Data Science in Statistical Methods using R

Md Sayeef Alam

21/09/2020

Day 1

Session 1: Application of Regression and Multiple Regression in Data Science

Dr. R. K. Jana, IIM Raipur

Simple addition in R

```
1+1
```

```
## [1] 2
```

Some packages to be installed

```
install.packages("matlib", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("corpcor", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("GPArotation", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("psych", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("FactoMineR", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("tseries", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("corrplot", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("tseries", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("ggpubr", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("tidyverse", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("Hmisc", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("dplyr", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("ggplot2", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("lattice", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("grid", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("DMwR", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("stats", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("nortest", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("MASS", dependencies = T,repos = "http://cran.us.r-project.org")
```

Adding the libraries corresponding to packages.

```
library(dplyr)
library(tseries)
library(matlib)
library(corpcor)
library(GPArotation)
library(psych)
library(FactoMineR)
library(corrplot)
library(ggpubr)
```

```
library(lattice)
library(grid)
library(nortest)
library(stats)
library(DMwR)
library(ggplot2)
library(MASS)
```

Reading xls and xlsx files

```
install.packages("gdata", dep = T,repos = "http://cran.us.r-project.org")
library(gdata)
xls.data = read.xls("file.xls")
```

You need to specify the sheetIndex (sheet number)

```
install.packages("xlsx", dep = T,repos = "http://cran.us.r-project.org")
library(xlsx)
xlsx.data = read.xlsx("file.xlsx", sheetIndex = 1)
```

Linear Regression

Simple Linear Regression

1 dependent (y)

1 independent (x)

Assumptions

1. Relationships between the above two must be linear
2. Residuals should be normally distributed
3. Residuals should be homoscedastic
4. Residuals should be independent

Homoscedasticity means same variance, error term (i.e. distance of the points from the fitted line) should be same across all values of the independent variables.

Heteroscedasticity is when the error varies with the values of the independent variables.

Several measures are there to check for homoscedasticity

```
library(datasets)
data(cars)
```

Lets check the variables inside the dataset

```
names(cars)
```

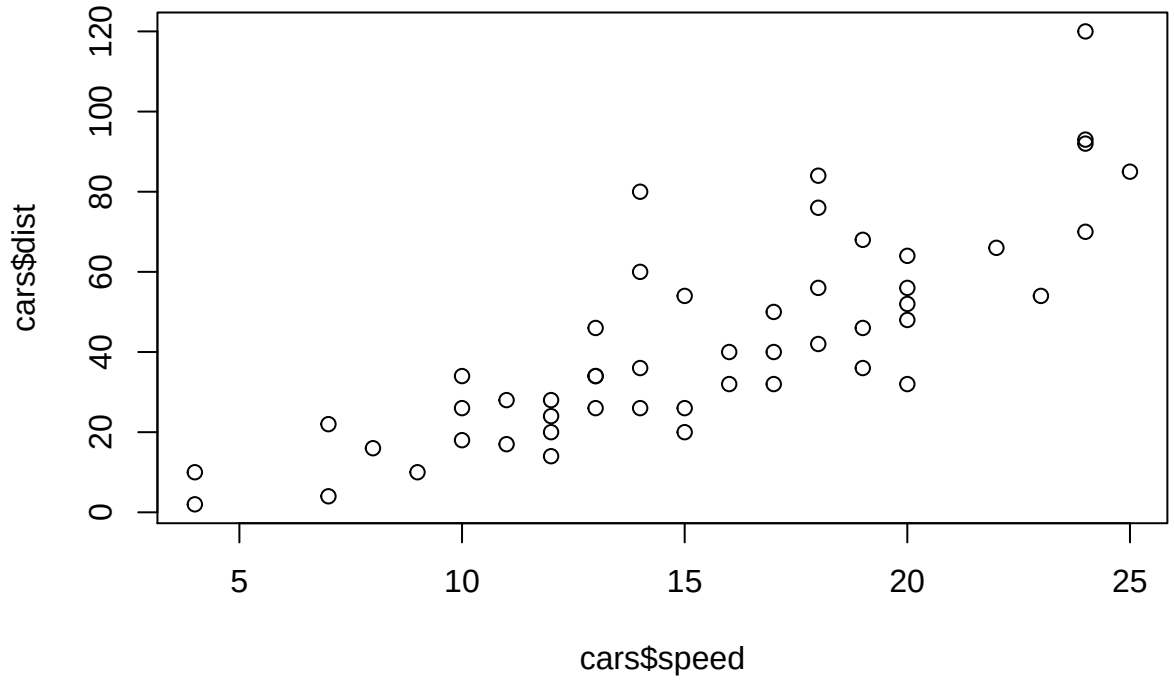
```
## [1] "speed" "dist"
```

```
head(cars)
```

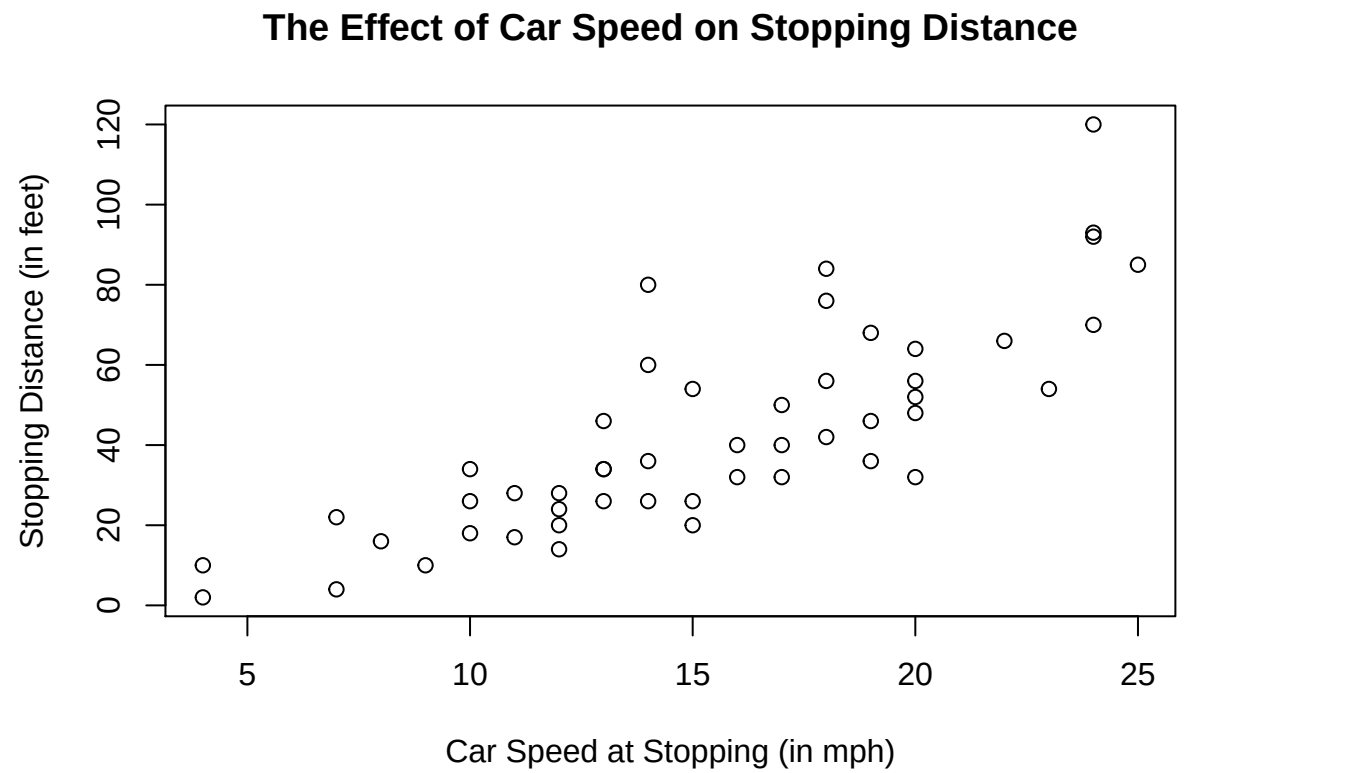
```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Lets plot some parameters specifically speed vs distance

