R-Linear Regression

David Walling

R-Linear Regression 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

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
```

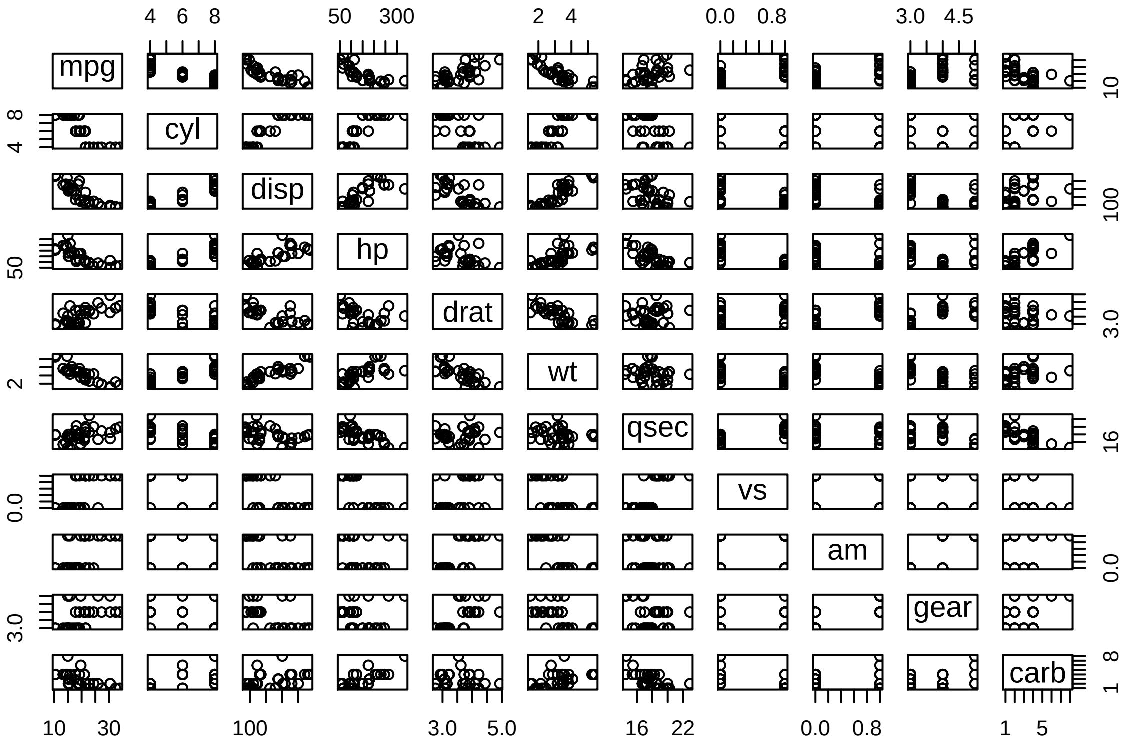
Plotting

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.

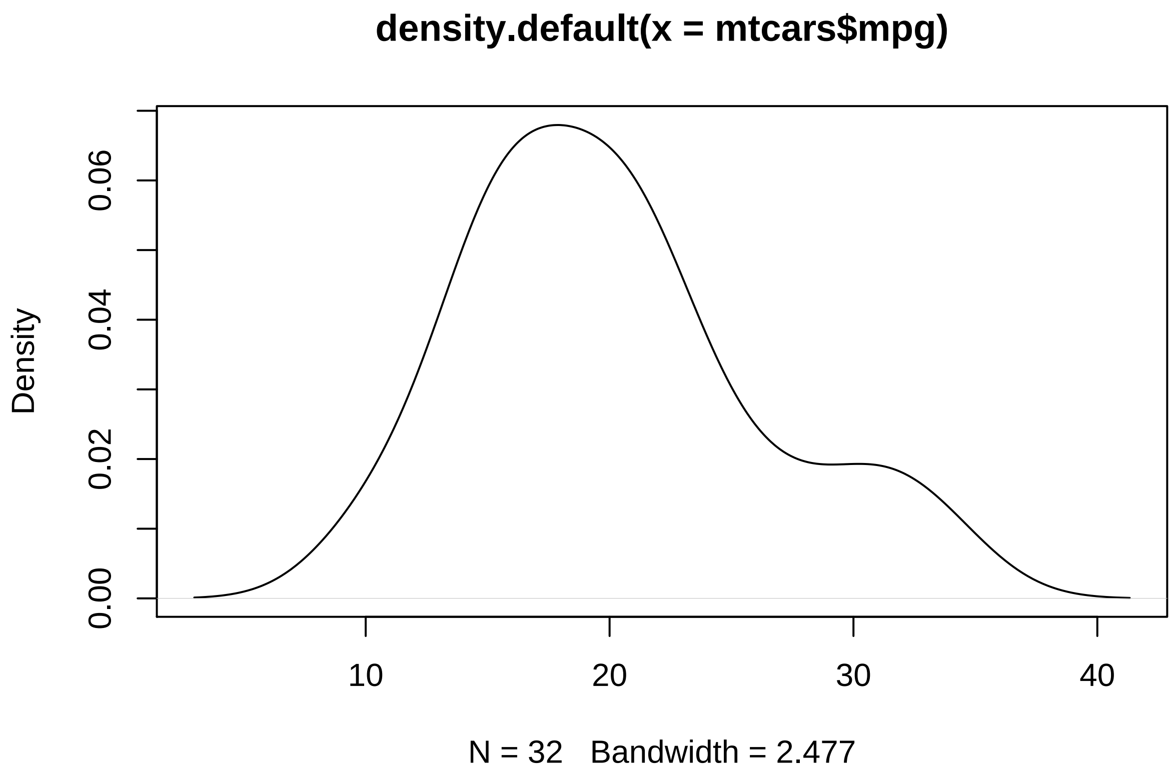
Plotting



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 24.4 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 2 1 1 1 1 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 24.4 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": 2 2 2 1 1 1 1 1 1 1 ...
```

