R Analytics Tutorial

David Walling

August 12, 2016

R Analytics Tutorial

The goal of this tutorial is to demonstrate basic data analytics using R.

Our primary objective is to determine if there is a statistically significant difference in gas mileage for cars with automatic vs manual transmissions.

We will use the pre-built data set 'mtcars' to first explore graphically our data and then peform basic regression analysis in R.

Throughout this tutorial, we will be exploring more detailed and advanced features of the R programming environment.

Included Datasets

First, lets explore which data sets are available by default in R.

```
data()
```

mtcars

```
Data sets in package 'datasets':

AirPassengers               Monthly Airline Passenger Numbers 1949-1960
BJsales                     Sales Data with Leading Indicator
BJsales.lead (BJsales)      Sales Data with Leading Indicator
BOD                         Biochemical Oxygen Demand
CO2                         Carbon Dioxide Uptake in Grass Plants
ChickWeight                 Weight versus age of chicks on different diets
DNase                       Elisa assay of DNase
EuStockMarkets              Daily Closing Prices of Major European Stock Indices,
                            1991-1998
Formaldehyde                Determination of Formaldehyde
HairEyeColor                Hair and Eye Color of Statistics Students
Harman23.cor                Harman Example 2.3
Harman74.cor                Harman Example 7.4
Indometh                    Pharmacokinetics of Indomethacin
InsectSprays                Effectiveness of Insect Sprays
JohnsonJohnson              Quarterly Earnings per Johnson & Johnson Share
LakeHuron                   Level of Lake Huron 1875-1972
LifeCycleSavings            Intercountry Life-Cycle Savings Data
Loblolly                    Growth of Loblolly pine trees
Nile                        Flow of the River Nile
Orange                      Growth of Orange Trees
OrchardSprays               Potency of Orchard Sprays
PlantGrowth                 Results from an Experiment on Plant Growth
Puromycin                   Reaction Velocity of an Enzymatic Reaction
Seatbelts                   Road Casualties in Great Britain 1969-84
Theoph                      Pharmacokinetics of Theophylline
Titanic                     Survival of passengers on the Titanic
ToothGrowth                 The Effect of Vitamin C on Tooth Growth in Guinea Pigs
```

mtcars

We will be using the 'mtcars' data set for this tutorial. Let's load it into our environment. And view additional help information about the data set.

```
data(mtcars)
?mtcars
```

mtcars

mtcars {datasets} R Documentation

Motor Trend Car Road Tests

Description

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).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 variables.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

Source

Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, **37**, 391–411.

Examples

```
require(graphics)
pairs(mtcars, main = "mtcars data")
coplot(mpg ~ disp | as.factor(cyl), data = mtcars,
       panel = panel.smooth, rows = 1)
```

[Package *datasets* version 3.2.3 Index]

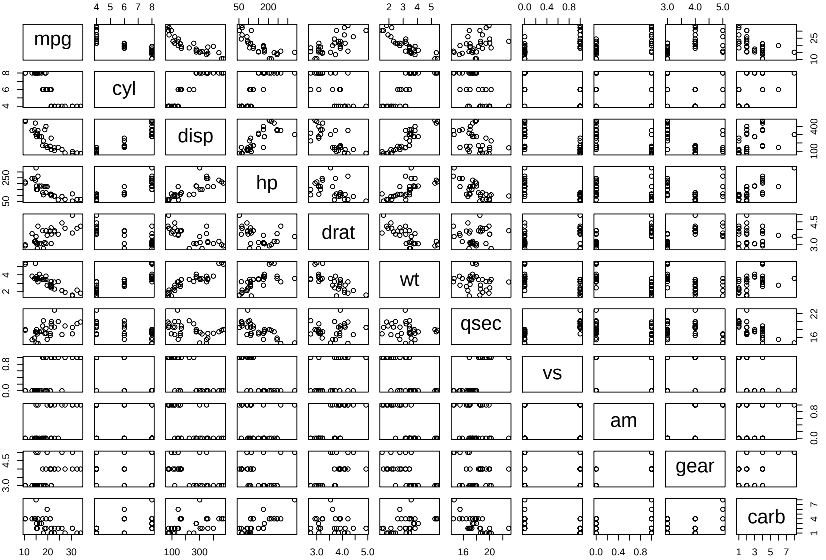
Plotting

Let's explore graphically the replationship between the variables.

The plot() function is an example of an 'overloaded' function in R. This means that its behavior differs depending on what object or parameters it is passed in. In this case, we are passing in a data.frame, and plot.data.frame will be called.