```
plot(cars$speed, cars$dist)
```

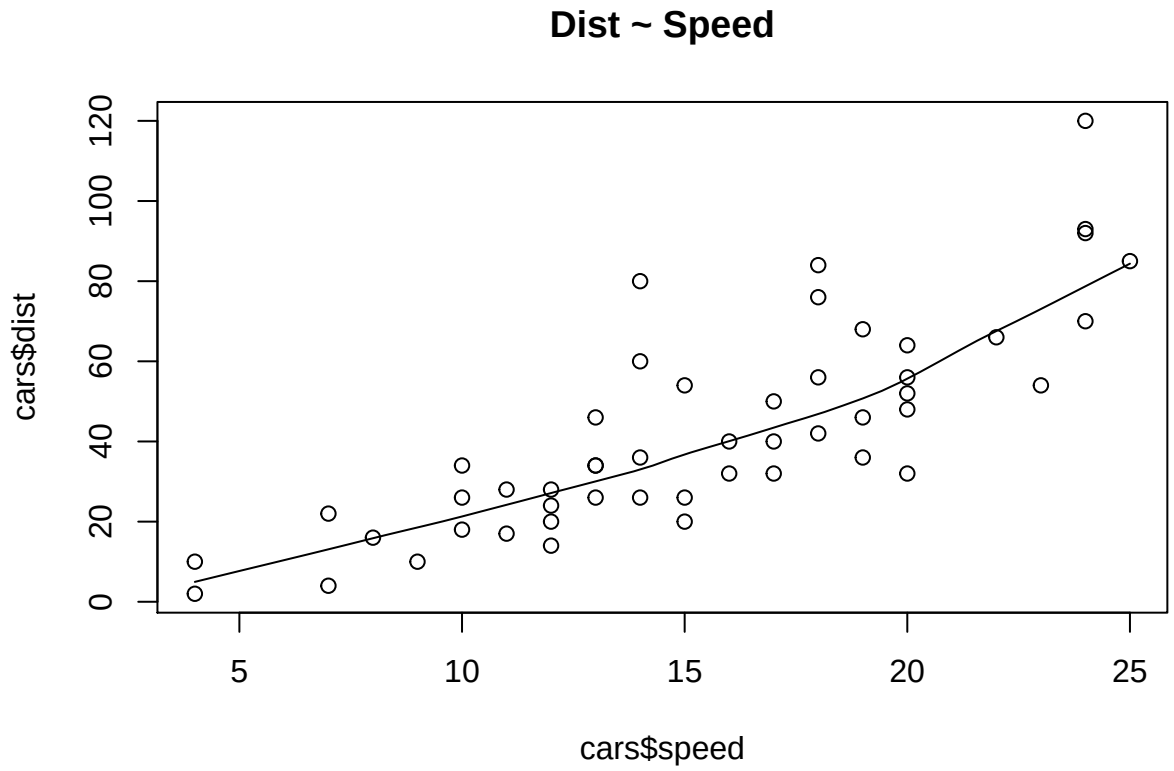



```
plot(cars$speed, cars$dist, xlab = "Car Speed at Stopping (in mph)",
     ylab = "Stopping Distance (in feet)", main = "The Effect of Car Speed on Stopping Distance")
```



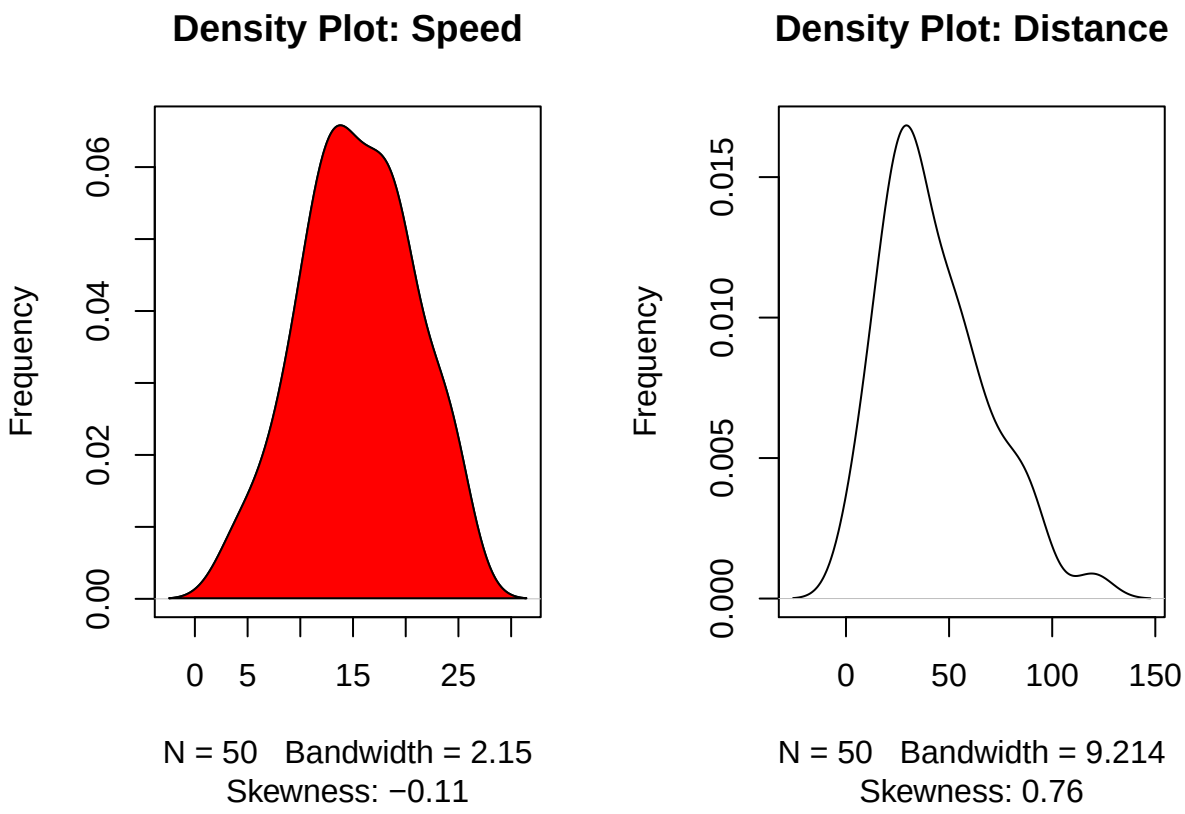
Fitting a smooth line

```
scatter.smooth(x=cars$speed, y=cars$dist, main="Dist ~ Speed")
```



Density plots for speed and distance

```
library(e1071)
par(mfrow=c(1, 2))

plot(density(cars$speed), main="Density Plot: Speed", ylab="Frequency", sub=paste("Skewness:", round(e1
polygon(density(cars$speed), col="red")

plot(density(cars$dist), main="Density Plot: Distance", ylab="Frequency", sub=paste("Skewness:", round(
```



Linear regression model fitting

```
carmod <- lm(dist ~ speed, data = cars)
summary(carmod)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

95% CI

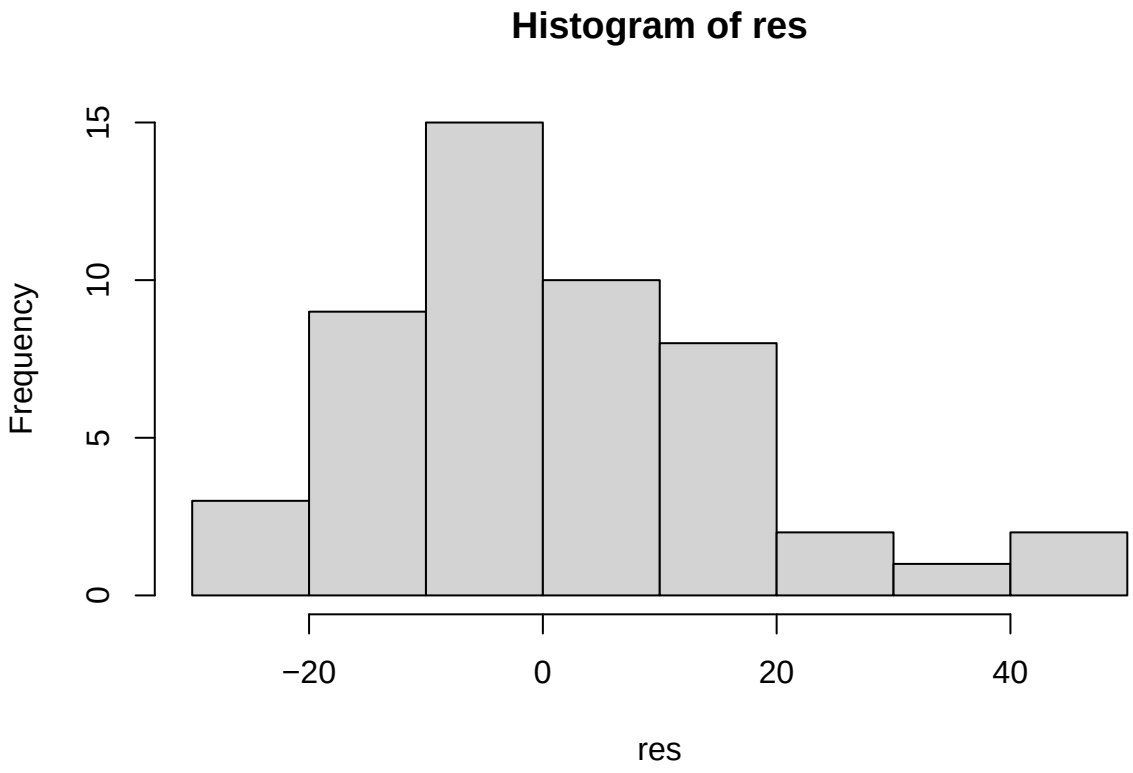
```
confint(carmod, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) -31.167850 -3.990340
```

```
## speed         3.096964  4.767853
```

Normality of residuals check

```
res = carmod$residuals
hist(res)
```



Interpretation

The coefficients in linear regression model states that with a unit change in x how much change is expected in y.

Session 2: Data Science & Sample Survey

Prof. G. N. Singh, IIT (ISM) Dhanbad

Word Statistics

In a literal sense

Plural sense some sort of data numerical figures in our day to day arising, runs and all figures are called statistics

In singular collection of methods and principles in a book,

Procedure to collection, analyse and interpret the data is called statistics

Statistics never claims 100% accuracy

Statistics is the science of decision making. As no decision is free from error.

Hope that PPTs will be provided soon.

Day 2

Session 3: Introduction to Statistical Methods in Data Science

Dr. A. K. Sinha, NIT Raipur

Theory and PPT will be available.

Session 4: Introduction to R

Dr. Anup Kumar Sharma, NIT Raipur

Theory and PPT will be available.

Session 5: Application of Data Science in Sample Survey

Prof. G. N. Singh, IIT (ISM) Dhanbad

Theory and PPT will be available.

Session 6: Graphical representation and normality testing in R

Dr. Dhaval Maheta, VNSGU Surat

```
mtcars
```

```
##                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic        30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla     33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona      21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger   15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin        15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28         13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird   19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9          27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2      26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa       30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L     15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino       19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
```

```
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

```
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Find the mean for all columns

the number in between is the parameter denoting the 1 = row and 2 = column. row mean is useless so we are looking at column mean

```
apply(mtcars,2,mean)
```

```
##        mpg        cyl       disp         hp       drat         wt       qsec
##  20.090625   6.187500 230.721875 146.687500   3.596563   3.217250  17.848750
##         vs         am       gear       carb
##   0.437500   0.406250   3.687500   2.812500
```

Now similarly for median and mode

```
apply(mtcars,2,median)
```

```
##     mpg     cyl    disp      hp    drat      wt    qsec      vs      am    gear
##  19.200   6.000 196.300 123.000   3.695   3.325  17.710   0.000   0.000   4.000
##    carb
##   2.000
```

```
apply(mtcars,2,mode)
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec        vs
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        am      gear      carb
## "numeric" "numeric" "numeric"
```

aggregate function helps to calculate the required function (mean/median/mode/sd) for each category of independent variable

```
aggregate(mpg~am,FUN = mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

```
aggregate(mpg~am,FUN = median)
```

```
##   am  mpg
## 1  0 17.3
## 2  1 22.8
```

```
aggregate(mpg~am,FUN = mode)
```

```
##   am     mpg
## 1  0 numeric
## 2  1 numeric
```

```
aggregate(mpg~am,FUN = sd)
```

```
##   am      mpg
## 1  0 3.833966
```

```
## 2  1 6.166504
```

Find 3 way table to summary statistics and describeBy (available in psych library)

```
aggregate(mpg~am+vs,FUN = mean)
```

```
##   am vs      mpg
## 1  0  0 15.05000
## 2  1  0 19.75000
## 3  0  1 20.74286
## 4  1  1 28.37143
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
describeBy(mpg,am)
```

```
##
##  Descriptive statistics by group
## group: 0
##    vars  n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 19 17.15 3.83   17.3   17.12 3.11 10.4 24.4    14 0.01     -0.8 0.88
## ------------------------------------------------------------
## group: 1
##    vars  n  mean   sd median trimmed  mad min  max range skew kurtosis   se
## X1    1 13 24.39 6.17   22.8   24.38 6.67  15 33.9  18.9 0.05    -1.46 1.71
```

best descriptive summarizer called the stargazer, the flip = T command helps to transpose the rows and columns

```
install.packages("stargazer", dependencies = T,repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/yp/0237rgk11t35swrh_2f9h_200000gn/T//Rtmpulv4lW/downloaded_packages
```

```
library(stargazer)

stargazer(mtcars,type = "text", title = "Descriptive Stats", digits = 1)
```

```
##
## Descriptive Stats
## ===================================================
## Statistic N  Mean  St. Dev. Min  Pctl(25) Pctl(75) Max
## ---------------------------------------------------
## mpg       32 20.1    6.0     10    15.4     22.8    34
## cyl       32  6.2    1.8     4      4        8      8
## disp      32 230.7  123.9    71   120.8     326    472
## hp        32 146.7   68.6    52    96.5     180    335
## drat      32  3.6    0.5    2.8    3.1      3.9    4.9
## wt        32  3.2    1.0    1.5    2.6      3.6    5.4
## qsec      32 17.8    1.8    14.5   16.9     18.9   22.9
## vs        32  0.4    0.5     0      0        1      1
## am        32  0.4    0.5     0      0        1      1
## gear      32  3.7    0.7     3      3        4      5
## carb      32  2.8    1.6     1      2        4      8
## ---------------------------------------------------
```

