ILLINOIS
Gies College of Business

"Dating with Data!" Understanding Data-driven Managerial Decision Making: Explanatory Modeling Concepts

Sridhar Seshadri

Overview

Explanatory models

Developing and estimating the model

Interpreting the output

Improvement possibilities

Explanatory models

What determines the price of a house?

What impacts cellular phone call performance?

What determines sales at a store in a mall?

What determines the success of a new product?

What helps explain whether a customer will repay a loan?

What parameters explain how reliable is this supplier?

Which factors explain the success of stores/branches which are not all doing equally well?

Developing and estimating a model

Examine the data

Write down the model

Estimate the model

Boston Housing

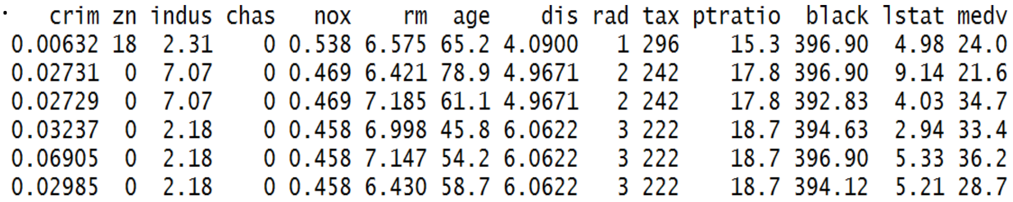
DESCRIPTION OF VARIABLES IN BOSTON HOUSING DATASET

crim	per capita crime rate by town.
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town.
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
nox	nitrogen oxides concentration (parts per 10 million).
rm	average number of rooms per dwelling.
age	proportion of owner-occupied units built prior to 1940.
dis	weighted mean of distances to five Boston employment centres.
rad	index of accessibility to radial highways.
tax	full-value property-tax rate per \$10,000.
ptratio	pupil-teacher ratio by town.
black	1000(Bk − 0.63)2 where Bk is the proportion of blacks by town.
lstat	lower status of the population (percent).
medv	median value of owner-occupied homes in \$1000s.

Open the mlbench Boston Housing from library

To get the table below click view.