Factors

Finally, for simplicity, lets drop the original values

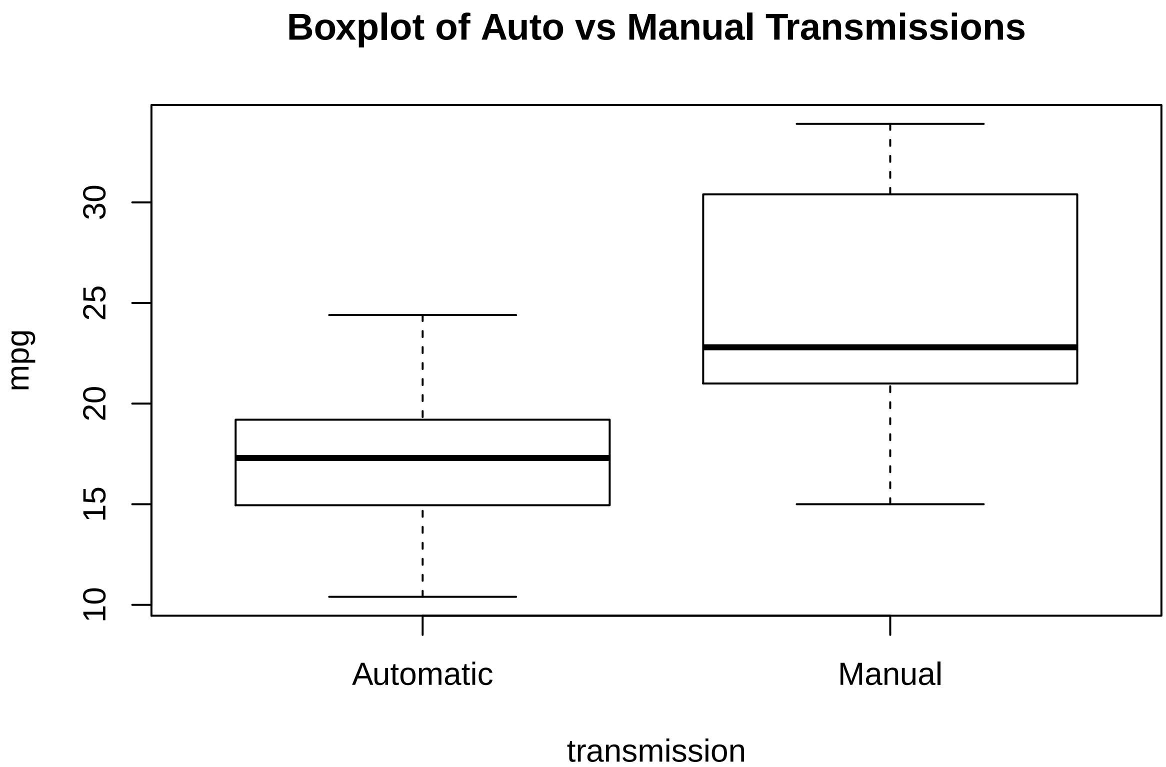
```
mtcars <- subset(mtcars, select=-c(am))
```

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



Linear Regression

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

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

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.