```
stargazer(mtcars,type = "text", title = "Descriptive Stats", digits = 1, flip = T)
```

```
##
## Descriptive Stats
## ============================================================
## Statistic mpg  cyl disp   hp   drat wt  qsec vs  am  gear carb
## ------------------------------------------------------------
## N          32  32   32    32    32  32   32  32  32   32   32
## Mean      20.1 6.2 230.7 146.7 3.6  3.2 17.8 0.4 0.4 3.7  2.8
## St. Dev.  6.0  1.8 123.9 68.6  0.5  1.0 1.8  0.5 0.5 0.7  1.6
## Min        10   4   71    52   2.8  1.5 14.5  0   0   3    1
## Pctl(25)  15.4  4  120.8 96.5  3.1  2.6 16.9  0   0   3    2
## Pctl(75)  22.8  8   326   180  3.9  3.6 18.9  1   1   4    4
## Max        34   8   472   335  4.9  5.4 22.9  1   1   5    8
## ------------------------------------------------------------
```

Try the following codes to obtain data like SPSS

```
install.packages("summarytools", dependencies = T,repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/yp/0237rgk11t35swrh_2f9h_200000gn/T//Rtmpulv4lW/downloaded_packages
```

```
install.packages("ellipsis", dependencies = T,repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/yp/0237rgk11t35swrh_2f9h_200000gn/T//Rtmpulv4lW/downloaded_packages
```

```
library(summarytools)
library(ellipsis)
```

```
attach(mtcars)
```

```
## The following objects are masked from mtcars (pos = 6):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt

## The following object is masked from package:ggplot2:
##
##     mpg
```

```
summarytools::descr(mtcars)
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

## Descriptive Statistics
## mtcars
## N: 32
##
##                         am     carb      cyl     disp     drat     gear        hp      mpg     qsec
## ----------------- -------- -------- -------- -------- -------- -------- -------- -------- --------
##              Mean     0.41     2.81     6.19   230.72     3.60     3.69   146.69    20.09    17.85
##           Std.Dev     0.50     1.62     1.79   123.94     0.53     0.74    68.56     6.03     1.79
##               Min     0.00     1.00     4.00    71.10     2.76     3.00    52.00    10.40    14.50
##                Q1     0.00     2.00     4.00   120.65     3.08     3.00    96.00    15.35    16.88
##            Median     0.00     2.00     6.00   196.30     3.70     4.00   123.00    19.20    17.71
##                Q3     1.00     4.00     8.00   334.00     3.92     4.00   180.00    22.80    18.90
##               Max     1.00     8.00     8.00   472.00     4.93     5.00   335.00    33.90    22.90
##               MAD     0.00     1.48     2.97   140.48     0.70     1.48    77.10     5.41     1.42
##               IQR     1.00     2.00     4.00   205.18     0.84     1.00    83.50     7.38     2.01
##                CV     1.23     0.57     0.29     0.54     0.15     0.20     0.47     0.30     0.10
##          Skewness     0.36     1.05    -0.17     0.38     0.27     0.53     0.73     0.61     0.37
##       SE.Skewness     0.41     0.41     0.41     0.41     0.41     0.41     0.41     0.41     0.41
##          Kurtosis    -1.92     1.26    -1.76    -1.21    -0.71    -1.07    -0.14    -0.37     0.34
##           N.Valid    32.00    32.00    32.00    32.00    32.00    32.00    32.00    32.00    32.00
##         Pct.Valid   100.00   100.00   100.00   100.00   100.00   100.00   100.00   100.00   100.00
##
## Table: Table continues below
##
##
##
##                         vs       wt
## ----------------- -------- --------
##              Mean     0.44     3.22
##           Std.Dev     0.50     0.98
##               Min     0.00     1.51
##                Q1     0.00     2.54
```

```
##             Median     0.00     3.33
##                 Q3     1.00     3.65
##                Max     1.00     5.42
##                MAD     0.00     0.77
##                IQR     1.00     1.03
##                 CV     1.15     0.30
##           Skewness     0.24     0.42
##        SE.Skewness     0.41     0.41
##           Kurtosis    -2.00    -0.02
##            N.Valid    32.00    32.00
##          Pct.Valid   100.00   100.00
```

```
summarytools::freq(am)
```

```
## Frequencies
## am
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           0     19     59.38          59.38     59.38          59.38
##           1     13     40.62         100.00     40.62         100.00
##        <NA>      0                               0.00         100.00
##       Total     32    100.00         100.00    100.00         100.00
```

```
summarytools::ctable(am,vs)
```

```
## Cross-Tabulation, Row Proportions
## am * vs
##
## ------- ---- ------------ ------------ -------------
##           vs            0            1         Total
##      am
##       0        12 (63.2%)    7 (36.8%)   19 (100.0%)
##       1         6 (46.2%)    7 (53.8%)   13 (100.0%)
##   Total        18 (56.2%)   14 (43.8%)   32 (100.0%)
## ------- ---- ------------ ------------ -------------
```

```
summarytools::dfSummary(mtcars)
```

```
## Data Frame Summary
## mtcars
## Dimensions: 32 x 11
## Duplicates: 0
##
## ---------------------------------------------------------------------------------------------------
## No   Variable    Stats / Values               Freqs (% of Valid)   Graph                   Valid
## ---- ----------- ---------------------------- -------------------- ----------------------- --------
## 1    mpg         Mean (sd) : 20.1 (6)         25 distinct values     :                     32
##      [numeric]   min < med < max:                                    : .                   (100%)
##                  10.4 < 19.2 < 33.9                               . : :
##                  IQR (CV) : 7.4 (0.3)                             : : :   .
##                                                                   : : : : :
##
## 2    cyl         Mean (sd) : 6.2 (1.8)        4 : 11 (34.4%)       IIIIII                  32
##      [numeric]   min < med < max:             6 :  7 (21.9%)       IIII                    (100%)
```

```
##                 4 < 6 < 8                  8 : 14 (43.8%)      IIIIIIII
##                 IQR (CV) : 4 (0.3)
##
## 3    disp       Mean (sd) : 230.7 (123.9)  27 distinct values    :                    32
##      [numeric]  min < med < max:                                 . :                    (100%)
##                 71.1 < 196.3 < 472                              : : :   : : :
##                 IQR (CV) : 205.2 (0.5)                          : : :   : : :   .
##                                                                 : : : . : : : . . :
##
## 4    hp         Mean (sd) : 146.7 (68.6)   22 distinct values   . :                    32
##      [numeric]  min < med < max:                                 : :                    (100%)
##                 52 < 123 < 335                                  : : : .
##                 IQR (CV) : 83.5 (0.5)                           : : : :
##                                                                 : : : : . .
##
## 5    drat       Mean (sd) : 3.6 (0.5)      22 distinct values      :                   32
##      [numeric]  min < med < max:                                  : :                   (100%)
##                 2.8 < 3.7 < 4.9                                  : : .
##                 IQR (CV) : 0.8 (0.1)                          . : : :
##                                                               : : : : .
##
## 6    wt         Mean (sd) : 3.2 (1)        29 distinct values       :                  32
##      [numeric]  min < med < max:                                    : :                 (100%)
##                 1.5 < 3.3 < 5.4                                    : :
##                 IQR (CV) : 1 (0.3)                            : : : : :     .
##                                                               : : : : : .   :
##
## 7    qsec       Mean (sd) : 17.8 (1.8)     30 distinct values       :                  32
##      [numeric]  min < med < max:                                    :                  (100%)
##                 14.5 < 17.7 < 22.9                                 : :
##                 IQR (CV) : 2 (0.1)                               . : : : :
##                                                               : : : : : : : :   .
##
## 8    vs         Min  : 0                   0 : 18 (56.2%)      IIIIIIIIIII            32
##      [numeric]  Mean : 0.4                 1 : 14 (43.8%)      IIIIIIII               (100%)
##                 Max  : 1
##
## 9    am         Min  : 0                   0 : 19 (59.4%)      IIIIIIIIIII            32
##      [numeric]  Mean : 0.4                 1 : 13 (40.6%)      IIIIIIII               (100%)
##                 Max  : 1
##
## 10   gear       Mean (sd) : 3.7 (0.7)      3 : 15 (46.9%)      IIIIIIIII              32
##      [numeric]  min < med < max:           4 : 12 (37.5%)      IIIIIII                (100%)
##                 3 < 4 < 5                  5 :  5 (15.6%)      III
##                 IQR (CV) : 1 (0.2)
##
## 11   carb       Mean (sd) : 2.8 (1.6)      1 :  7 (21.9%)      IIII                   32
##      [numeric]  min < med < max:           2 : 10 (31.2%)      IIIIII                 (100%)
##                 1 < 2 < 8                  3 :  3 ( 9.4%)      I
##                 IQR (CV) : 2 (0.6)         4 : 10 (31.2%)      IIIIII
##                                            6 :  1 ( 3.1%)
##                                            8 :  1 ( 3.1%)
## ---------------------------------------------------------------------------------------------------
```

Graphical representation of data

Using a new data set called “Orange”

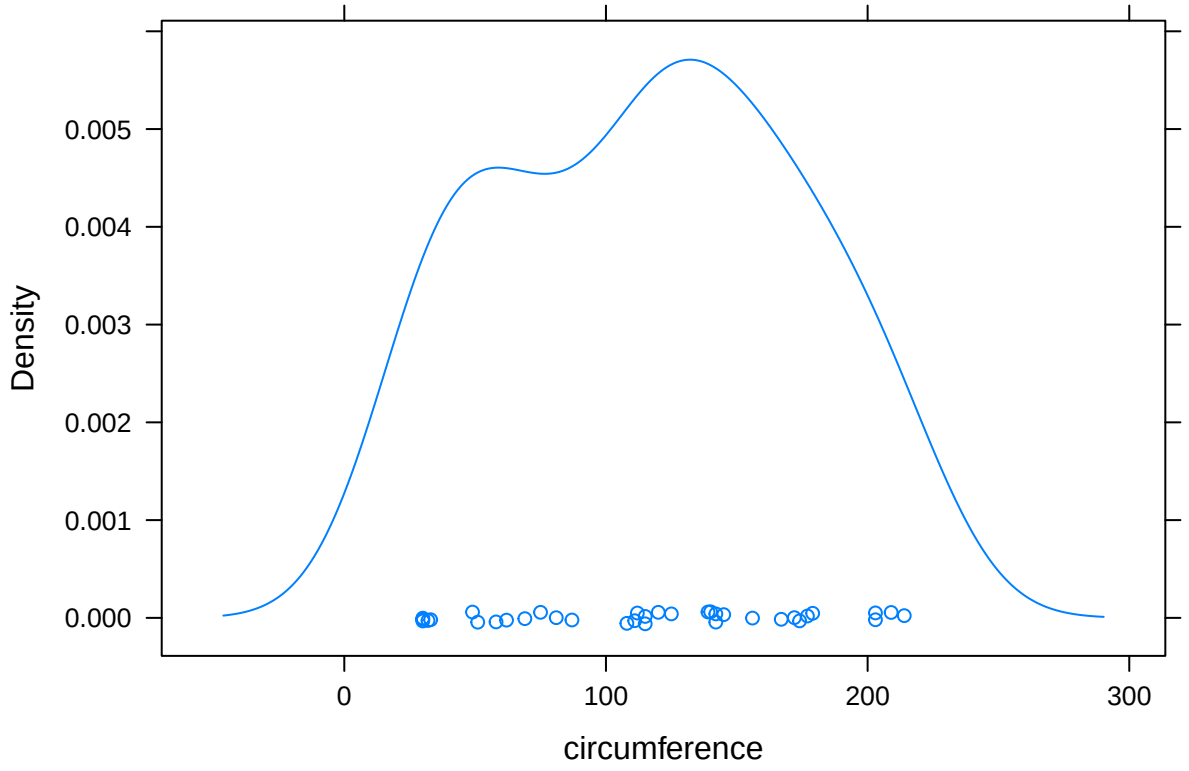
```
Orange
```