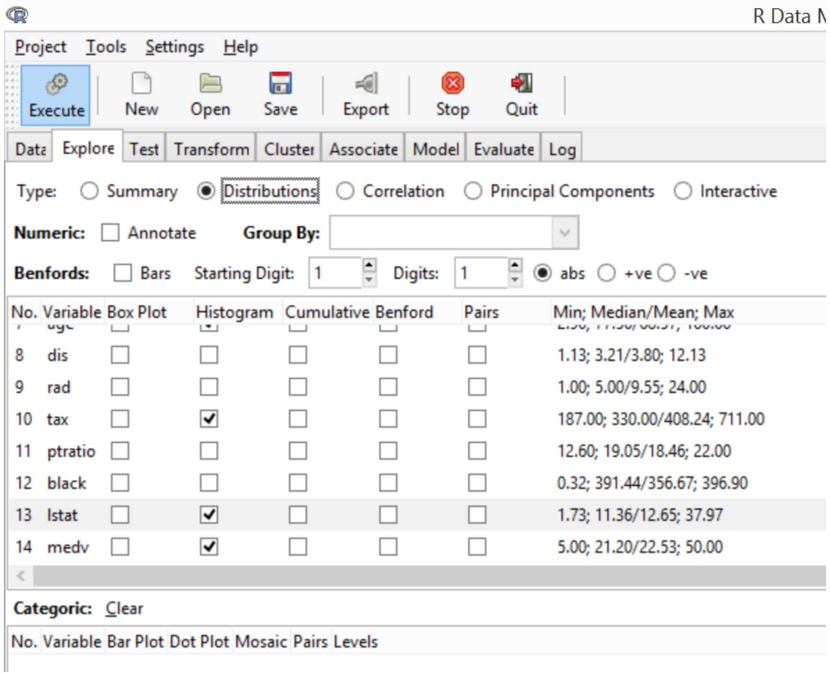
Don’t try to edit the data, it might hang up



```
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
  0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
  0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
  0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
  0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
  0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
  0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

Attribution: Ripley et al. / GPL-2 | GPL-3 / https://bit.ly/1E6z7w6

Visualization – Univariate



R Data M
Project Tools Settings Help
Execute
New
Open
Save
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ○ Summary ● Distributions ○ Correlation ○ Principal Components ○ Interactive

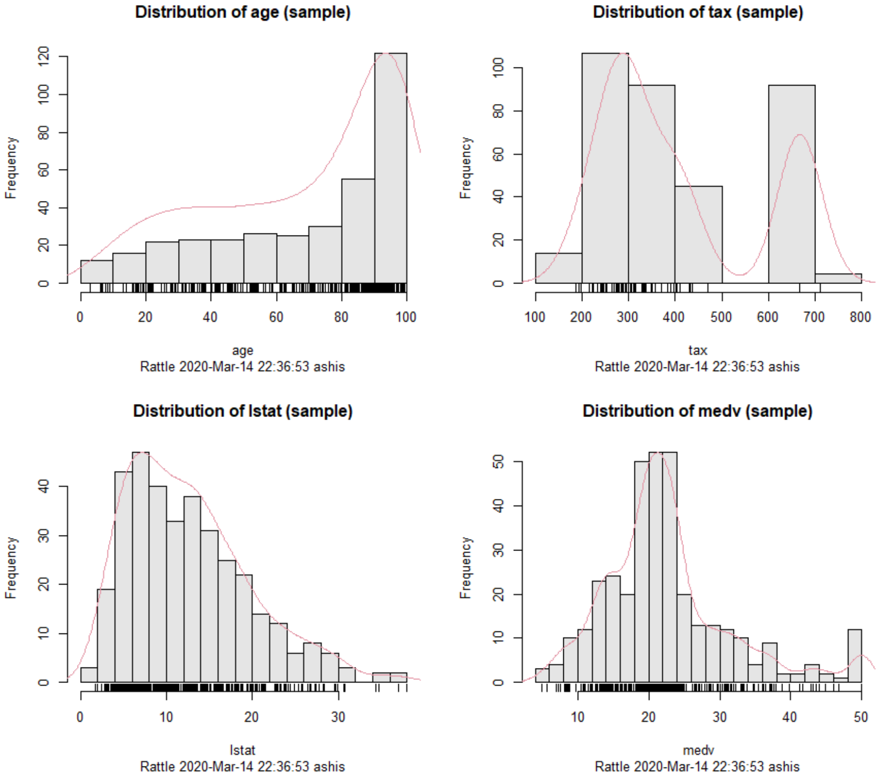
Numeric: □ Annotate Group By:

Benfords: □ Bars Starting Digit: 1 Digits: 1 ● abs ○ +ve ○ -ve

No.	Variable	Box Plot	Histogram	Cumulative	Benford	Pairs	Min; Median/Mean; Max
7	age	□	☑	□	□	□	2.90; 77.50/68.57; 100.00
8	dis	□	□	□	□	□	1.13; 3.21/3.80; 12.13
9	rad	□	□	□	□	□	1.00; 5.00/9.55; 24.00
10	tax	□	☑	□	□	□	187.00; 330.00/408.24; 711.00
11	ptratio	□	□	□	□	□	12.60; 19.05/18.46; 22.00
12	black	□	□	□	□	□	0.32; 391.44/356.67; 396.90
13	lstat	□	☑	□	□	□	1.73; 11.36/12.65; 37.97
14	medv	□	☑	□	□	□	5.00; 21.20/22.53; 50.00

Categoric: Clear

No.	Variable	Bar Plot	Dot Plot	Mosaic	Pairs	Levels



Distribution of age (sample)

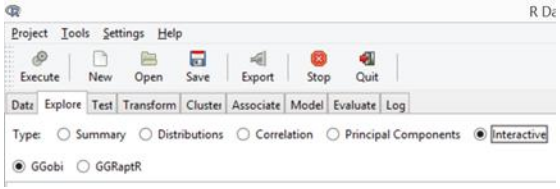
Distribution of tax (sample)

Distribution of lstat (sample)

Distribution of medv (sample)

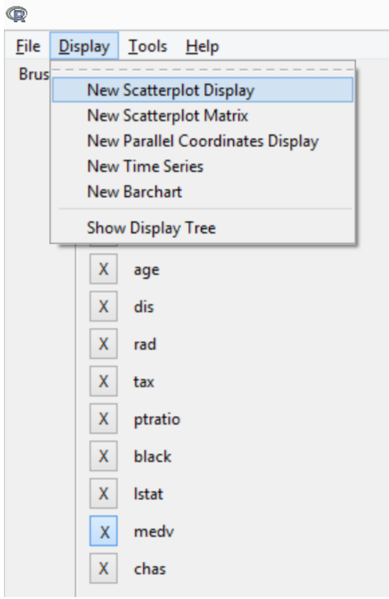
Source: Rattle GUI / Togaware

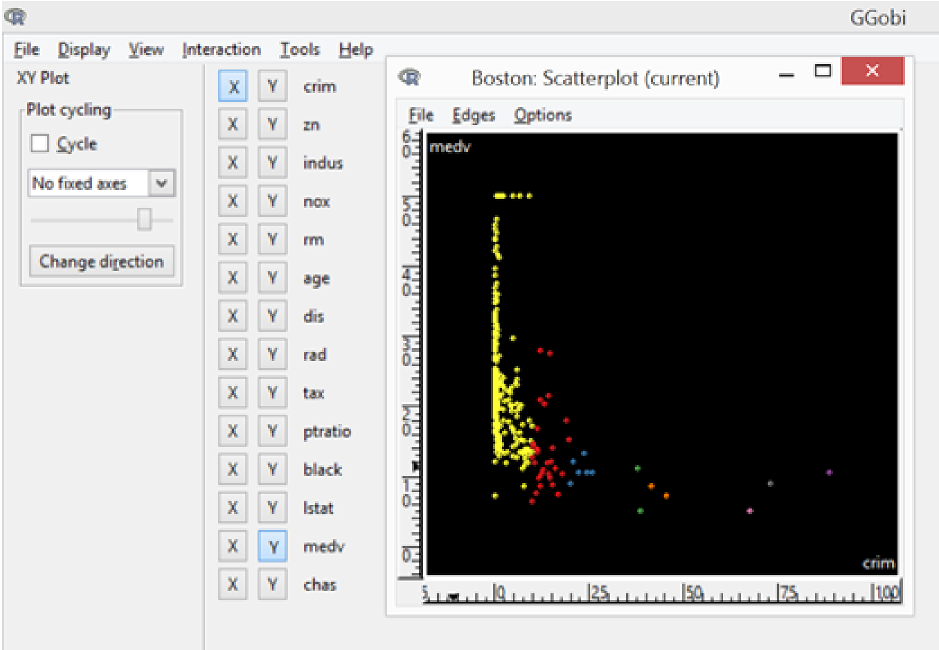
Bivariate - Scatterplot



Type: ○ Summary ○ Distributions ○ Correlation ○ Principal Components ● Interactive

● GGobi ○ GGRaptR



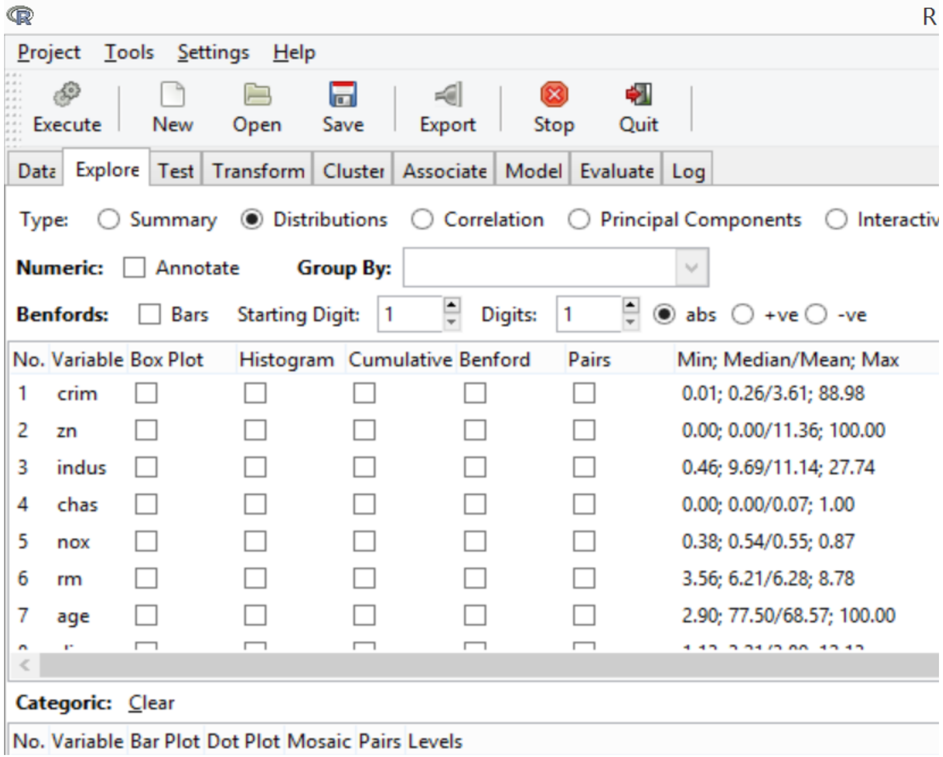


Source: Rattle GUI / Togaware

Windows use RGGOBI to create
Mac use Ggraptr

Bivariate – Correlation

Please do not attempt on Macs


R
Project Tools Settings Help
Execute
New
Open
Save
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ○ Summary ● Distributions ○ Correlation ○ Principal Components ○ Interactiv

Numeric: □ Annotate **Group By:**

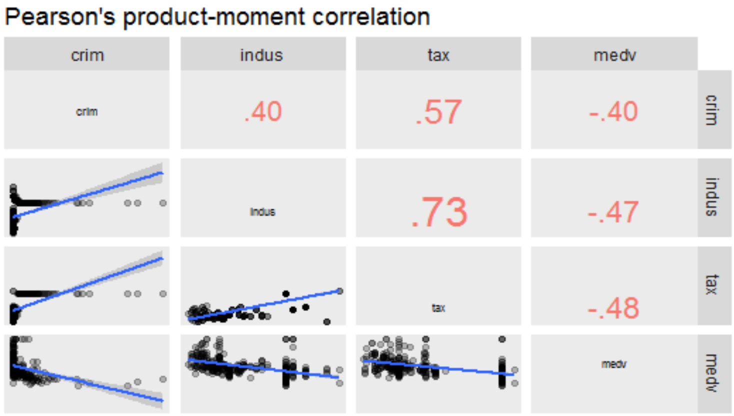
Benfords: □ Bars Starting Digit: 1 Digits: 1 ● abs ○ +ve ○ -ve

No.	Variable	Box Plot	Histogram	Cumulative	Benford	Pairs	Min; Median/Mean; Max
1	crim	□	□	□	□	□	0.01; 0.26/3.61; 88.98
2	zn	□	□	□	□	□	0.00; 0.00/11.36; 100.00
3	indus	□	□	□	□	□	0.46; 9.69/11.14; 27.74
4	chas	□	□	□	□	□	0.00; 0.00/0.07; 1.00
5	nox	□	□	□	□	□	0.38; 0.54/0.55; 0.87
6	rm	□	□	□	□	□	3.56; 6.21/6.28; 8.78
7	age	□	□	□	□	□	2.90; 77.50/68.57; 100.00

Categoric: Clear

No.	Variable	Bar Plot	Dot Plot	Mosaic	Pairs	Levels

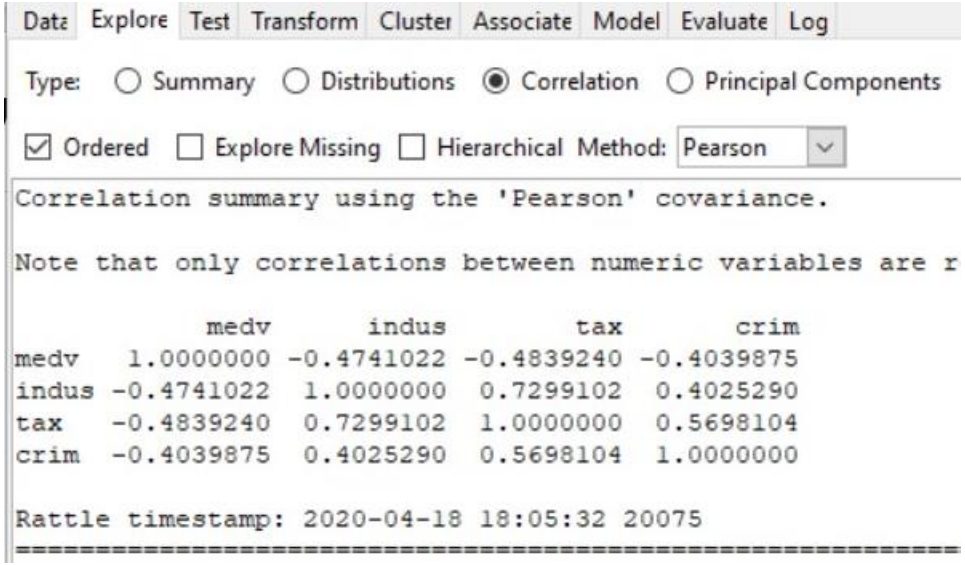
Advanced Graphics Select Explore → Distribution (don’t select any variable) → Execute



We select only these four variables as input and ignore others, for this matrix. We can select other variables to plot relationship among them.

Source: Rattle GUI / Togaware

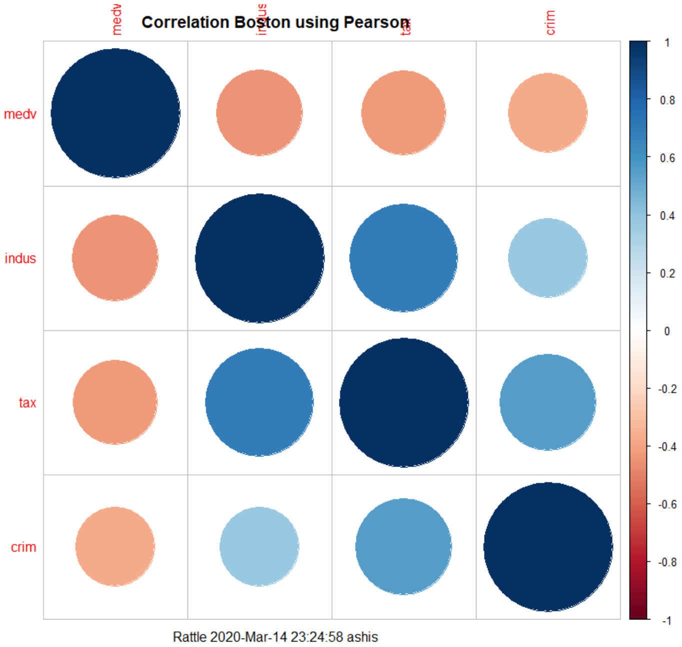
Bivariate – Correlation



Type: ○ Summary ○ Distributions ● Correlation ○ Principal Components

☑ Ordered □ Explore Missing □ Hierarchical Method: Pearson

