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 perform 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, let's explore which data sets are available by default in R.

data()
## mtcars

Data sets in package ‘datasets’:

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

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

Source

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 relationship 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.
Distribution

How are the values of MPG distributed?

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

![Density plot of MPG values](image)
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`.

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

```r
## 'data.frame': 32 obs. of 12 variables:
##  $ mpg   : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
##  $ 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 95 123 ...  
##  $ drat  : num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
##  $ 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 1 1 1 1 1 1 1 1 ...  
```
Factors

Reset values to something more readable

```r
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 24.4 22.8 19.2 ... 
## $ 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 95 123 ... 
## $ drat : num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ... 
## $ 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, let's drop the original values

```r
mtcars <- subset(mtcars, select=-c(am))
```
Now let’s break down the distribution between Automatic vs Manual transmissions.

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

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

\[ Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \ldots + \beta_n * X_{ni} + \epsilon_i \]

\( Y = \text{mpg} \)
\( \beta_0 = \text{intercept} \)
\( \beta_1 - \beta_n = \text{effect of each predictor} \)
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
First, we create a linear regression model using just the transmission type.

```r
model_simple <- lm(mpg~transmission, data=mtcars)
```
Simple Model - Verification

Before we interpret the results, let's 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.

```r
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 ' ' 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.
Let’s throw all the available variables into the model.

```r
model_kitchensink <- lm(mpg~., data=mtcars)
summary(model_kitchensink)
```
Call:
`lm(formula = mpg ~ ., data = mtcars)`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-3.4506</td>
<td>-1.6044</td>
<td>-0.1196</td>
<td>1.2193</td>
<td>4.6271</td>
</tr>
</tbody>
</table>

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

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

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

```text
## Analysis of Variance Table

## Model 1: mpg ~ transmission
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + gear + carb + transmission

<table>
<thead>
<tr>
<th></th>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>720.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>147.49</td>
<td>9</td>
<td>573.4</td>
<td>9.0711</td>
<td>1.779e-05 ***</td>
</tr>
</tbody>
</table>

---

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

```r
#install.packages('car')
library(car)
```
Variance Inflation Factors

Now run vif and use the heuristic that you want values where \( \sqrt{\text{vif}} \leq 2 \).

\[
\begin{array}{cccccc}
\text{vif(model_kitchensink)} \\
\hline
## & cyl & disp & hp & drat & wt \\
## & qsec & vs & gear & carb & transmission \\
## & 7.527958 & 4.965873 & 5.357452 & 7.908747 & 4.648487 \\
\end{array}
\]

\[
\begin{array}{cccccc}
\text{sqrt(vif(model_kitchensink)) > 2} \\
\hline
## & cyl & disp & hp & drat \\
## & TRUE & TRUE & TRUE & FALSE \\
## & qsec & vs & gear & carb & transmission \\
## & TRUE & TRUE & TRUE & TRUE \\
\end{array}
\]
So, this model is no good. Let’s try and find a compromise between one that is too simple and one that is overly complex.

Again, we’ll use 3rd party package ‘leaps’ to automatically select the appropriate variables using backward selection and the BIC selection criteria. BIC penalizes the model for each additional variable.
library(leaps)
result <- regsubsets(mpg~., data=mtcars, method="backward")
plot(result, scale="bic")
Let's build our final model and repeat the basic validation.

```r
model_final <- lm(mpg~wt+qsec+transmission, data=mtcars)
```
Let's re-check for multi-collinearity.

```
vif(model_final)
```

<table>
<thead>
<tr>
<th></th>
<th>wt</th>
<th>qsec</th>
<th>transmission</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.482952</td>
<td>1.364339</td>
<td>2.541437</td>
</tr>
</tbody>
</table>
**Final Model - Summary**

```r
summary(model_final)
```

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

**Residuals:**

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-3.4811</td>
<td>-1.5555</td>
<td>-0.7257</td>
<td>1.4110</td>
<td>4.6610</td>
</tr>
</tbody>
</table>

**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!!!