```
## Grouped Data: circumference ~ age | Tree
##    Tree  age circumference
## 1     1  118            30
## 2     1  484            58
## 3     1  664            87
## 4     1 1004           115
## 5     1 1231           120
## 6     1 1372           142
## 7     1 1582           145
## 8     2  118            33
## 9     2  484            69
## 10    2  664           111
## 11    2 1004           156
## 12    2 1231           172
## 13    2 1372           203
## 14    2 1582           203
## 15    3  118            30
## 16    3  484            51
## 17    3  664            75
## 18    3 1004           108
## 19    3 1231           115
## 20    3 1372           139
## 21    3 1582           140
## 22    4  118            32
## 23    4  484            62
## 24    4  664           112
## 25    4 1004           167
## 26    4 1231           179
## 27    4 1372           209
## 28    4 1582           214
## 29    5  118            30
## 30    5  484            49
## 31    5  664            81
## 32    5 1004           125
## 33    5 1231           142
## 34    5 1372           174
## 35    5 1582           177
```

```
attach(Orange)
```

Density plot of circumference

```
densityplot(~circumference)
```

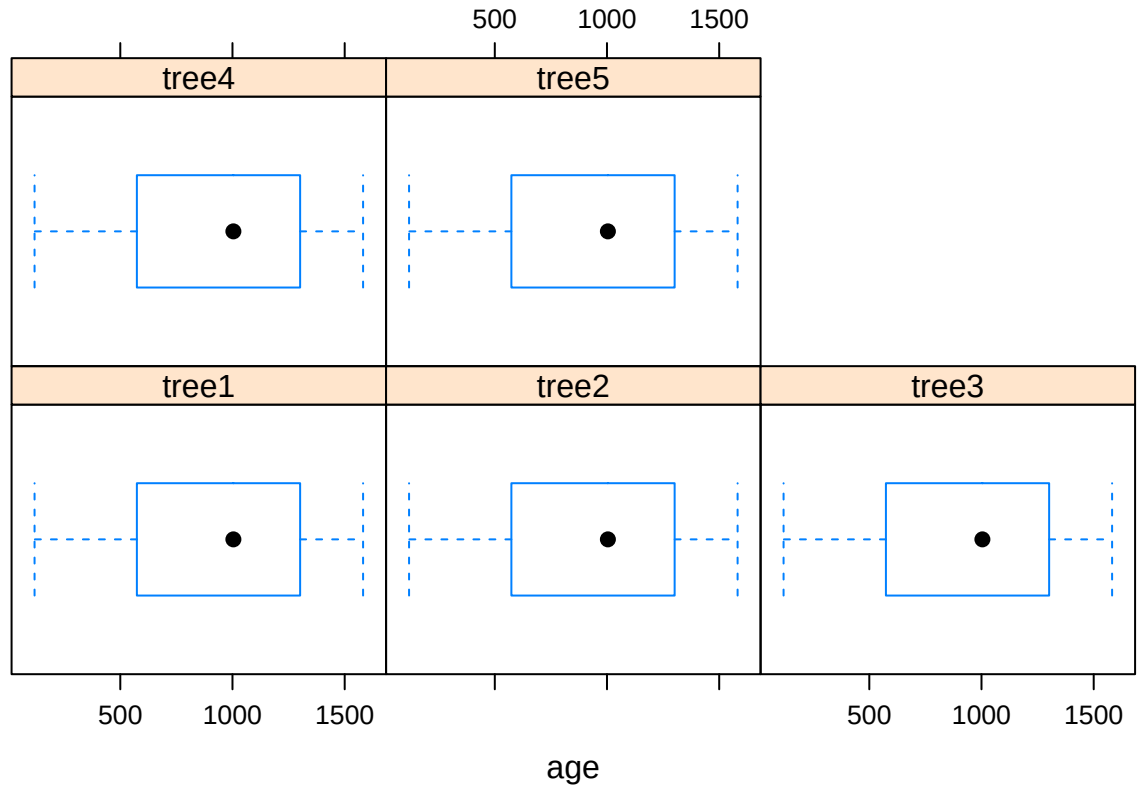


Converting into categorical/factor variable

```
tree.f = factor(Tree, levels = c(1,2,3,4,5), labels = c("tree1","tree2","tree3","tree4","tree5"))
```

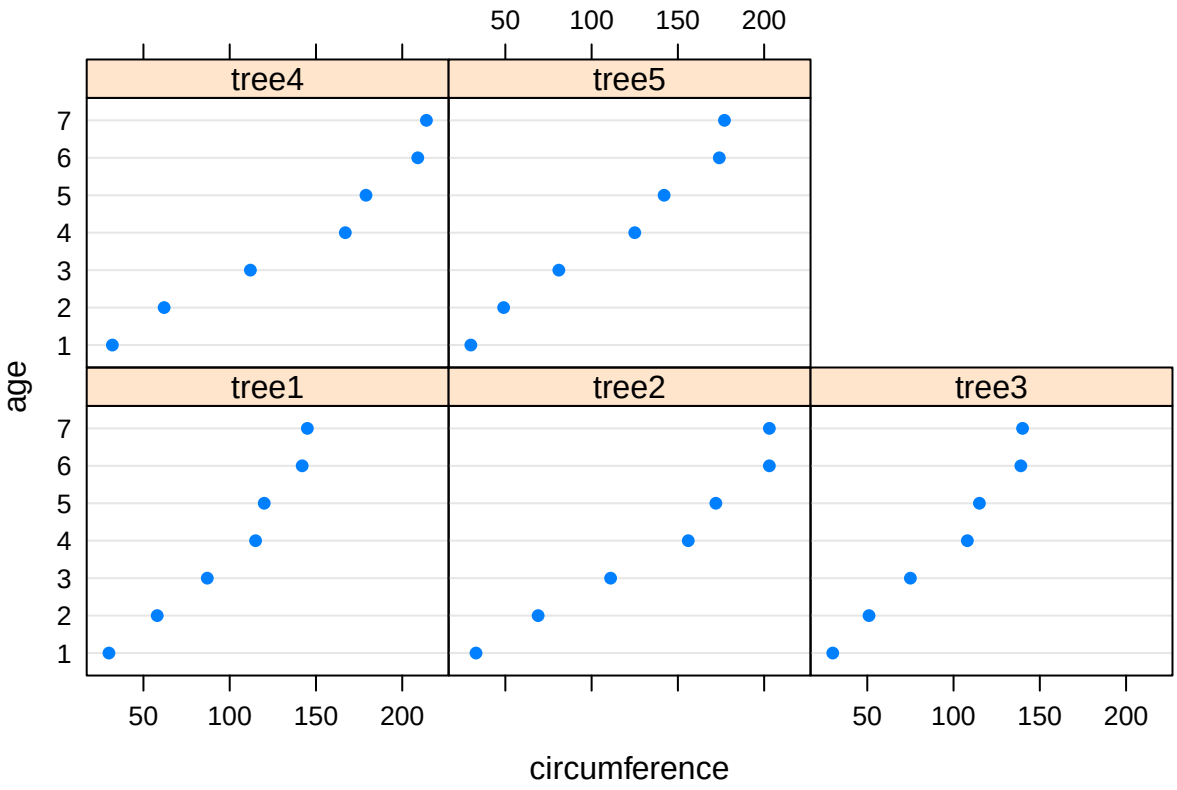
Boxplot of age of trees