See ?plot.data.frame for details.

Plotting

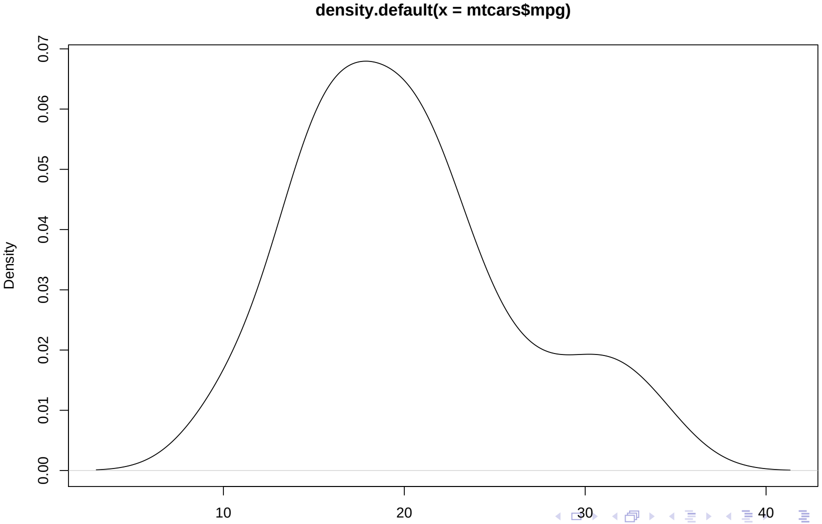


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
4 5 6 7 8
50 200
2 3 4 5
0.0 0.4 0.8
3.0 4.0 5.0
10 20 30
100 300
3.0 4.0 5.0
16 20
0.0 0.4 0.8
1 3 5 7
10 25
8 6 4
100 400
50 250
3.0 4.5
2 4
16 22
0.0 0.8
0.0 0.8
3.0 4.5
1 4 7

Distribution

How are the values of MPG distributed?

```
plot(density(mtcars$mpg))
```



Factors

Do we *see* a difference between automatic and manual transmissions?

First, we note that the variable representing the auto vs. manual is a numeric. We want to model this as categorical. R has a special 'class' of variable for representing categorical variables known as 'factor'.

Factors

Use as.factor to add a new variable to the data.frame.

```
mtcars$transmission <- as.factor(mtcars$am)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  12 variables:
##  $ mpg         : num  21 21 22.8 21.4 18.7 18.1 14.3 24.
##  $ cyl         : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp        : num  160 160 108 258 360 ...
##  $ hp          : num  110 110 93 110 175 105 245 62 95 1
##  $ drat        : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3
##  $ wt          : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec        : num  16.5 17 18.6 19.4 17 ...
##  $ vs          : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am          : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear        : num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb        : num  4 4 1 1 2 1 4 2 2 4 ...
##  $ transmission: Factor w/ 2 levels "0","1": 2 2 2 1 1 1
```

Factors

Reset values to something more readable

```
levels(mtcars$transmission) <- c('Automatic', 'Manual')
str(mtcars)
```

```
## 'data.frame':    32 obs. of  12 variables:
##  $ mpg         : num  21 21 22.8 21.4 18.7 18.1 14.3 24.
##  $ cyl         : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp        : num  160 160 108 258 360 ...
##  $ hp          : num  110 110 93 110 175 105 245 62 95 1
##  $ drat        : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3
##  $ wt          : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec        : num  16.5 17 18.6 19.4 17 ...
##  $ vs          : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am          : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear        : num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb        : num  4 4 1 1 2 1 4 2 2 4 ...
##  $ transmission: Factor w/ 2 levels "Automatic","Manual'
```

Factors

Finally, for simplicity, lets drop the original values

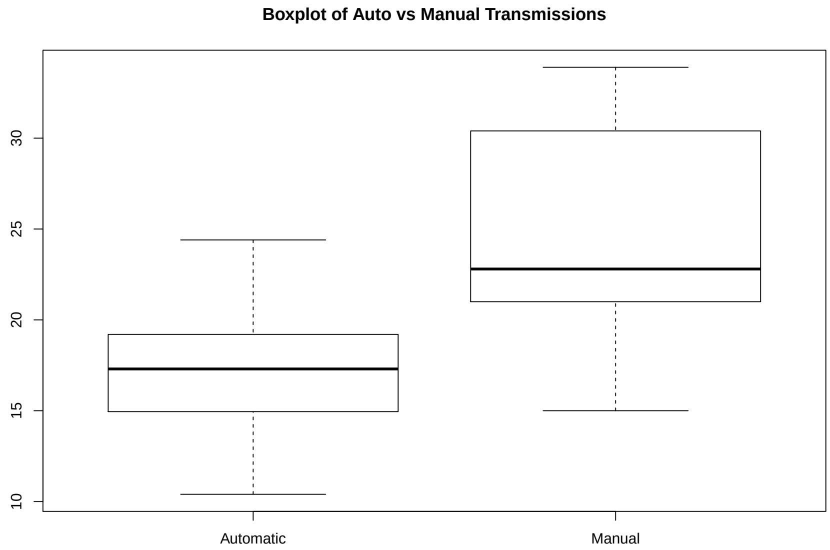
```
mtcars <- subset(mtcars, select=-c(am))
```