```
Correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are r

            medv      indus        tax       crim
medv   1.0000000 -0.4741022 -0.4839240 -0.4039875
indus -0.4741022  1.0000000  0.7299102  0.4025290
tax   -0.4839240  0.7299102  1.0000000  0.5698104
crim  -0.4039875  0.4025290  0.5698104  1.0000000

Rattle timestamp: 2020-04-18 18:05:32 20075
=============================================================
```

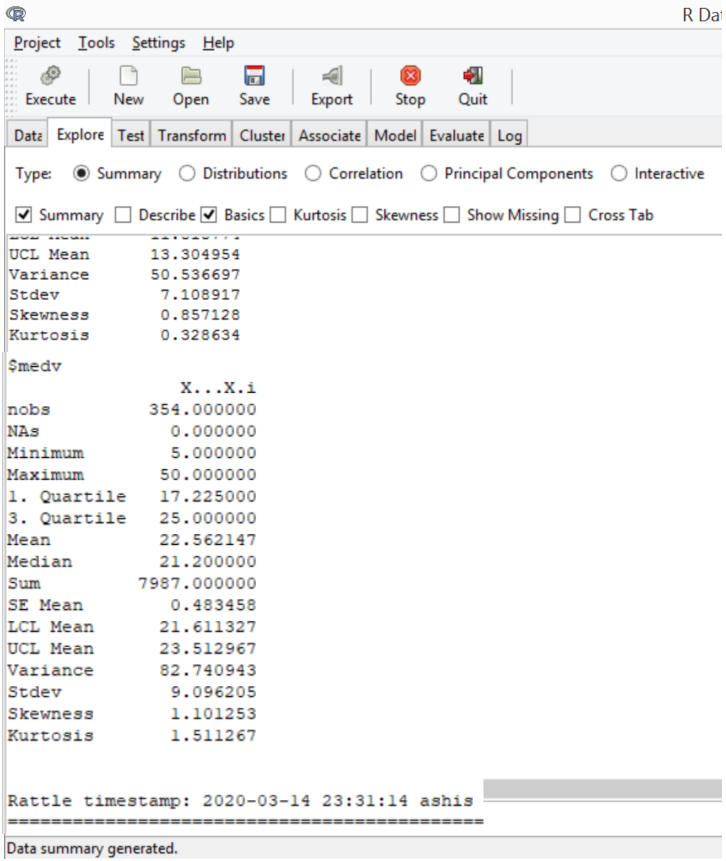


Correlation Boston using Pearson

Rattle 2020-Mar-14 23:24:58 ashis

Source: Rattle GUI / Togaware

Summary Statistics



Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

☑ Summary □ Describe ☑ Basics □ Kurtosis □ Skewness □ Show Missing □ Cross Tab

```
UCL Mean      13.304954
Variance      50.536697
Stdev          7.108917
Skewness       0.857128
Kurtosis       0.328634

$medv
                 X...X.i
nobs           354.000000
NAs              0.000000
Minimum          5.000000
Maximum         50.000000
1. Quartile     17.225000
3. Quartile     25.000000
Mean            22.562147
Median          21.200000
Sum           7987.000000
SE Mean          0.483458
LCL Mean        21.611327
UCL Mean        23.512967
Variance        82.740943
Stdev            9.096205
Skewness         1.101253
Kurtosis         1.511267