```
bwplot(~age|tree.f)
```



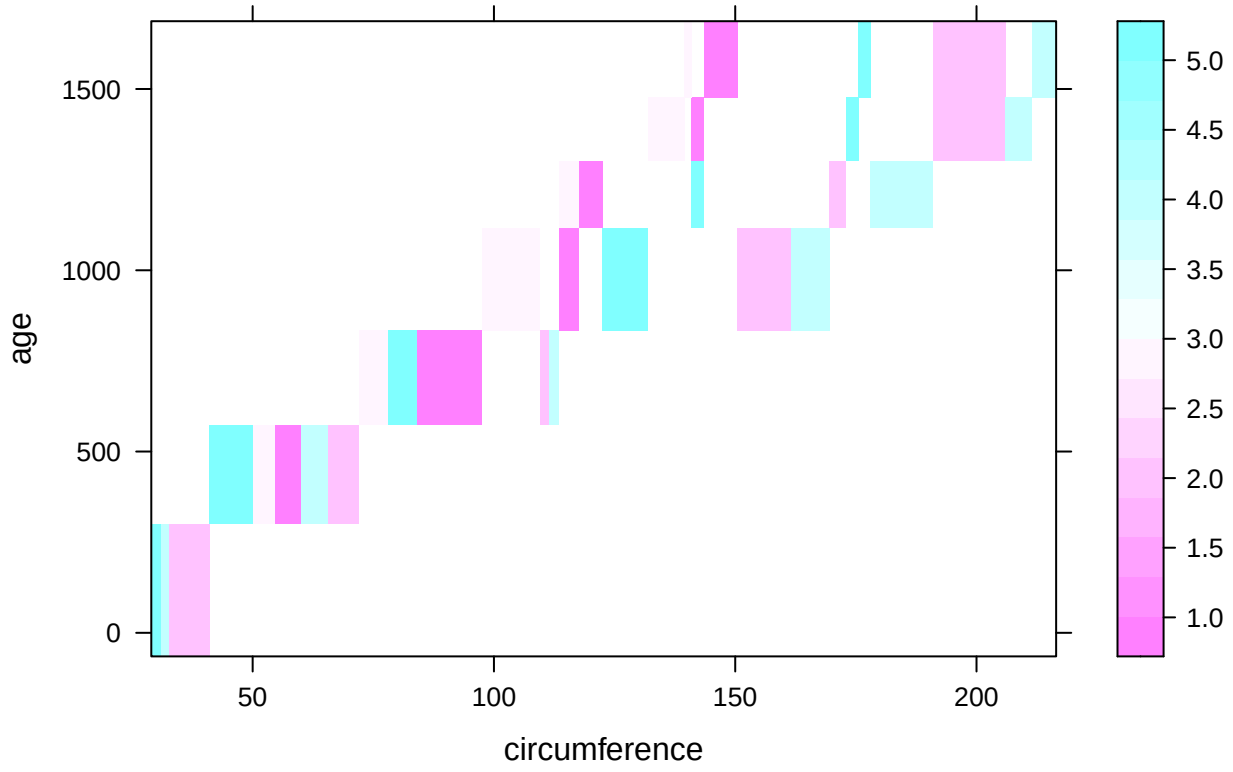
Dot plot

```
dotplot(age~circumference|tree.f)
```



Level plot

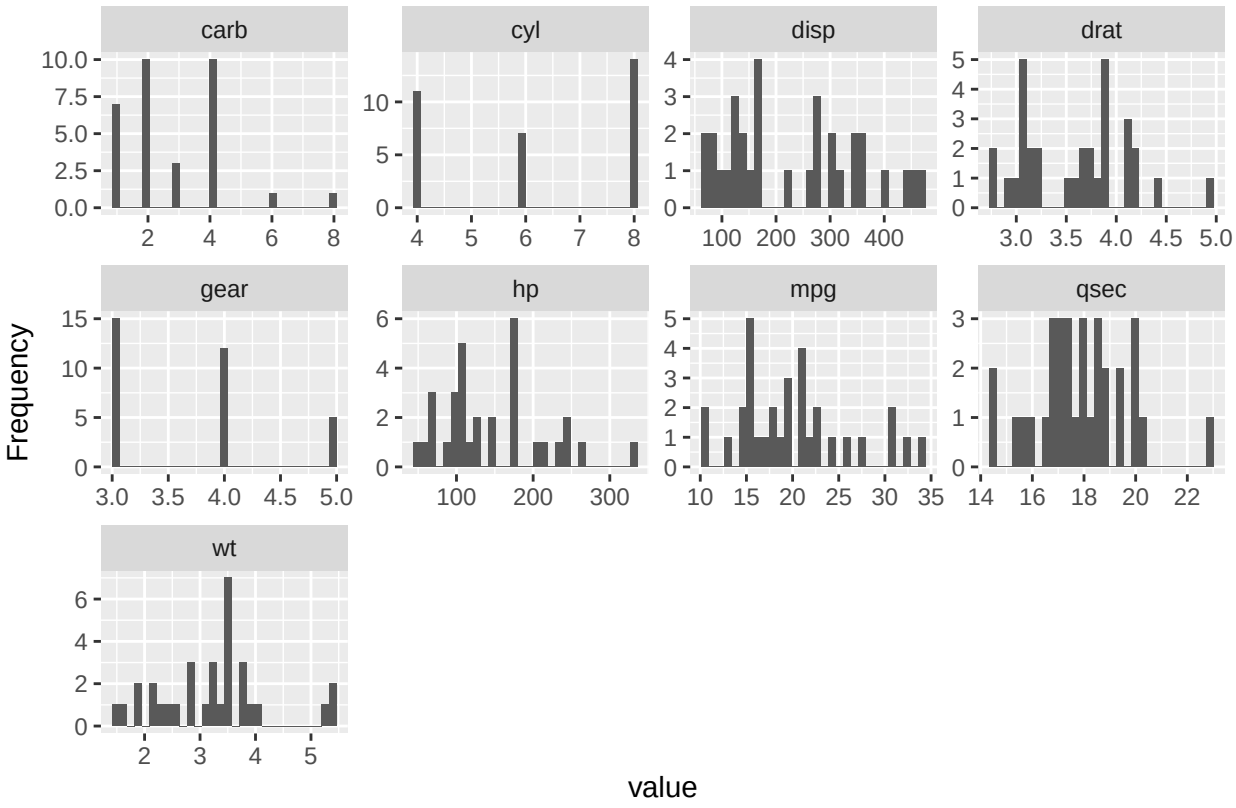
```
levelplot(tree.f~circumference*age)
```




```
install.packages("DataExplorer", dependencies = T,repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/yp/0237rgk11t35swrh_2f9h_200000gn/T//Rtmpulv4lW/downloaded_packages
```

```
library(DataExplorer)

plot_histogram(mtcars)
```



Not even interested to write a single line of command, this is very sexy and appealing for data cleaning

```
install.packages("esquisse", dependencies = T,repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/yp/0237rgk11t35swrh_2f9h_200000gn/T//Rtmpulv4lW/downloaded_packages
```

```
library(esquisse)

esquisse::esquisser(mtcars)
```

Even new packages click and play, contingency tables, summary stats

```
install.packages("Rcmdr", dependencies = T,repos = "http://cran.us.r-project.org")
library(Rcmdr)
```

Lets use our own dataset Employee dataset but for now use mtcars

```
mtcars
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
```

```
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

```
attach(mtcars)
```

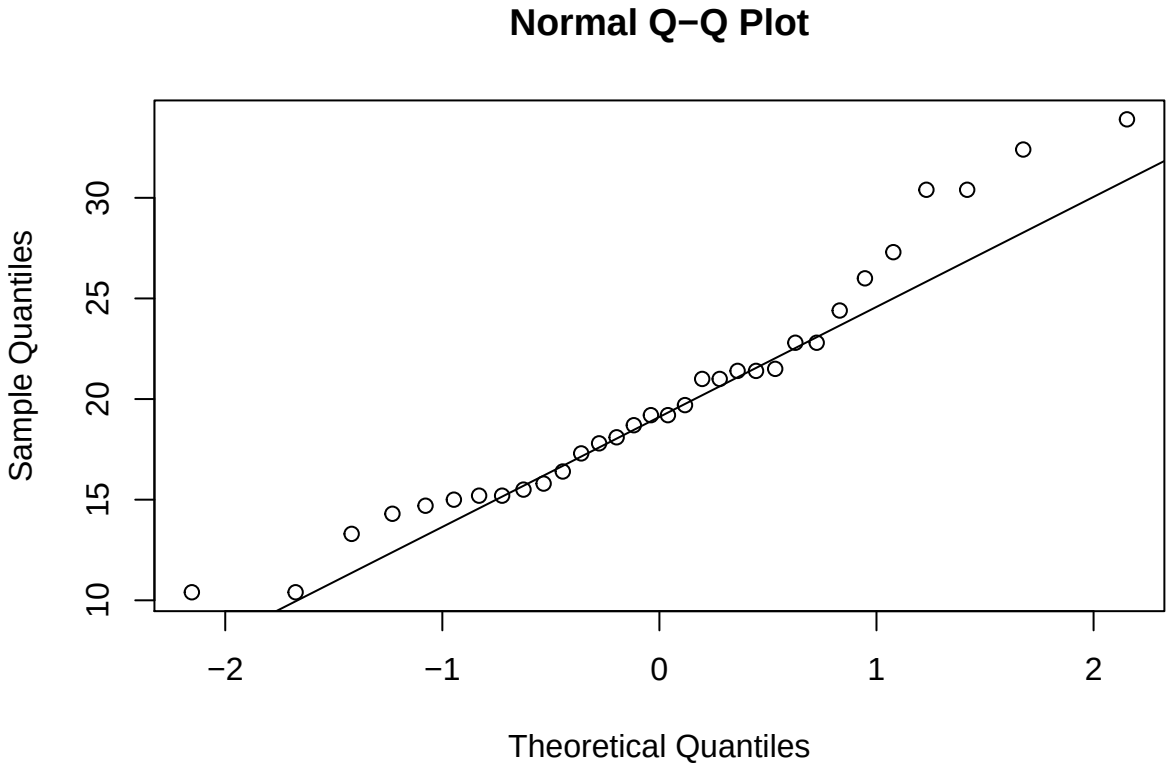
```
## The following objects are masked from mtcars (pos = 14):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following objects are masked from mtcars (pos = 18):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Normality Checks using graphics but graphics is not 100% so we use test for rejection of h0: normal distribution and h1: not normal ; hence if p value < 0.05 then the data is not normal

```
qqnorm(mpg)
qqline(mpg)
```




```
shapiro.test(mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mpg
## W = 0.94756, p-value = 0.1229
```

tm, quanteda: for unstructured data

tseries: for timeseries

animate: can be used to animate any plot type, written by Yihui Xie

gganimate: used to specifically animate ggplot graphics, written by Thomas Lin Pedersen

plotly: an interactive plotting library which has animation features

googlevis: has a flash based motion chart option

plspm for SEM

Day 3

Session 7: Regression and Multiple Regression in R

Dr. R. K. Jana, IIM Raipur

Before beginning with regression simple visualization is necessary to see association of variables

Install the package "MASS" for dataset mtcars

Looking at the dataset

head() : function to display first 6 rows

tail() : function to display last 6 rows

names() : function to display names of all the columns

attach() : function to attach the dataset helping us to not rewrite when accessing its columns(variables)