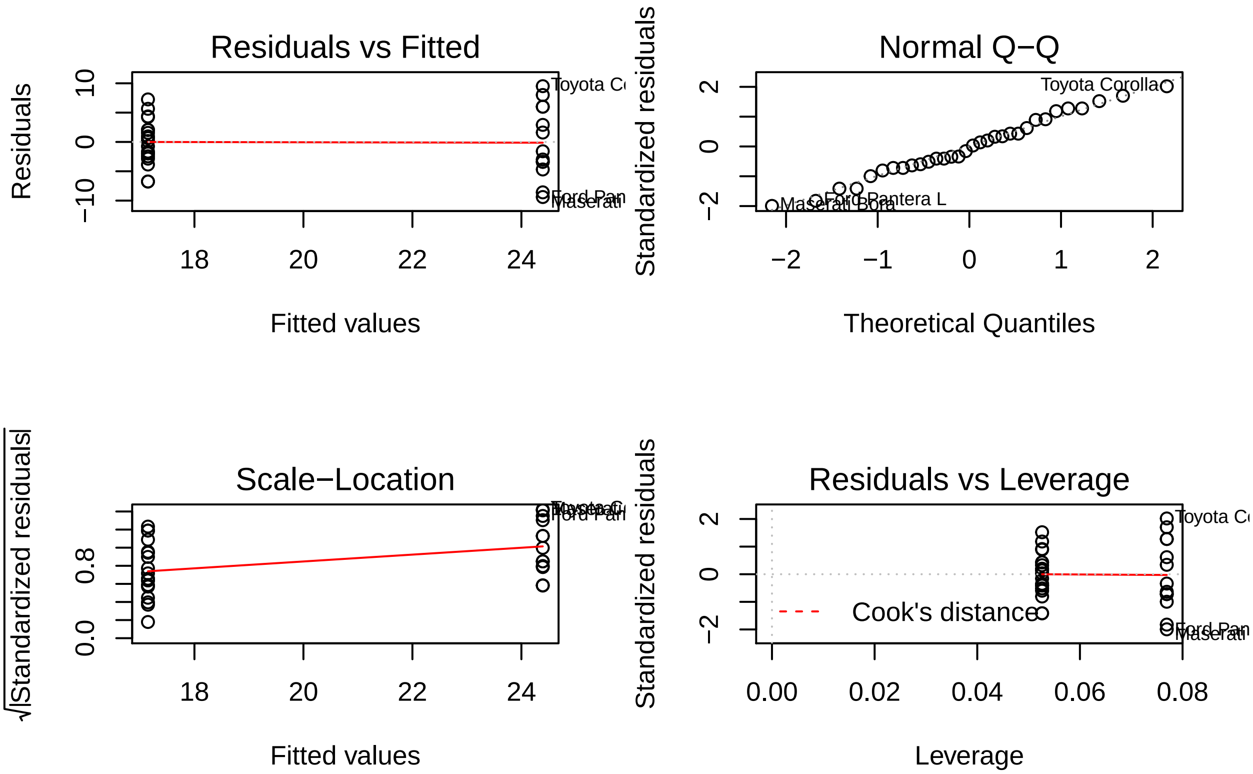
```
model_simple <- lm(mpg~transmission, data=mtcars)
```

Simple Model - Verification

Before we interpret the results, lets verify our assumptions.