Rattle timestamp: 2020-03-14 23:31:14 ashis
==============================================
```

Data summary generated.

Source: Rattle GUI / Togaware

Model(s) We May Like to Start With

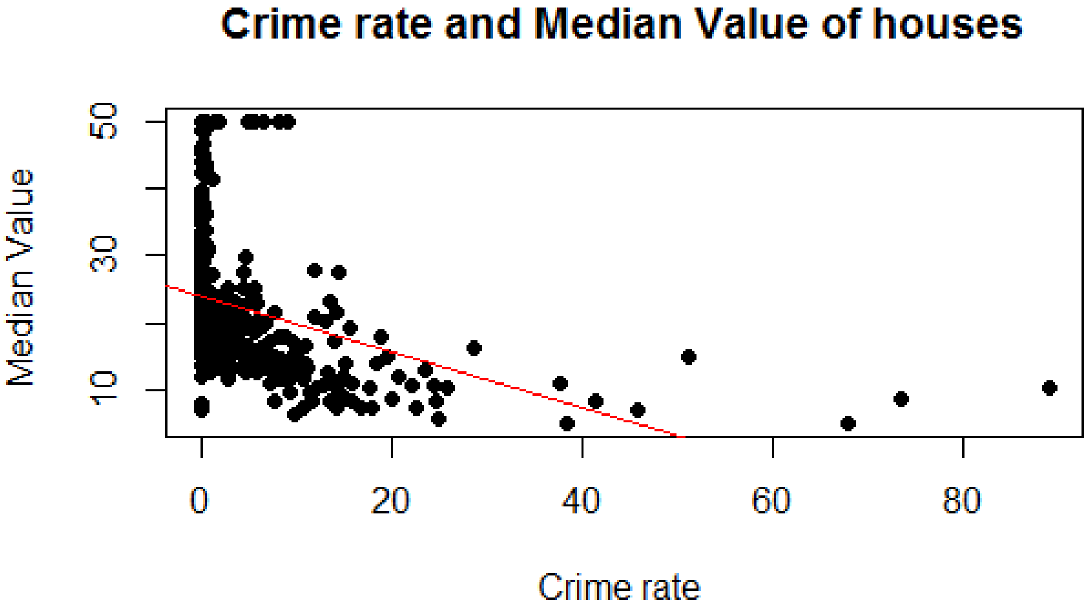
Medv = b0 + b1 * crim + error

Medv1 = b0 + b1 * crim1 + error1

$$\sum_i error^2$$

Errori = medvi – (b0 + b1 * crimi)

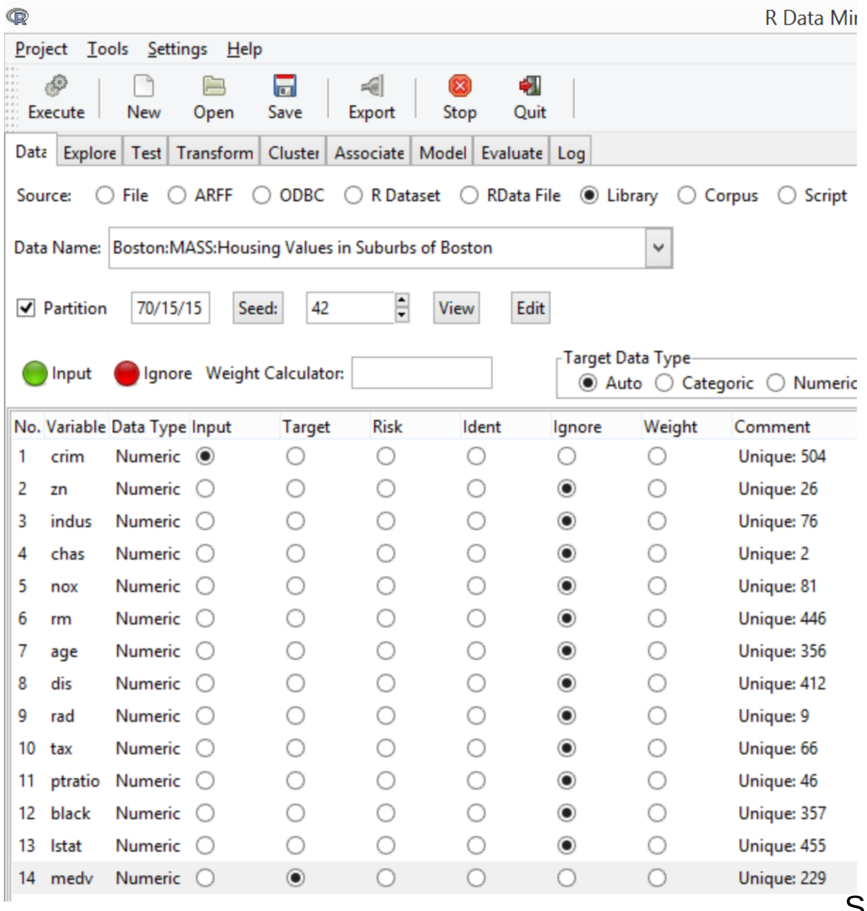
Model Visualization



Source: Rattle GUI / Togaware

You may create on RGgobi or ggraptr

Estimating the Model



Source: ○ File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ● Library ○ Corpus ○ Script

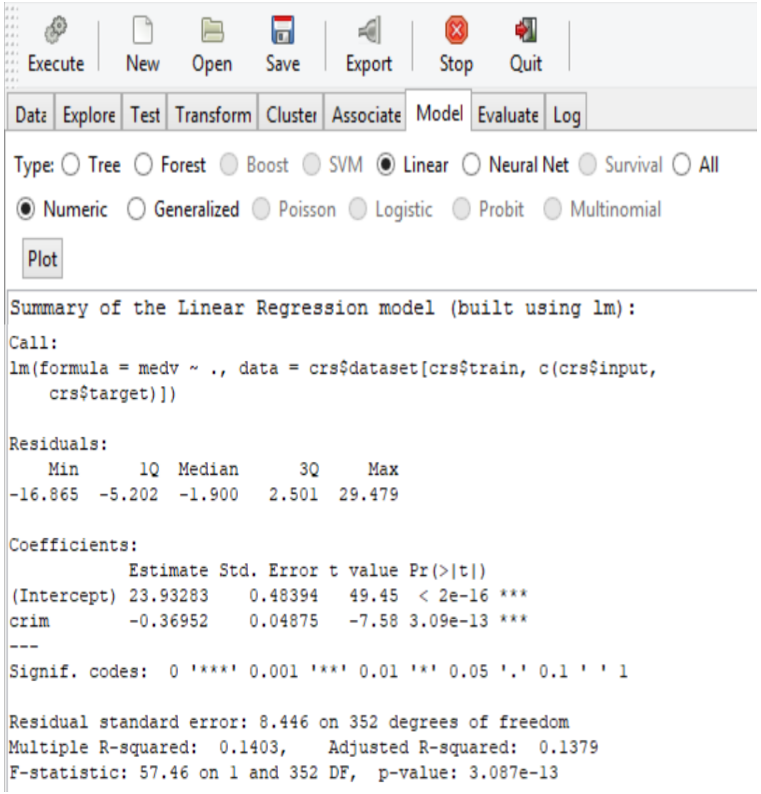
Data Name: Boston:MASS:Housing Values in Suburbs of Boston

☑ Partition 70/15/15 Seed: 42 View Edit

● Input ● Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	crim	Numeric	●	○	○	○	○	○	Unique: 504
2	zn	Numeric	○	○	○	○	●	○	Unique: 26
3	indus	Numeric	○	○	○	○	●	○	Unique: 76
4	chas	Numeric	○	○	○	○	●	○	Unique: 2
5	nox	Numeric	○	○	○	○	●	○	Unique: 81
6	rm	Numeric	○	○	○	○	●	○	Unique: 446
7	age	Numeric	○	○	○	○	●	○	Unique: 356
8	dis	Numeric	○	○	○	○	●	○	Unique: 412
9	rad	Numeric	○	○	○	○	●	○	Unique: 9
10	tax	Numeric	○	○	○	○	●	○	Unique: 66
11	ptratio	Numeric	○	○	○	○	●	○	Unique: 46
12	black	Numeric	○	○	○	○	●	○	Unique: 357
13	lstat	Numeric	○	○	○	○	●	○	Unique: 455
14	medv	Numeric	○	●	○	○	○	○	Unique: 229



Type: ○ Tree ○ Forest ○ Boost ○ SVM ● Linear ○ Neural Net ○ Survival ○ All

● Numeric ○ Generalized ○ Poisson ○ Logistic ○ Probit ○ Multinomial

```
Summary of the Linear Regression model (built using lm):
Call:
lm(formula = medv ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)])