```
data(mtcars)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
names(mtcars)
```

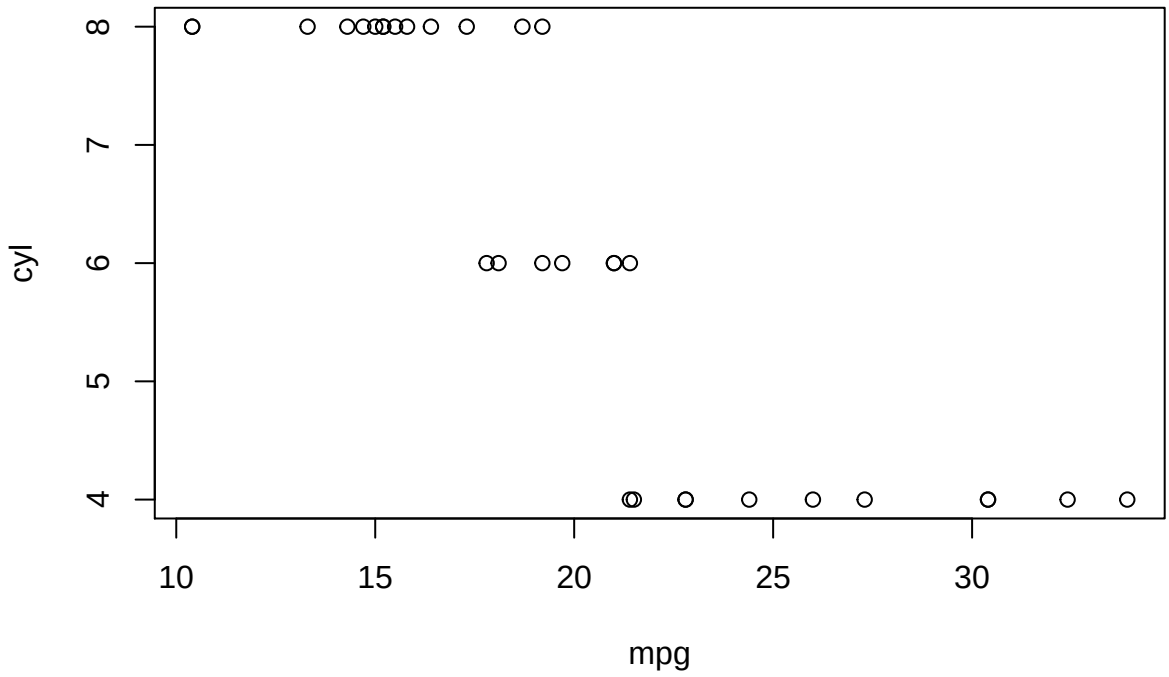
```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
attach(mtcars)
```

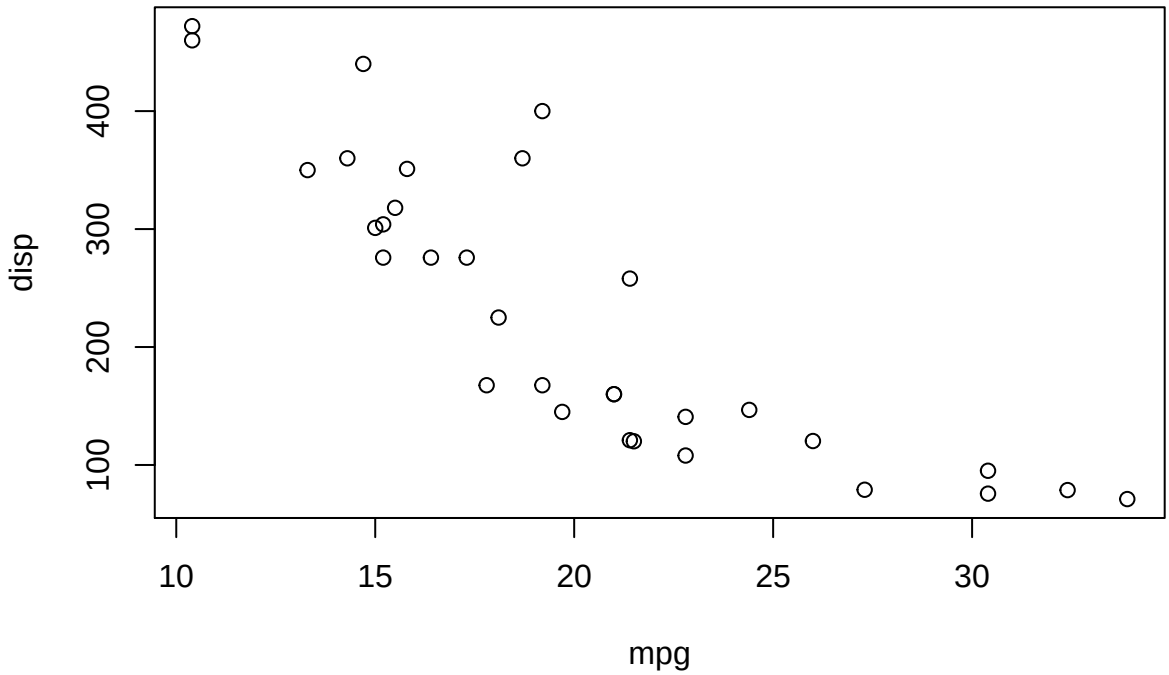
```
## The following objects are masked from mtcars (pos = 3):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt

## The following objects are masked from mtcars (pos = 15):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt

## The following objects are masked from mtcars (pos = 19):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt

## The following object is masked from package:ggplot2:
##
##     mpg
```

Pairwise plotting of the variables of the dataset

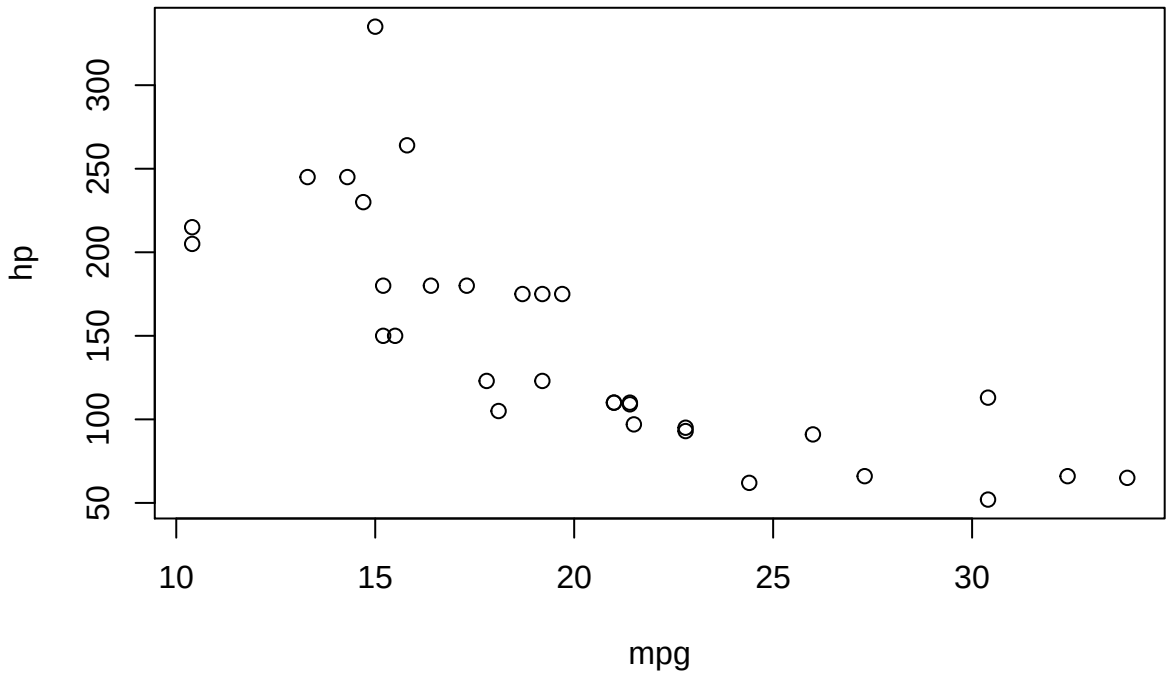
```
plot(mpg, cyl)
```



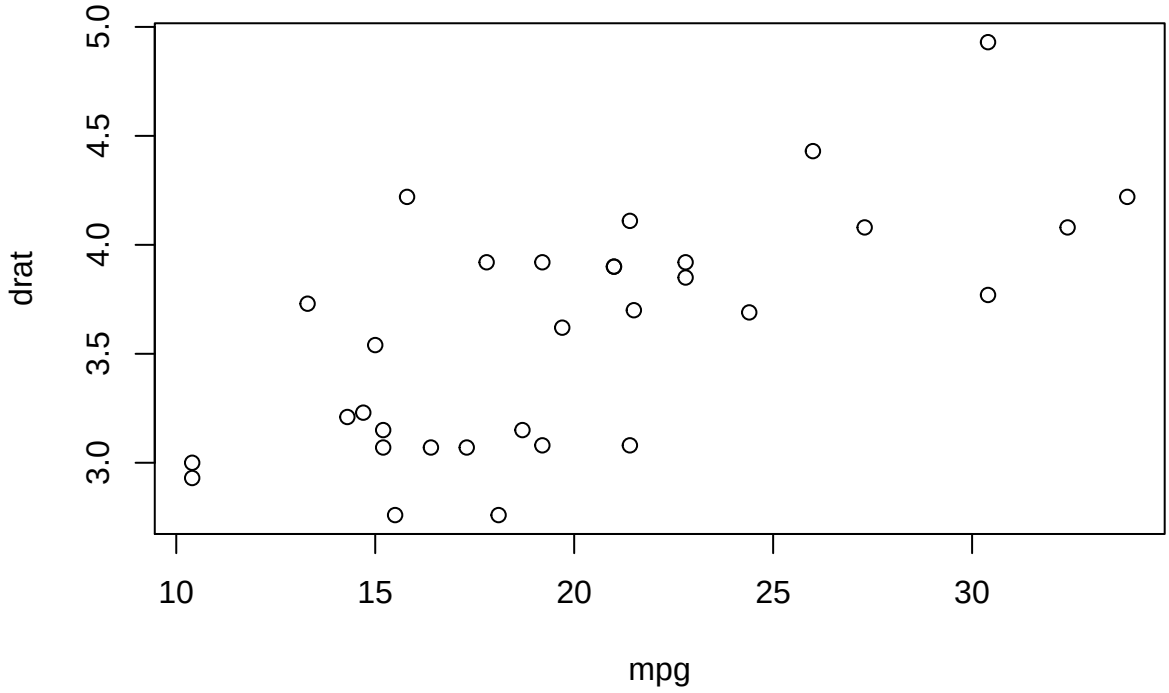