Boxplot

Now let's break down the distribution between Automatic vs Manual transmissions.

```
boxplot(mpg~transmission,
        data=mtcars,
        main='Boxplot of Auto vs Manual Transmissions')
```

Boxplot



Boxplot of Auto vs Manual Transmissions
10
15
20
25
30
Automatic
Manual

Linear Regression

We want to examine effects of other variables on the outcome of intest, MPG.

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + .. + \beta_n * X_{ni} + \epsilon_i$$

Y = mpg
$\beta_0 = intercept$
$\beta_1 - \beta_n =$ effect of each predictor

Linear Regression: Assumptions

Linear regression has the following assumptions:

- Linear relationship, i.e. a linear combination of predictor variable
- Residuals are normally distributed
- Residuals are independent
- Residuals variance constant

Simple Model

First, we create a linear regression model using just the transmission type.

```
model_simple <- lm(mpg~transmission, data=mtcars)
```

Simple Model - Verification

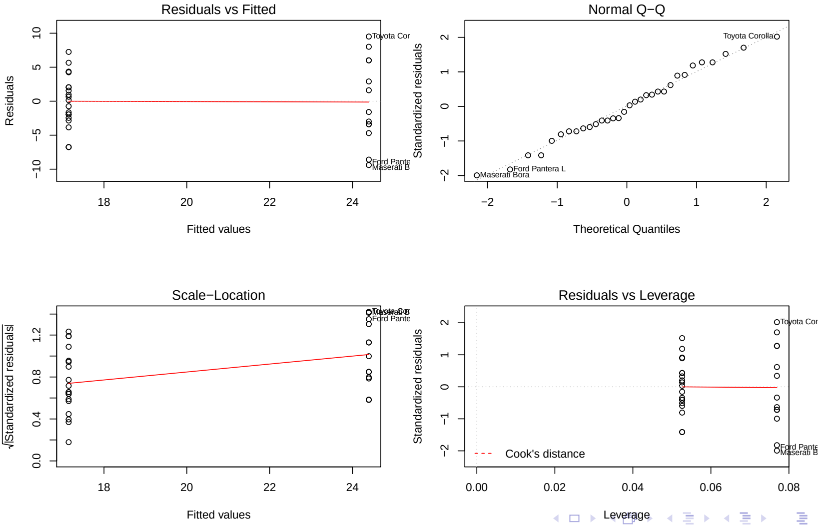
Before we interpret the results, lets verify our assumptions.

We can use the general graphics par() function to set a variety of graphical parameters. In this case, we want the 4 plots produced by the plot.lm function (remember function overloading!) to print to the same output.

```
par(mfrow=c(2,2))
plot(model_simple)
```

Simple Model - Verification

```
par(mfrow=c(2,2))
plot(model_simple)
```



Simple Model - Results

```
summary(model_simple)
```

```
##
## Call:
## lm(formula = mpg ~ transmission, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          17.147      1.125  15.247 1.13e-15
## transmissionManual    7.245      1.764   4.106 0.000285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 4.902 on 30 degrees of freedom
```

Simple Model - Results

$\beta_0 = 17.147$
$\beta_{transmissionManual} = 7.245$

Our model is telling us that we expect a manual transmission to get 7.25 MPG better than automatic.

However, our model only explains 34% of the variance seen in the data.

What might be a problem with this model?

Confounding

In our simple model, we are not considering the effects of the other variables, which are essentially unknown to our model.

Let's try adding them in.