Residuals:
    Min      1Q  Median      3Q     Max
-16.865  -5.202  -1.900   2.501  29.479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.93283    0.48394   49.45  < 2e-16 ***
crim        -0.36952    0.04875   -7.58 3.09e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.446 on 352 degrees of freedom
Multiple R-squared:  0.1403,    Adjusted R-squared:  0.1379
F-statistic: 57.46 on 1 and 352 DF,  p-value: 3.087e-13
```

Source: Rattle GUI / Togaware

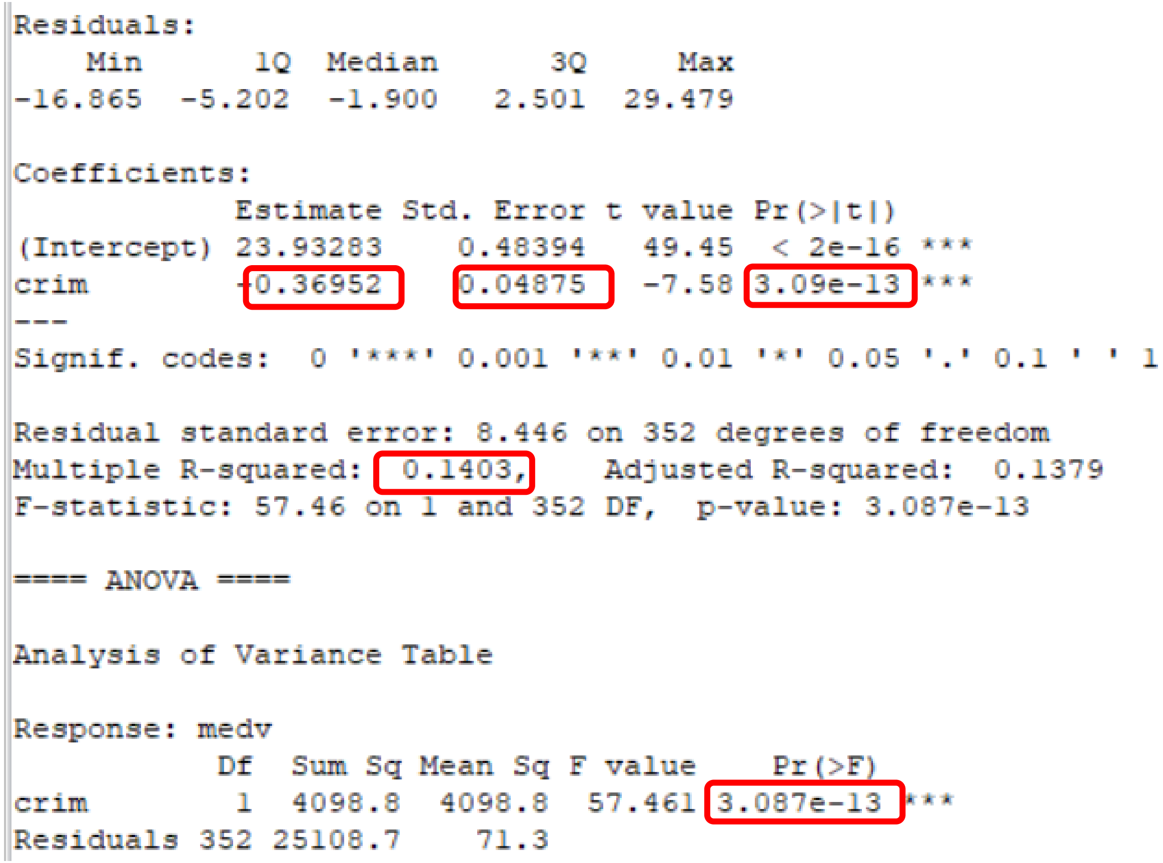
Interpret the Output

R square value and coefficients of the line

Visual examination of fit

Residuals

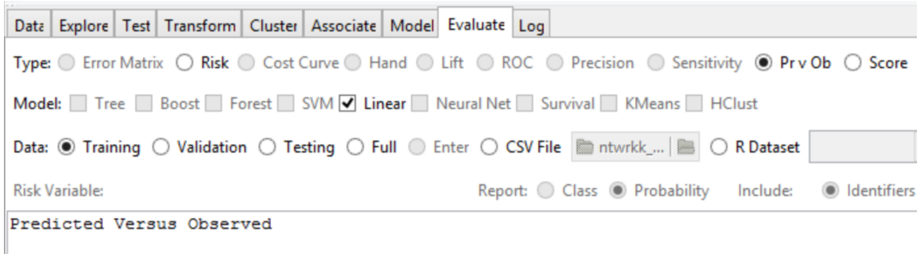
R Square and Coefficients Estimate



```
Residuals:
    Min      1Q  Median      3Q     Max
-16.865  -5.202  -1.900   2.501  29.479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.93283    0.48394   49.45  < 2e-16 ***
crim        -0.36952    0.04875   -7.58 3.09e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.446 on 352 degrees of freedom
Multiple R-squared:  0.1403,    Adjusted R-squared:  0.1379
F-statistic: 57.46 on 1 and 352 DF,  p-value: 3.087e-13

==== ANOVA ====

Analysis of Variance Table

Response: medv
           Df  Sum Sq Mean Sq F value    Pr(>F)
crim        1  4098.8  4098.8  57.461 3.087e-13 ***
Residuals 352 25108.7    71.3
```

Source: Rattle GUI / Togaware

Prediction vs Observed



Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

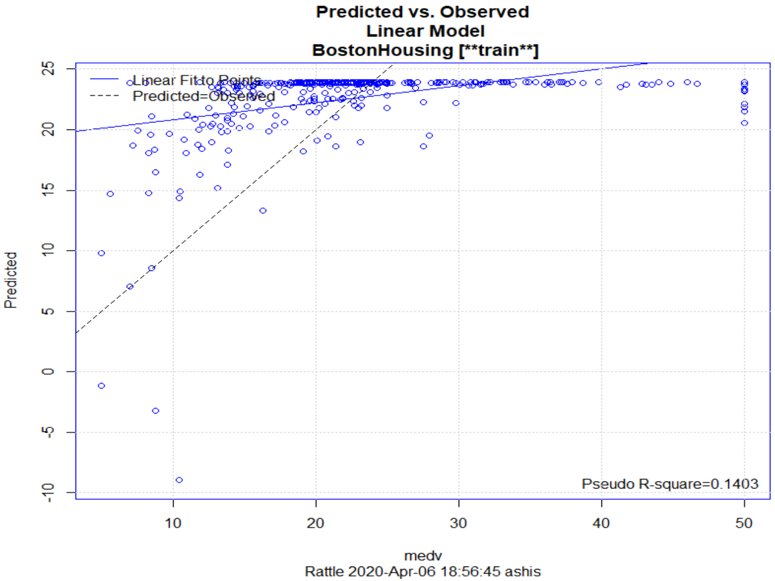
Type: ○ Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ● Pr v Ob ○ Score

Model: □ Tree □ Boost □ Forest □ SVM ☑ Linear □ Neural Net □ Survival □ KMeans □ HClust

Data: ● Training ○ Validation ○ Testing ○ Full ○ Enter ○ CSV File ntwrkk_... ○ R Dataset

Risk Variable: Report: ○ Class ● Probability Include: ● Identifiers