```
plot(mpg, disp)
```



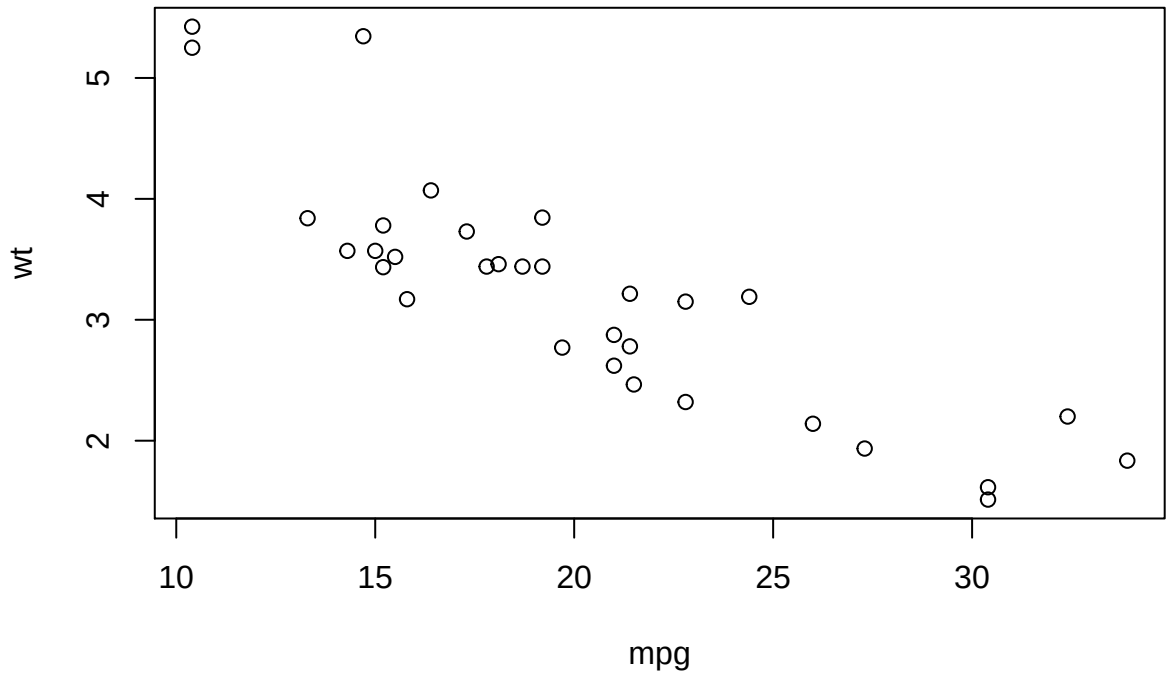

```
plot(mpg, hp)
```



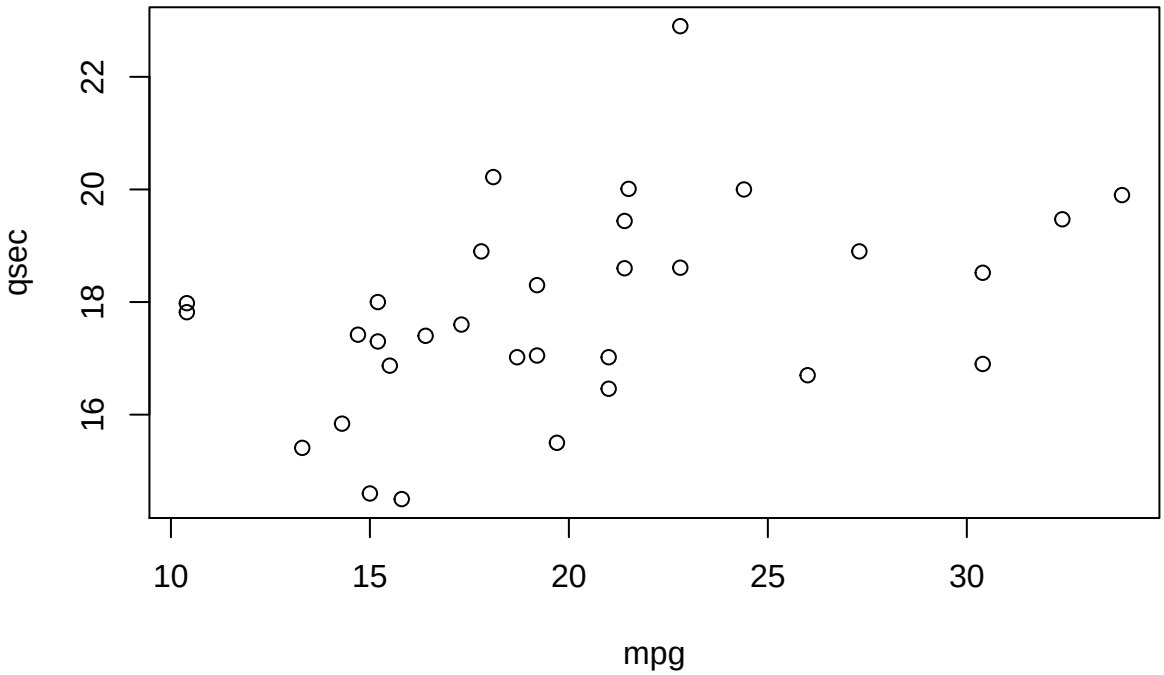

```
plot(mpg, drat)
```



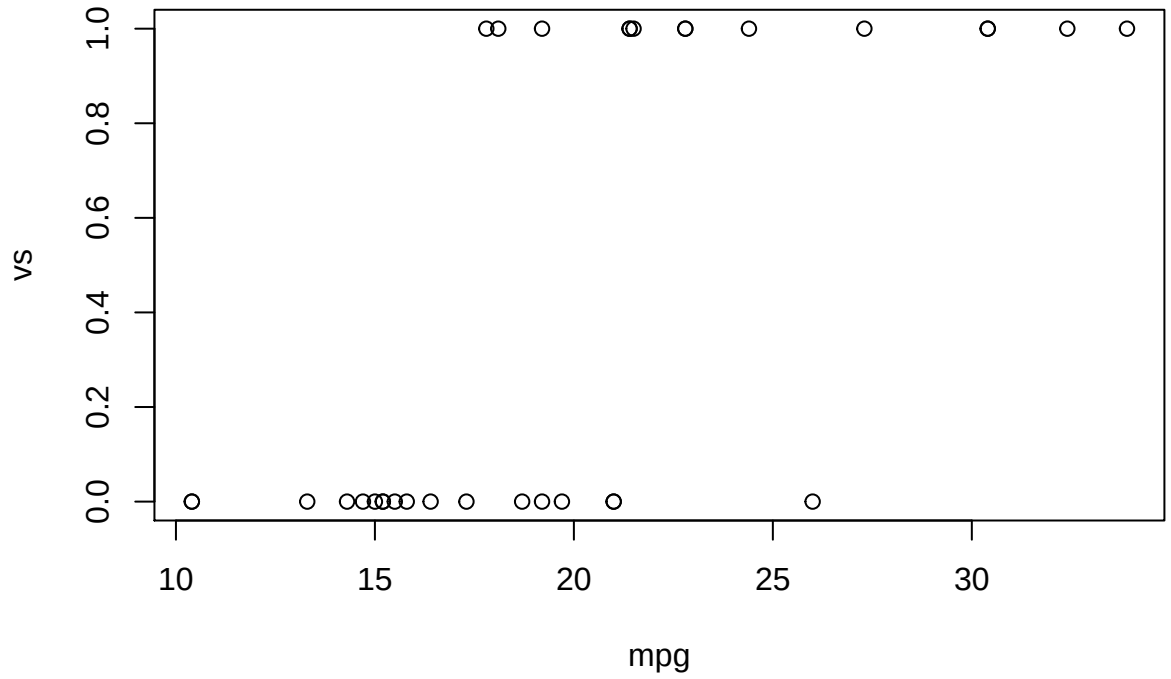

```
plot(mpg, wt)
```



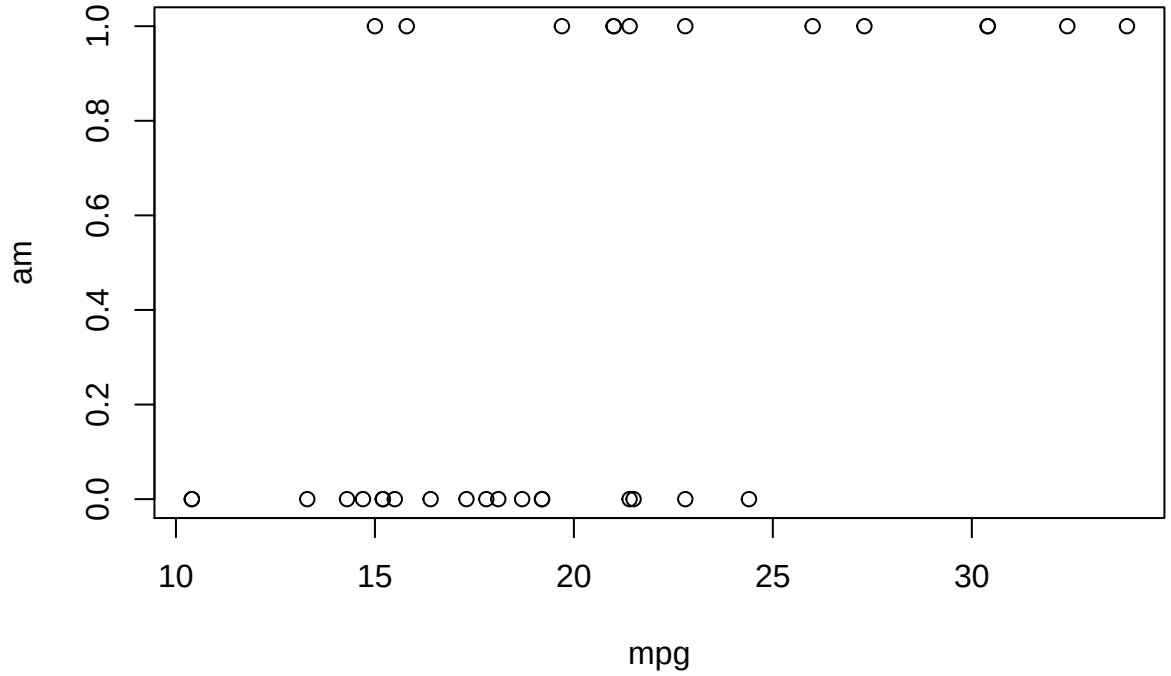

```
plot(mpg, qsec)
```



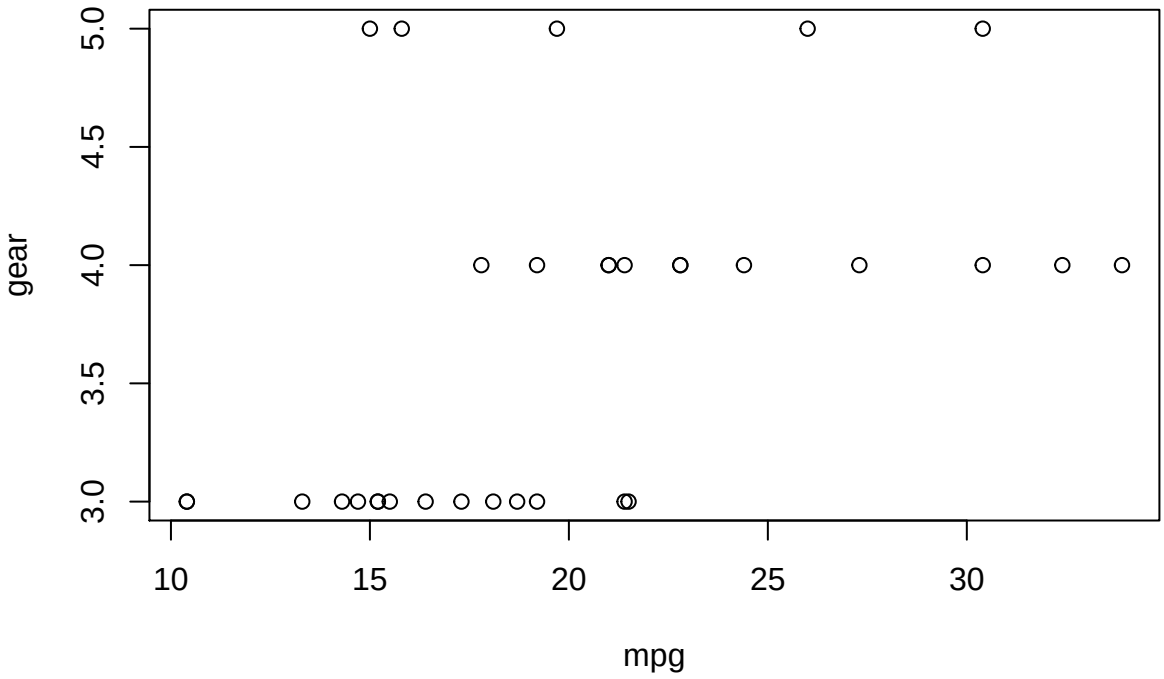

```
plot(mpg, vs)
```



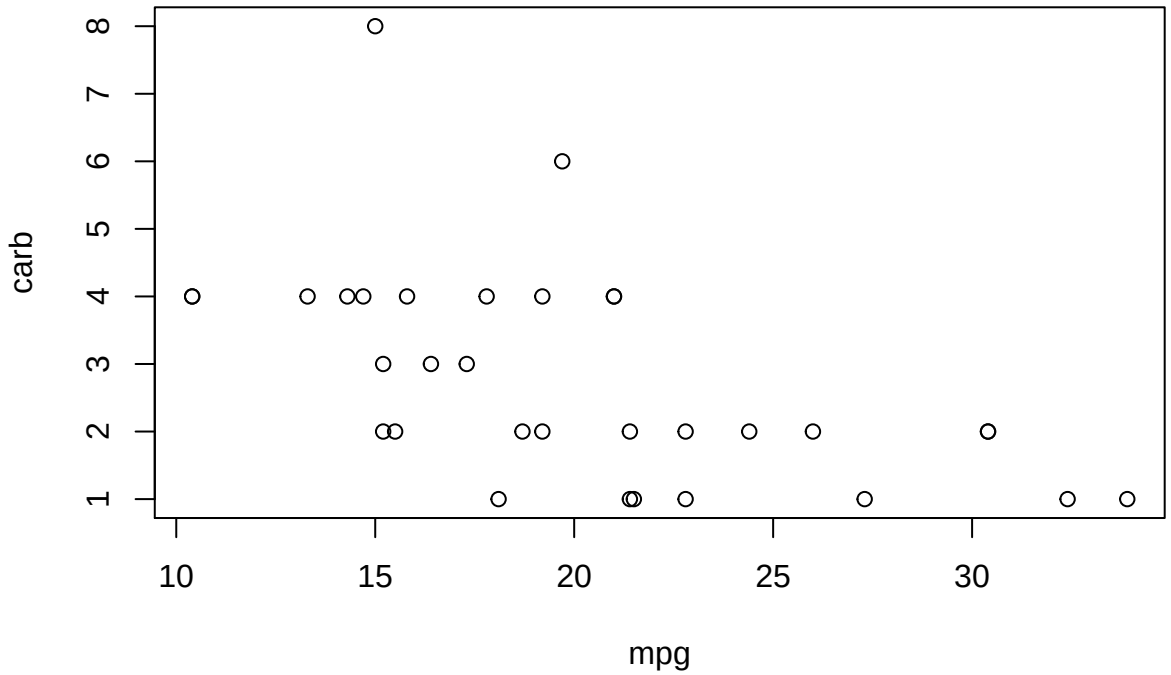

```
plot(mpg, am)
```




```
plot(mpg, gear)
```




```
plot(mpg, carb)
```



Multiple Linear Regression model fittiing

```
model = lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb)
model
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
##     am + gear + carb)
##
```

```
## Coefficients:
## (Intercept)          cyl         disp           hp         drat           wt
##    12.30337     -0.11144      0.01334     -0.02148      0.78711     -3.71530
##        qsec           vs           am         gear         carb
##     0.82104      0.31776      2.52023      0.65541     -0.19942
```

```
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