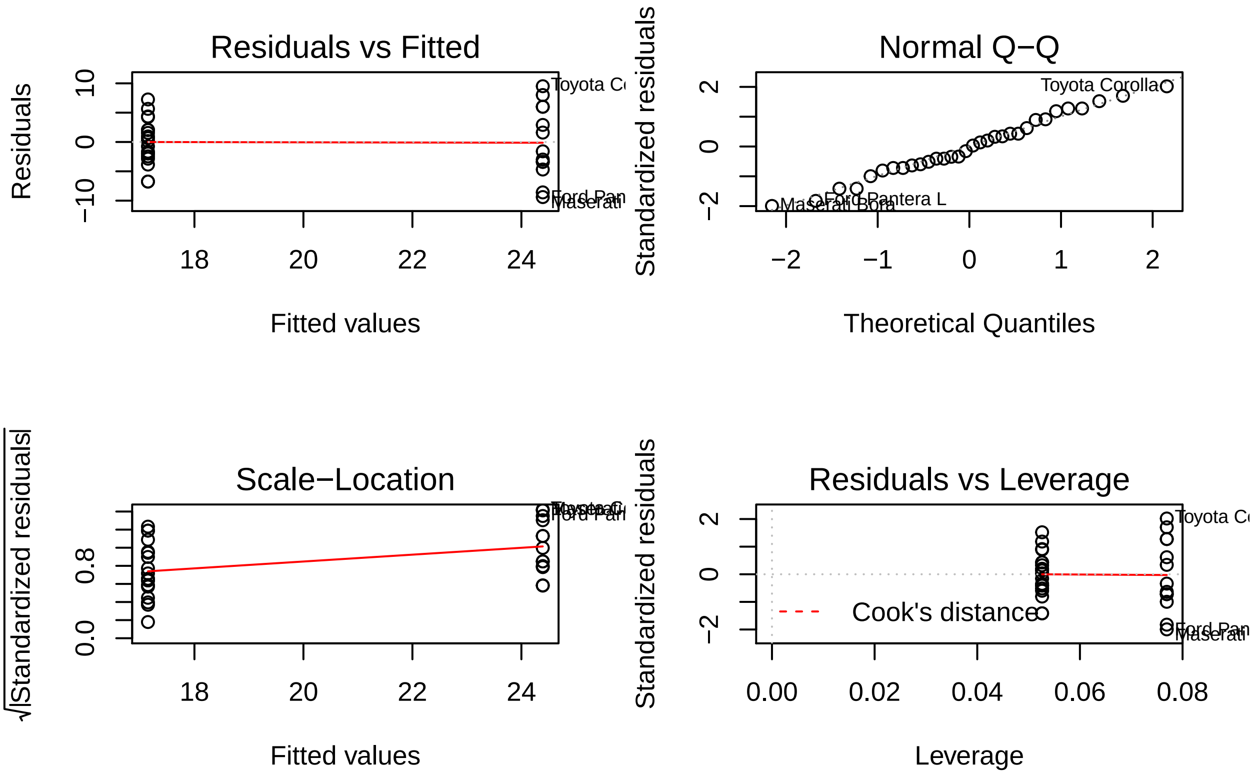
We can use the general graphics par() function to set a variety of graphical parameters. In this case, we want the 4 plots produced by the plot.lm function (remember function overloading!) to print to the same output.

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



Simple Model - Verification

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



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)
```

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 + gear + carb +
##     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 ' ' 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 packages in your environment.

```
install.packages(c('car', 'leaps'))
```

Variance Inflation Factors

Now load the library in your environment.

```
suppressMessages(library(car))
```

Variance Inflation Factors

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

```
(vif = vif(model_kitchensink))
```

```
##          cyl         disp           hp         drat           wt
##    15.373833    21.620241     9.832037     3.374620    15.164887
##         qsec           vs         gear         carb transmission
##     7.527958     4.965873     5.357452     7.908747     4.648487
```

Variance Inflation Factors

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