```
Predicted Versus Observed
```



“Predicted = Observed” line is around which we should (ideally) find the points.

“Linear fit to points” line shows how well are points scattered. We see that at lower values of medv we are over-predicting, while at higher values of medv we are under-predicting. This indicates that model could be improved.

Source: Rattle GUI / Togaware

Summary of Single Regressions

Model	Constant	Slope	R Squared	Correlation
Crime	23.93283	-0.36952	0.1403	-0.374
Indus	29.28308	-0.58728	0.1942	-0.440
tax	31.89735	-0.022851	0.1802	-0.424

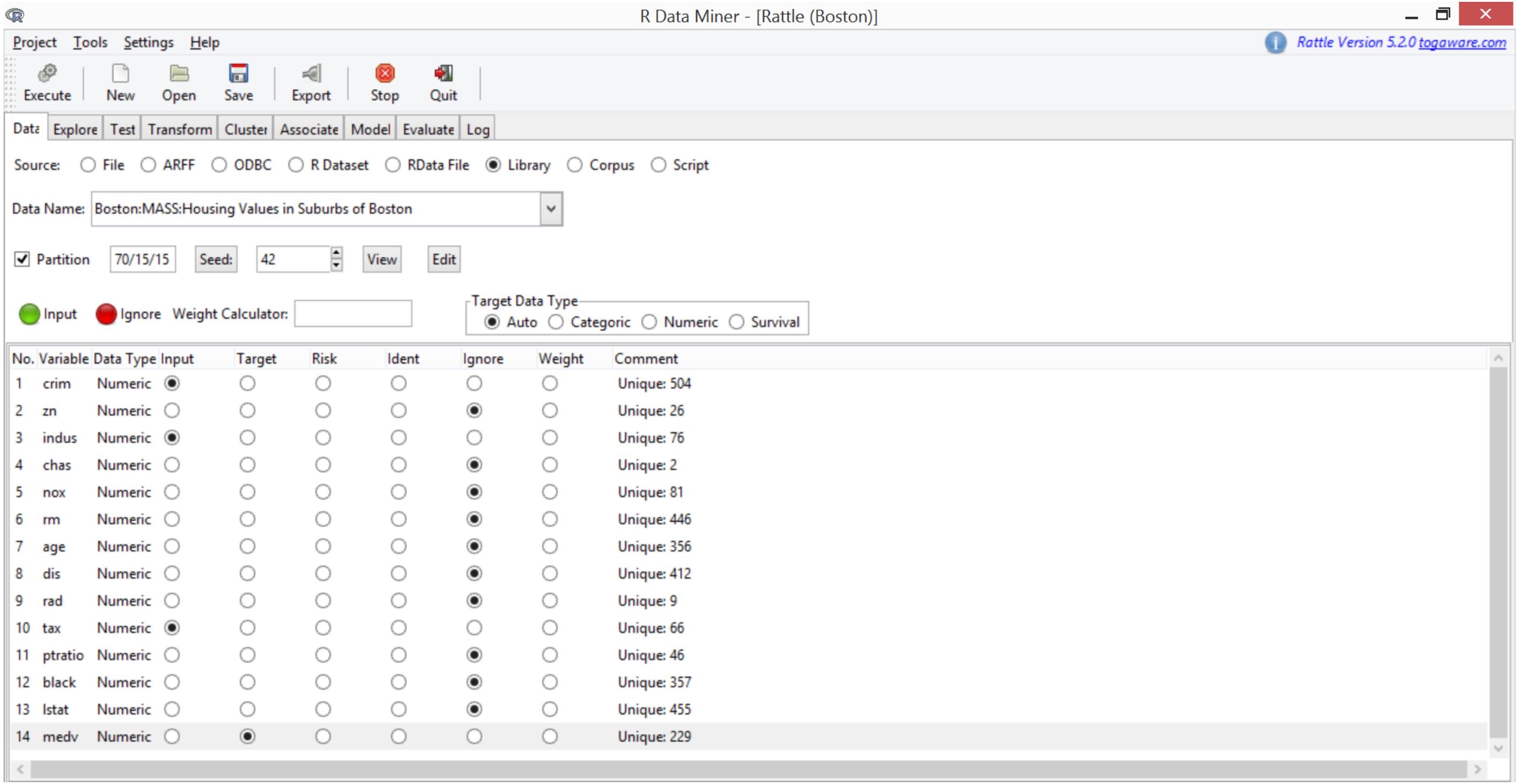
Improving the Model

Adding or removing variables

Transforming variables

Changing the nature of the fit

Adding More Variables to a Model


R Data Miner - [Rattle (Boston)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute
New
Open
Save
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log

Source: ○ File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ● Library ○ Corpus ○ Script

Data Name: Boston:MASS:Housing Values in Suburbs of Boston

☑ Partition 70/15/15 Seed: 42 View Edit

🟢 Input 🔴 Ignore Weight Calculator:

Target Data Type
● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	crim	Numeric	●	○	○	○	○	○	Unique: 504
2	zn	Numeric	○	○	○	○	●	○	Unique: 26
3	indus	Numeric	●	○	○	○	○	○	Unique: 76
4	chas	Numeric	○	○	○	○	●	○	Unique: 2
5	nox	Numeric	○	○	○	○	●	○	Unique: 81
6	rm	Numeric	○	○	○	○	●	○	Unique: 446
7	age	Numeric	○	○	○	○	●	○	Unique: 356
8	dis	Numeric	○	○	○	○	●	○	Unique: 412
9	rad	Numeric	○	○	○	○	●	○	Unique: 9
10	tax	Numeric	●	○	○	○	○	○	Unique: 66
11	ptratio	Numeric	○	○	○	○	●	○	Unique: 46
12	black	Numeric	○	○	○	○	●	○	Unique: 357
13	lstat	Numeric	○	○	○	○	●	○	Unique: 455
14	medv	Numeric	○	●	○	○	○	○	Unique: 229

Source: Rattle GUI / Togaware

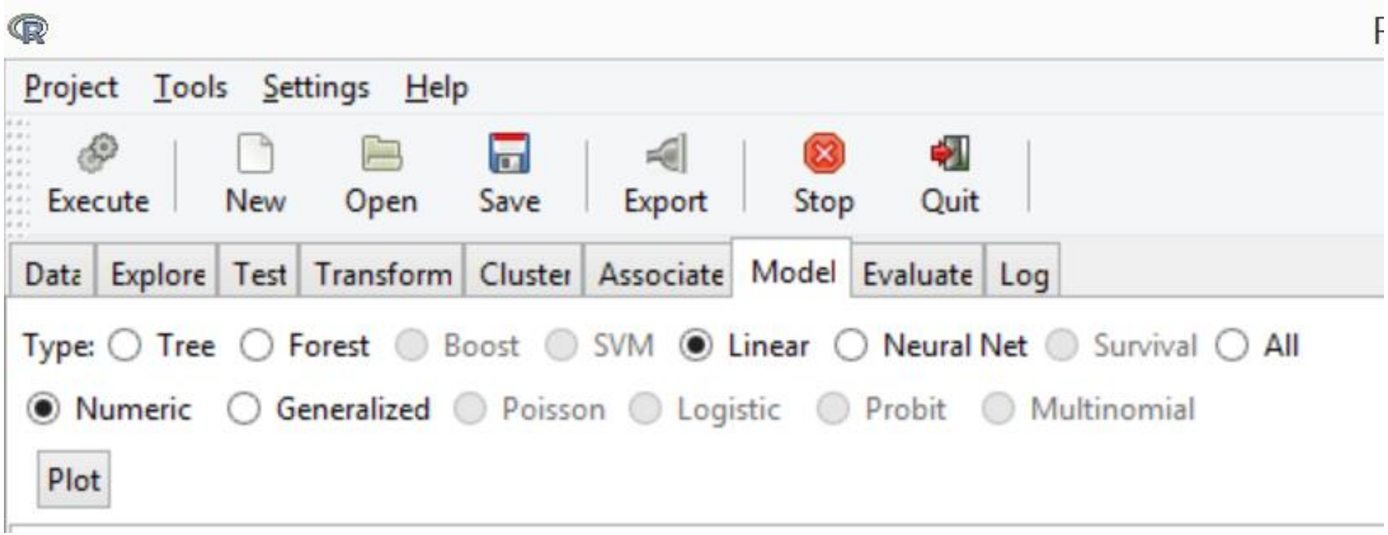
Improving the Model

Medv = b0 + b1 * crim + b2 * indus + b3 * tax + error

Medvi = b0 + b1 * crimi + b2 * indusi + b3 * taxi + errori

Minimize sum squares of errori

Estimating the Model

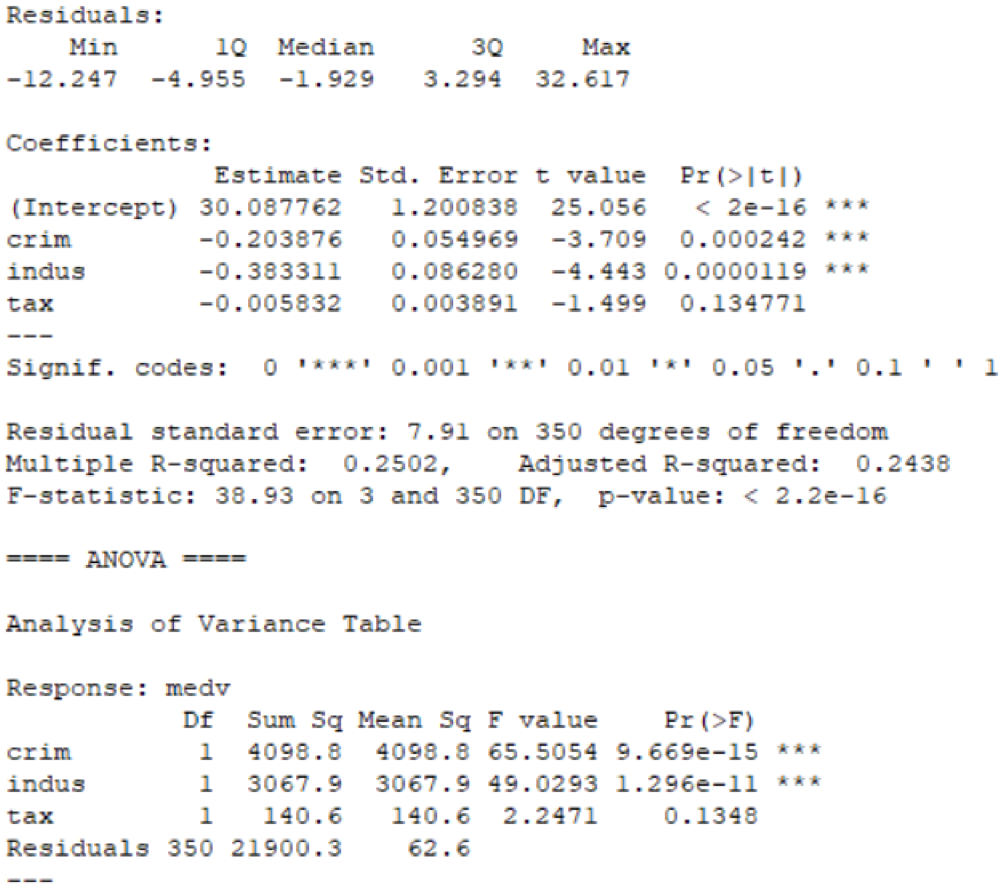

Project Tools Settings Help
Execute
New
Open
Save
Export
Stop
Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Plot

Type: ○ Tree ○ Forest ○ Boost ○ SVM ● Linear ○ Neural Net ○ Survival ○ All

● Numeric ○ Generalized ○ Poisson ○ Logistic ○ Probit ○ Multinomial

Source: Rattle GUI / Togaware

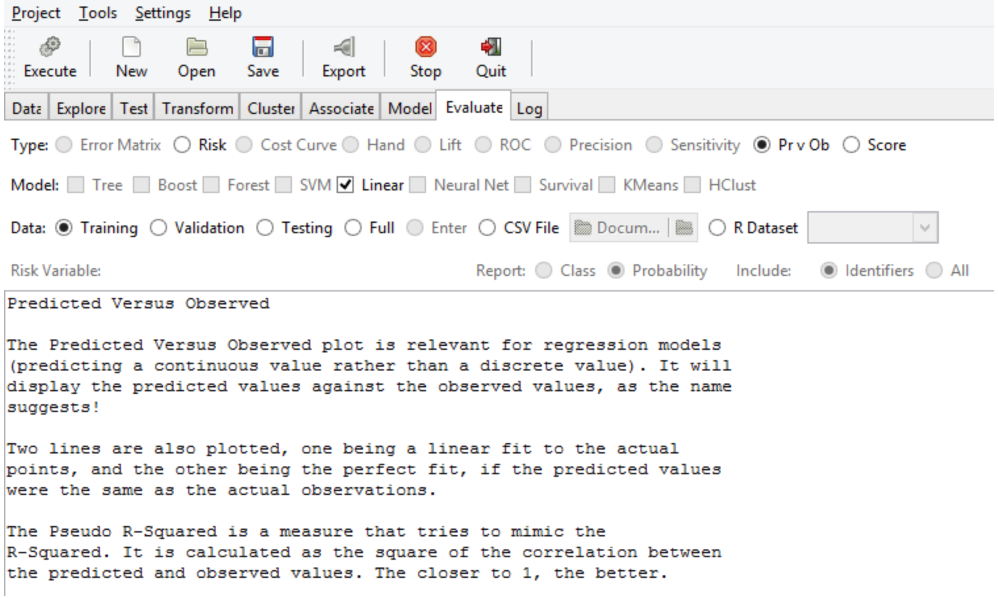
Interpreting the Output