Kitchen Sink

Let's throw all the available variables into the model.

```
model_kitchensink <- lm(mpg~., data=mtcars)
```

```
summary(model_kitchensink)
```

Kitchen Sink

```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        12.30337   18.71788   0.657   0.5181
cyl                -0.11144    1.04502  -0.107   0.9161
disp                0.01334    0.01786   0.747   0.4635
hp                 -0.02148    0.02177  -0.987   0.3350
drat                0.78711    1.63537   0.481   0.6353
wt                 -3.71530    1.89441  -1.961   0.0633 .
qsec                0.82104    0.73084   1.123   0.2739
vs                  0.31776    2.10451   0.151   0.8814
gear                0.65541    1.49326   0.439   0.6652
carb               -0.19942    0.82875  -0.241   0.8122
transmissionManual  2.52023    2.05665   1.225   0.2340
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

ANOVA

Is this model 'better'? We can use anova to test this.

The Null Hypothesis is that the two models are equally good.

```
anova(model_simple, model_kitchensink)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ transmission
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs -
##     transmission
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     21 147.49  9     573.4 9.0711 1.779e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

ANOVA

Conclusion: the kitchen sink model is an improvement.

However, we still see that the model is having trouble distinguishing the influence of each variable as all the beta p-values are > 0.05

Variance Inflation Factors

We'll use another 3rd party package 'car' (no relation) to check for multi-collinearity in our model.

First, you may need to install the package in your environment and then load it.

```
#install.packages('car')
library(car)
```

Variance Inflation Factors

Now run vif and use the heuristic that you want values where sqrt(vif) <= 2.

```r
vif(model_kitchensink)
```

```
##       cyl      disp        hp      drat     1
##  15.373833  21.620241   9.832037   3.374620    1
##      qsec        vs      gear      carb tran
##   7.527958   4.965873   5.357452   7.908747
```

```r
sqrt(vif(model_kitchensink)) > 2
```

```
##       cyl      disp        hp      drat
##      TRUE      TRUE      TRUE     FALSE
##      qsec        vs      gear      carb tran
##      TRUE      TRUE      TRUE      TRUE
```

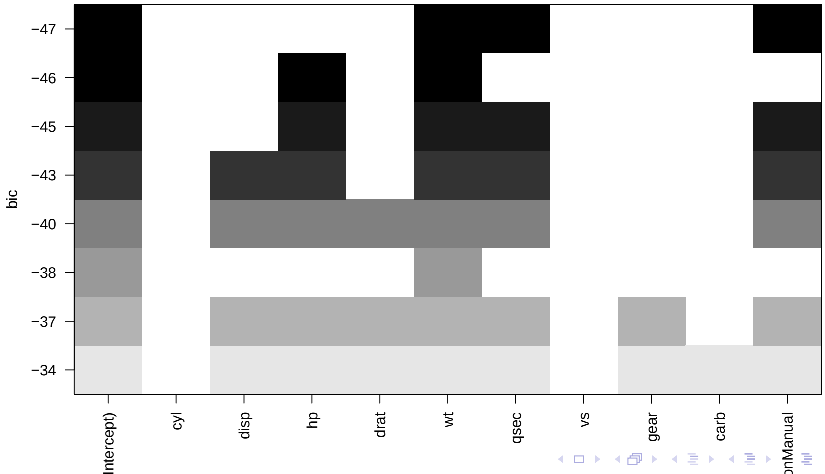
Variable Selection

So, this model is no good. Let's try and find a compromise between one that is too simple and one that is overally complex.

Again, we'll use 3rd party package 'leaps' to automatically select the appropriate variables using backward selection and the BIC selection critera. BIC penalizes the model for each additional variable.

Variable Selection

```
library(leaps)
result <- regsubsets(mpg~., data=mtcars,
                     method='backward')
plot(result, scale="bic")
```



Final Model

Let's build are final model and repeat the basic validation.

```
model_final <- lm(mpg~wt+qsec+transmission, data=mtcars)
```

Final Model - Check

Let's re-check for multi-collinearity.

```
vif(model_final)
```

```
##           wt         qsec transmission
##     2.482952     1.364339     2.541437
```

Final Model - Summary

```r
summary(model_final)
```

```

Call:
lm(formula = mpg ~ wt + qsec + transmission, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4811 -1.5555 -0.7257  1.4110  4.6610 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)          9.6178     6.9596   1.382 0.177915    
wt                  -3.9165     0.7112  -5.507 6.95e-06 ***
qsec                 1.2259     0.2887   4.247 0.000216 ***
transmissionManual   2.9358     1.4109   2.081 0.046716 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,	Adjusted R-squared:  0.8336 
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Final Model - Conclusion

Our model accounts for 83% of the variance seen in the data.

Holding qsec and wt equal, a manual transmission is expected to achieve 2.93 MPG better than an automatic.

Conclusion

Break for lunch!!!