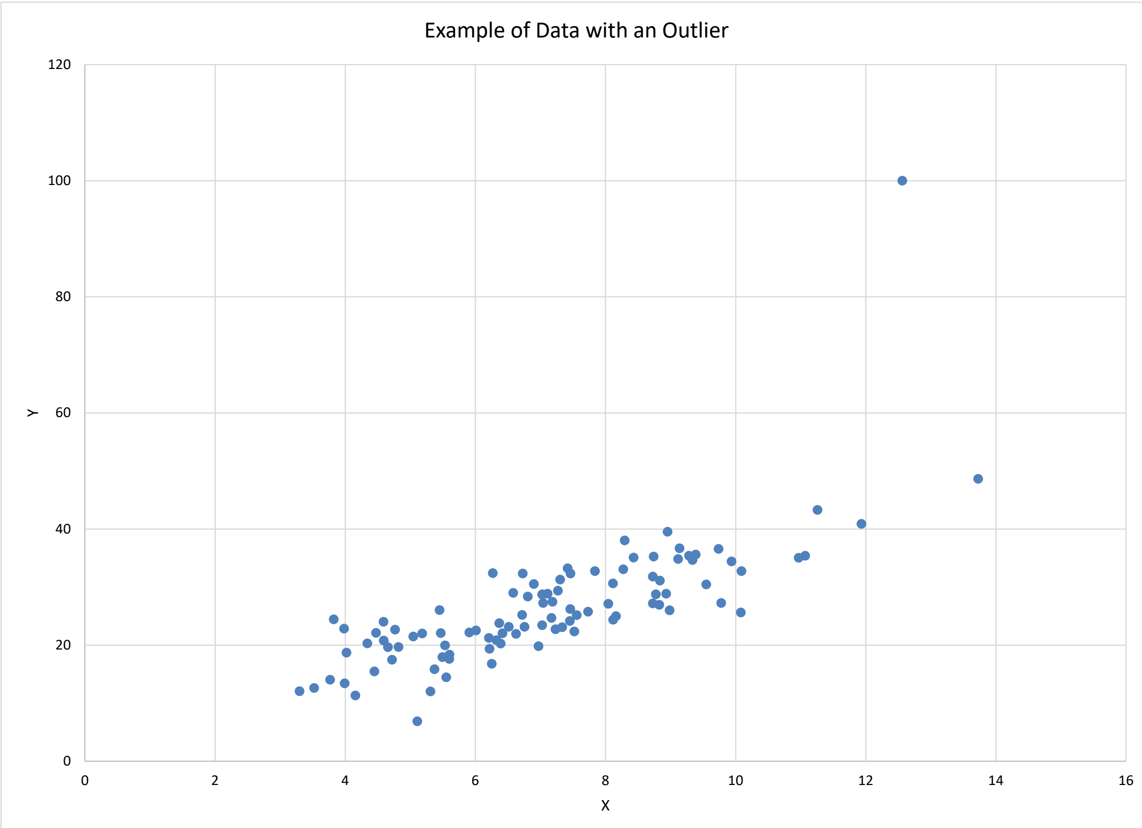
*There are several techniques for correction of serial correlation, including Cochrane-Orcutt (*Cochrane, D.; Orcutt, G. H. (1949)), Prais-Winsten (Prais, S. J.; Winsten, C. B. (1954)) and rho differencing.

```
install.packages("prais",dependencies=TRUE)
library(prais)
pw <- prais_winsten(logmove ~ AGE60 + INCOME + invprice, data=oj)
summary(pw)
```



Assumption #3c: There are no outliers (violation: 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.



Effect of outliers

If outliers exist in the data, the coefficients (slopes) will be incorrect.

Test for Outliers

Bonferroni outlier test

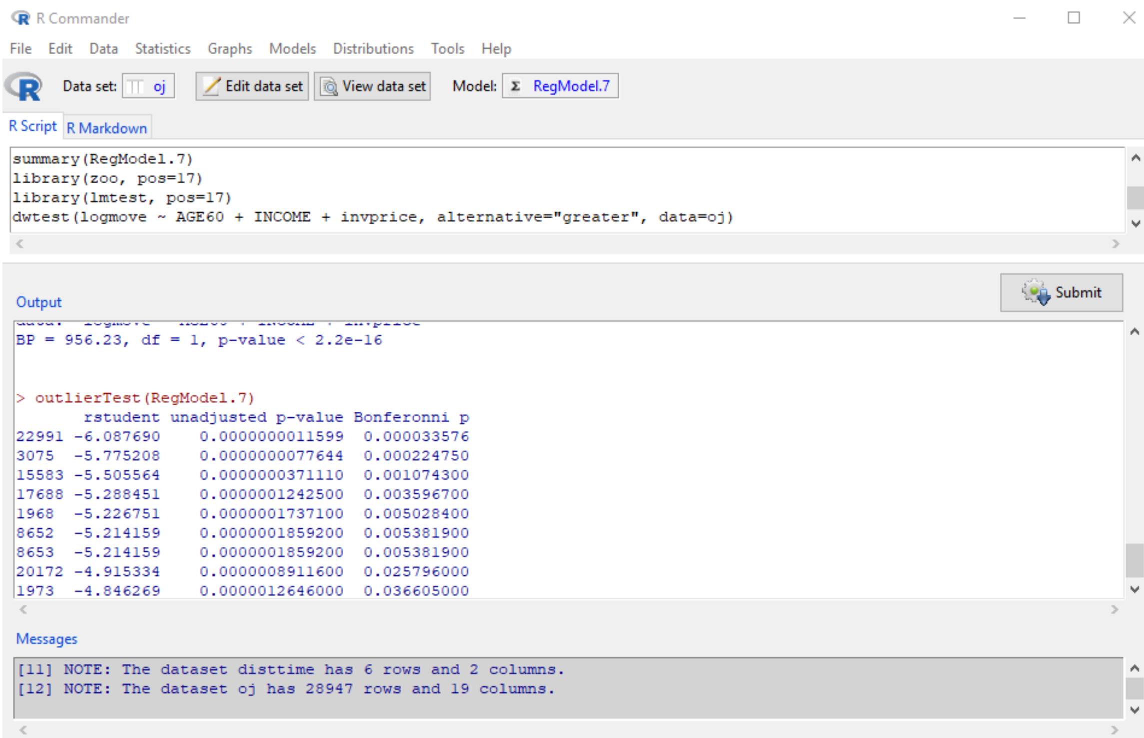
Solution to Outliers

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

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



2. Outliers have a Bonferonni p value < 0.05
3. In this example, there are several outliers. It is usually best to remove these data points from your data and retest the model. Always document that you removed outliers.