```
Residuals:
    Min      1Q  Median      3Q     Max 
-12.247  -4.955  -1.929   3.294  32.617 

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 30.087762   1.200838  25.056   < 2e-16 ***
crim        -0.203876   0.054969  -3.709  0.000242 ***
indus       -0.383311   0.086280  -4.443 0.0000119 ***
tax         -0.005832   0.003891  -1.499  0.134771    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.91 on 350 degrees of freedom
Multiple R-squared:  0.2502,    Adjusted R-squared:  0.2438 
F-statistic: 38.93 on 3 and 350 DF,  p-value: < 2.2e-16

==== ANOVA ====

Analysis of Variance Table

Response: medv
           Df  Sum Sq Mean Sq F value    Pr(>F)    
crim        1  4098.8  4098.8 65.5054 9.669e-15 ***
indus       1  3067.9  3067.9 49.0293 1.296e-11 ***
tax         1   140.6   140.6  2.2471    0.1348    
Residuals 350 21900.3    62.6                      
---
```

Source: Rattle GUI / Togaware

Visual Inspection: Prediction vs Observed



Type: ○ Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ○ ROC ○ Precision ○ Sensitivity ● Pr v Ob ○ Score

Model: □ Tree □ Boost □ Forest □ SVM ☑ Linear □ Neural Net □ Survival □ KMeans □ HClust

Data: ● Training ○ Validation ○ Testing ○ Full ○ Enter ○ CSV File ○ R Dataset

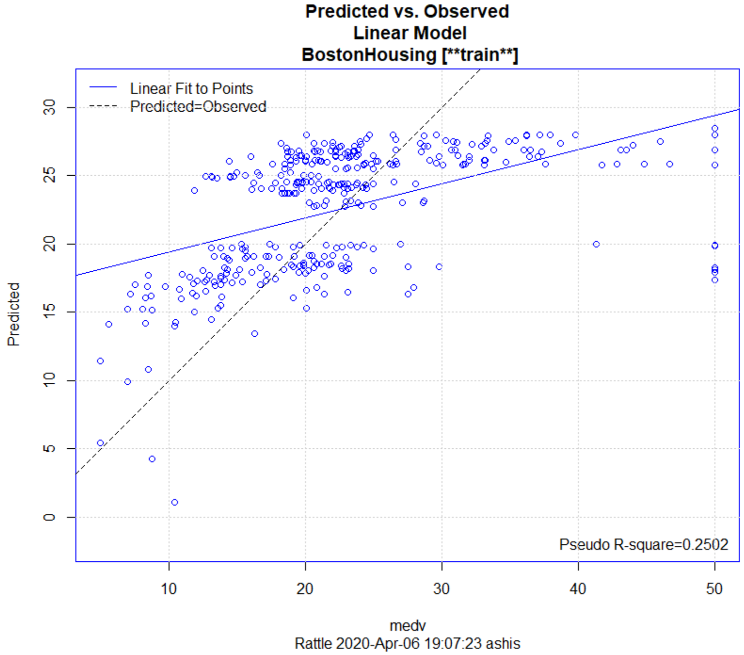
Risk Variable: Report: ○ Class ● Probability Include: ● Identifiers ○ All

Predicted Versus Observed

The Predicted Versus Observed plot is relevant for regression models (predicting a continuous value rather than a discrete value). It will display the predicted values against the observed values, as the name suggests!

Two lines are also plotted, one being a linear fit to the actual points, and the other being the perfect fit, if the predicted values were the same as the actual observations.

The Pseudo R-Squared is a measure that tries to mimic the R-Squared. It is calculated as the square of the correlation between the predicted and observed values. The closer to 1, the better.



Source: Rattle GUI / Togaware

Model Improvement (More)

More data?

More or less features?

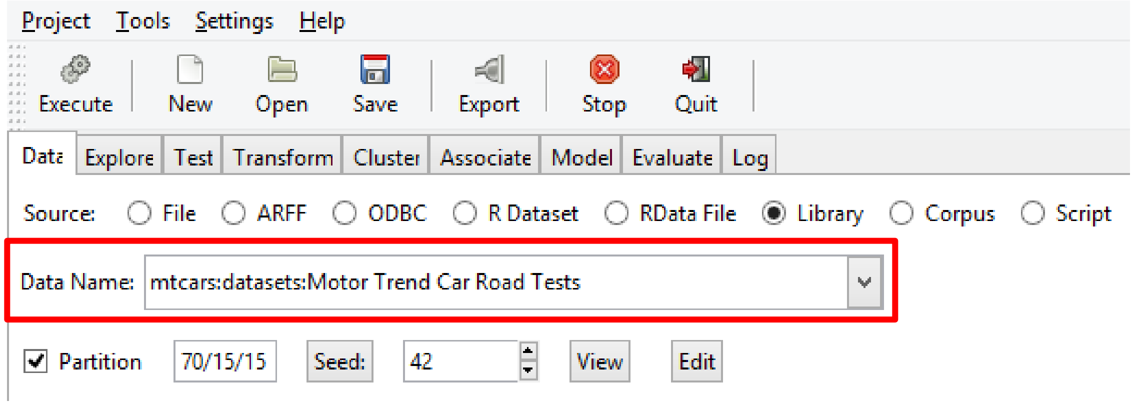
Others:

- Outliers
- Missing data
- …

What Impacts a Car’s Mileage(mpg)?

Data and features

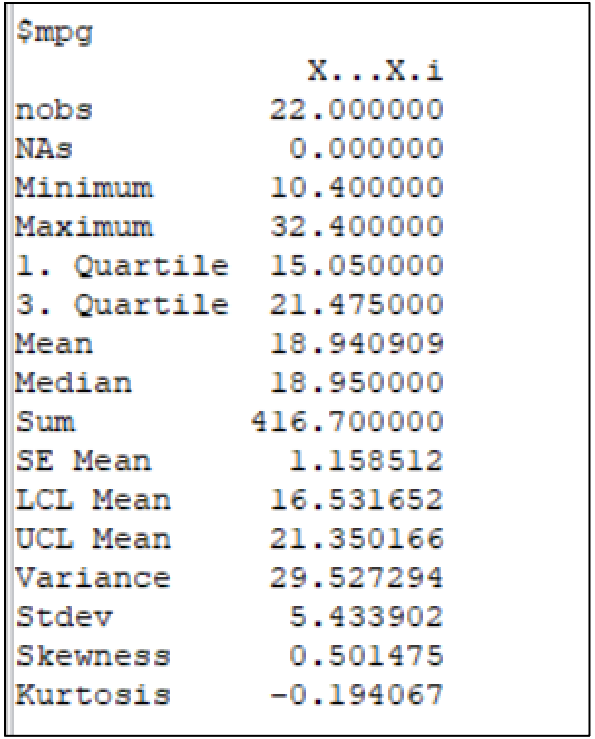
The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973 - 74 models)



Source: ○ File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ● Library ○ Corpus ○ Script

Data Name: mtcars:datasets:Motor Trend Car Road Tests

☑ Partition 70/15/15 Seed: 42

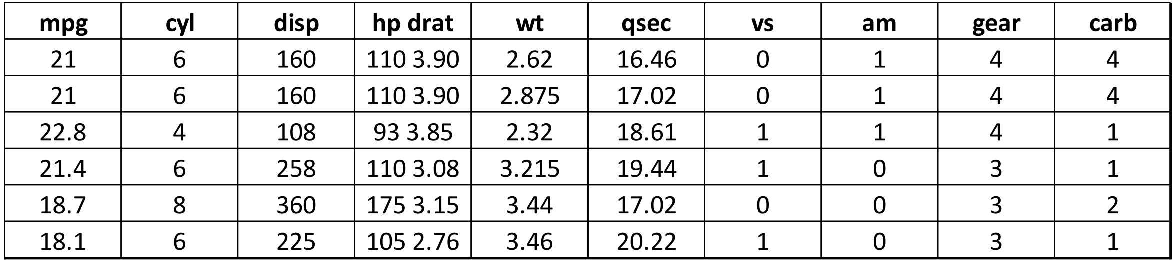