```
sqrt(vif) > 2
```

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

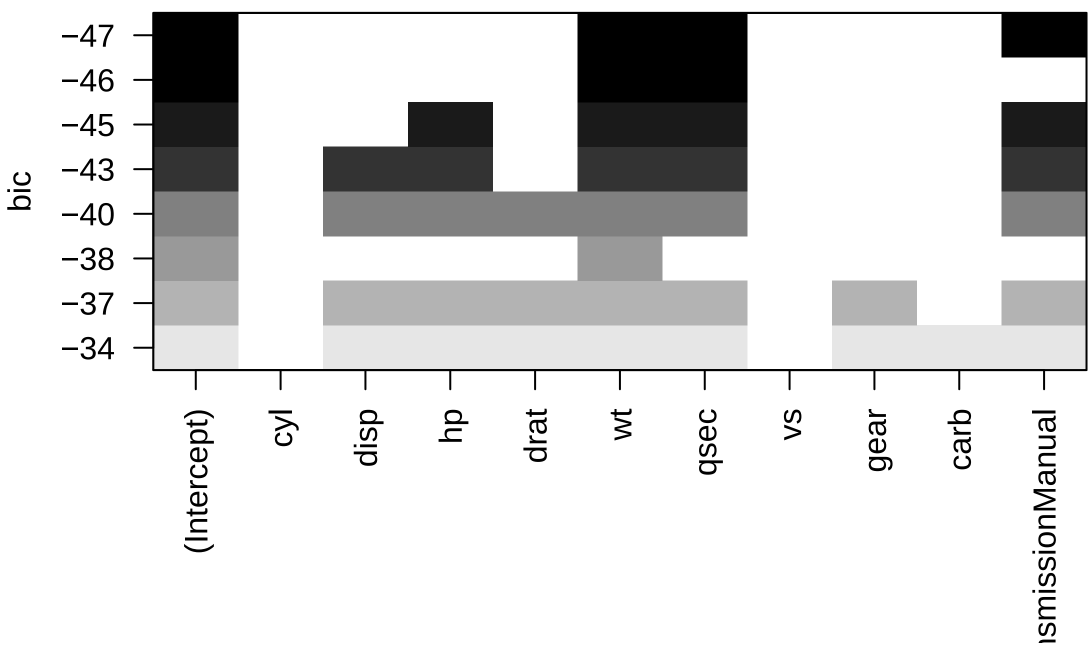
Variable Selection

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

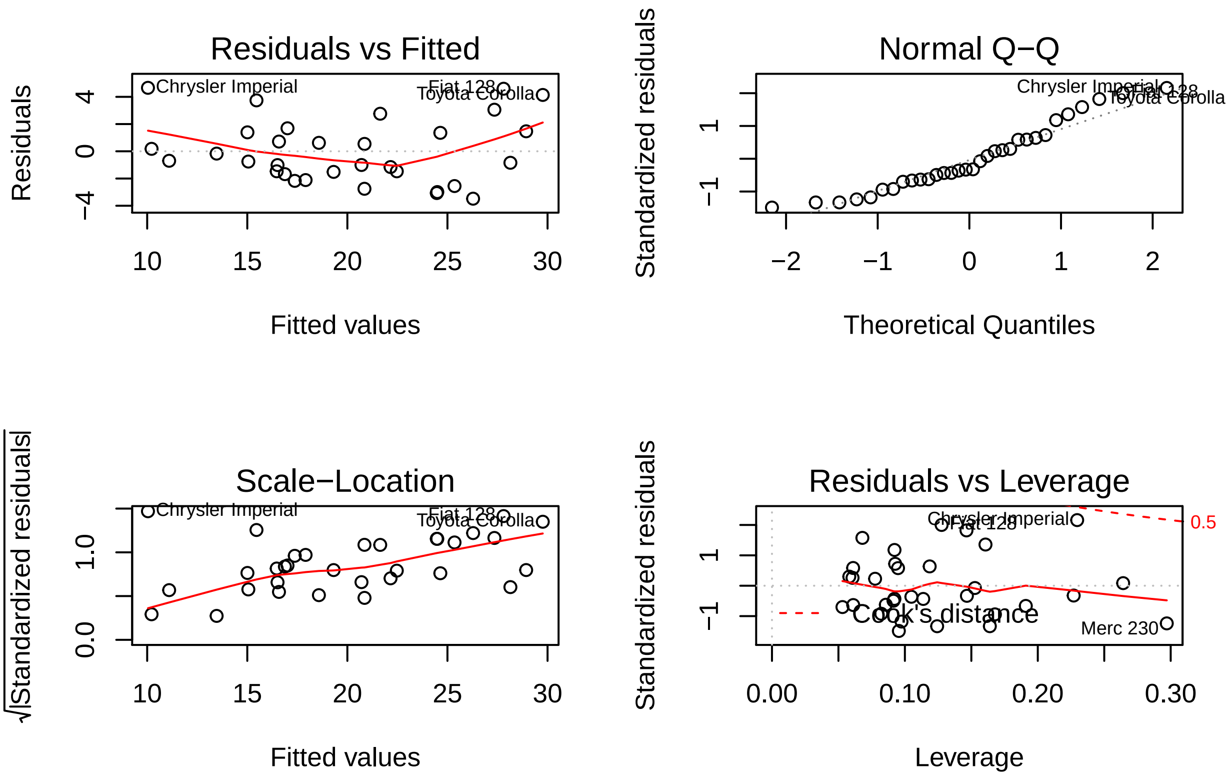
Let's build are final model and repeat the basic validation.

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

Final Model - Verification

Verify linear regression assumptions.

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



Final Model - VIF

Let's re-check for multi-collinearity.

```
vif(model_final)
```

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

Final Model - Summary

```
summary(model_final)
```

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

This example demonstrated basic regression analysis using a mostly clean dataset. In practice, much of the effort is spent cleaning up datasets, for which R is also a very powerful tool.

We also saw how 3rd party packages play an integral role in the R data analysis workflows.