```
$mpg
                X...X.i
nobs          22.000000
NAs            0.000000
Minimum       10.400000
Maximum       32.400000
1. Quartile   15.050000
3. Quartile   21.475000
Mean          18.940909
Median        18.950000
Sum          416.700000
SE Mean        1.158512
LCL Mean      16.531652
UCL Mean      21.350166
Variance      29.527294
Stdev          5.433902
Skewness       0.501475
Kurtosis      -0.194067
```

Source: Rattle GUI / Togaware

What Impacts a Car’s Mileage(mpg)?

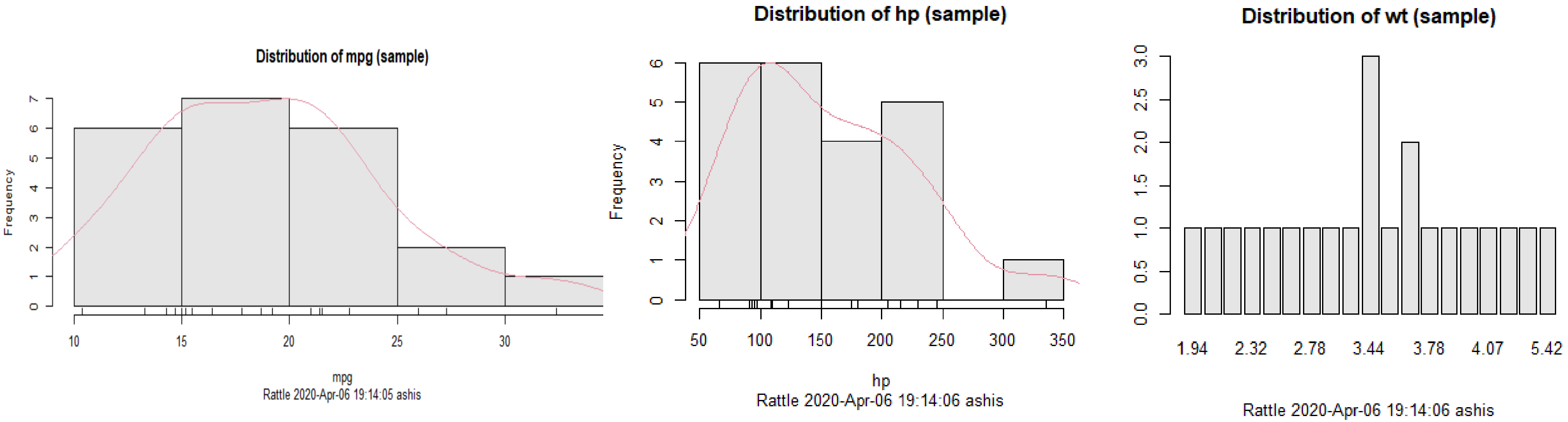
Variable	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	Engine (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors



mpg	cyl	disp	hp drat	wt	qsec	vs	am	gear	carb
21	6	160	110 3.90	2.62	16.46	0	1	4	4
21	6	160	110 3.90	2.875	17.02	0	1	4	4
22.8	4	108	93 3.85	2.32	18.61	1	1	4	1
21.4	6	258	110 3.08	3.215	19.44	1	0	3	1
18.7	8	360	175 3.15	3.44	17.02	0	0	3	2
18.1	6	225	105 2.76	3.46	20.22	1	0	3	1

Source: Rattle GUI / Togaware

Data Visualization



Please check these distributions with your knowledge about cars

Distribution of Miles_per_Gallon, Weight and Horse_Power

Source: Rattle GUI / Togaware

Marginal effects

The **Weight – Miles_per_Gallon(mpg)** model estimates a **decrease** in mileage of 4.88 miles per gallon with 1 unit increase in weight, and about 77 % of the variation in percentage of Miles_per_Gallon is associated with variation in weight.

Marginal effects

The **Horse_Power – Miles_per_Gallon (mpg)** model estimates a decrease in mileage of 0.06 miles per gallon with 1 unit increase in horse power, and about 60% of the variation in percentage of Miles_per_Gallon is associated with variation in horse power.

Marginal effects

Verify these statements by running individual regressions (use full data, all 30 observations). *Your answers may vary due to random partitions and the partitioning chosen. We have used 80-20-0 partition and random number = 42.*

Your answers may be slightly different due to different R versions

Multiple Regression

Two Predictors. The “regression model” is now the *plane (instead of line)* that best fits the points in 3-D.

The generic mathematical representation is:

$$Y=b_0+b_1X_1 + b_2X_2$$

Interpret the output

Estimate the model and verify your answer
(**mpg** = 37.22- 0.03***hp -** 3.87* **wt,** R-squared =82.7%)

Comment on the degree of Fit and the fitted parameters

Why do you think the joint estimate produces different estimates for the effect of Horse Power and Weight when compared to individual regressions?

Perform the Visual Test of fit

Others

Summary

What are explanatory models?

Data visualization and scatter plots

Estimating a model

Interpreting the output

Improving a model

References

Rattle
GUI / Togaware (https://rattle.togaware.com/)

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2019, April 26). Package "MASS". Retrieved from https://bit.ly/1E6z7w6