##     am + gear + carb)
##
## 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
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## 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
```

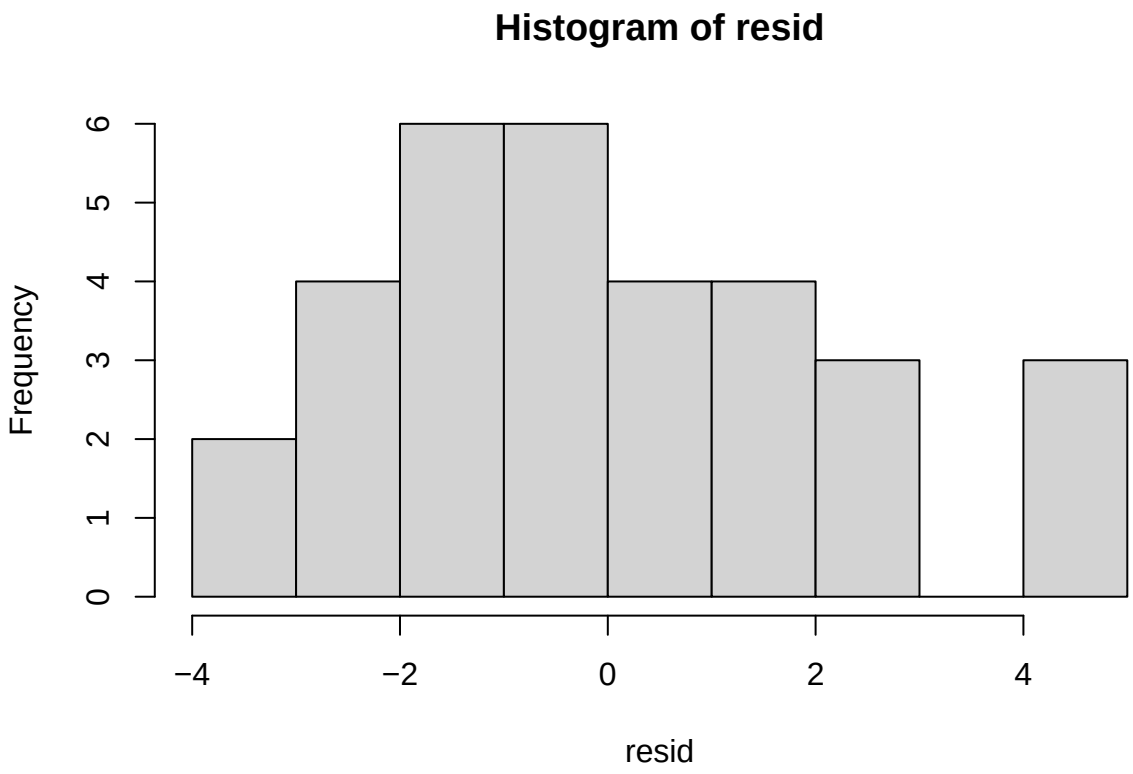
Coefficients confidence intervals

```
confint(model, level=.95)
```

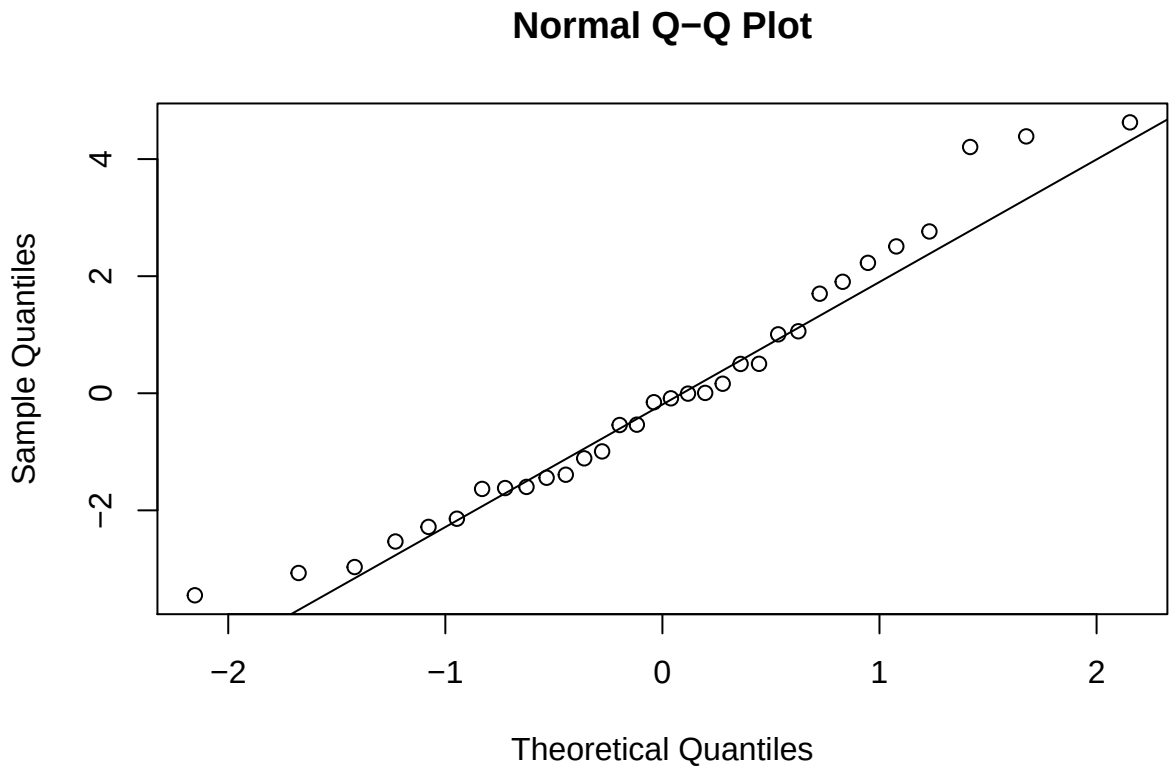
```
##                    2.5 %      97.5 %
## (Intercept) -26.62259745 51.22934576
## cyl          -2.28468553  2.06180457
## disp         -0.02380146  0.05047194
## hp           -0.06675236  0.02378812
## drat         -2.61383350  4.18805545
## wt           -7.65495413  0.22434628
## qsec         -0.69883421  2.34091571
## vs           -4.05880242  4.69432805
## am           -1.75681208  6.79726585
## gear         -2.44999107  3.76081711
## carb         -1.92290442  1.52406591
```

Checking normality of residuals and Residual histogram

```
resid<- model$residuals
hist(resid)
```

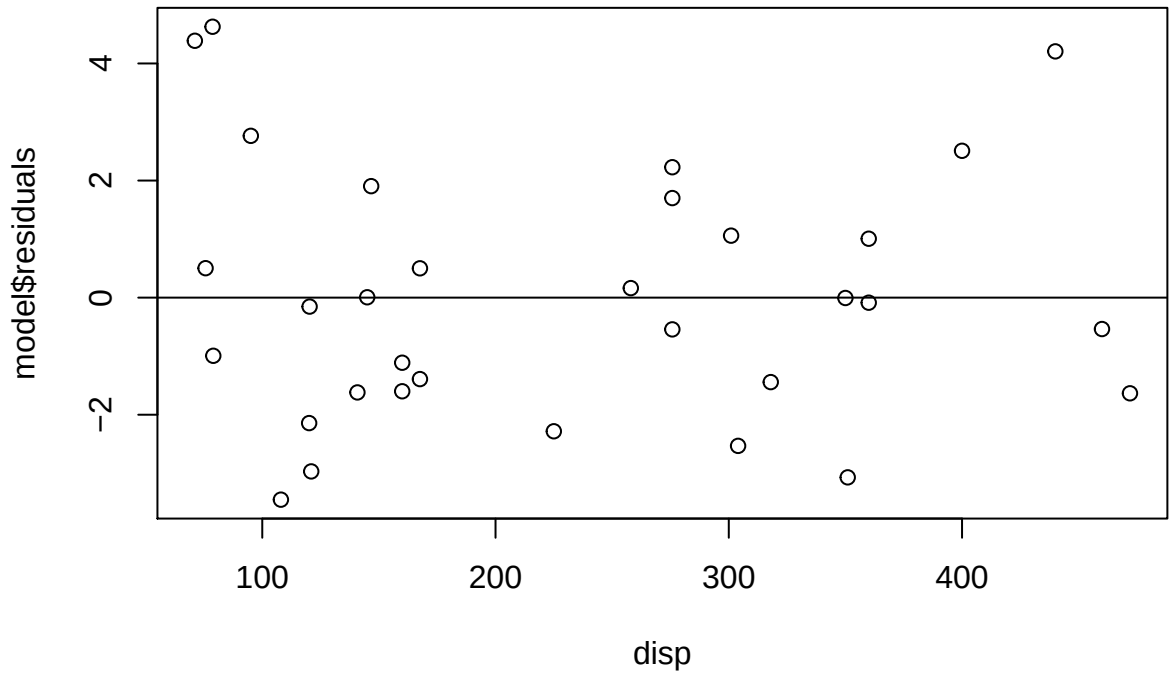



```
qqnorm(resid)
qqline(resid)
```



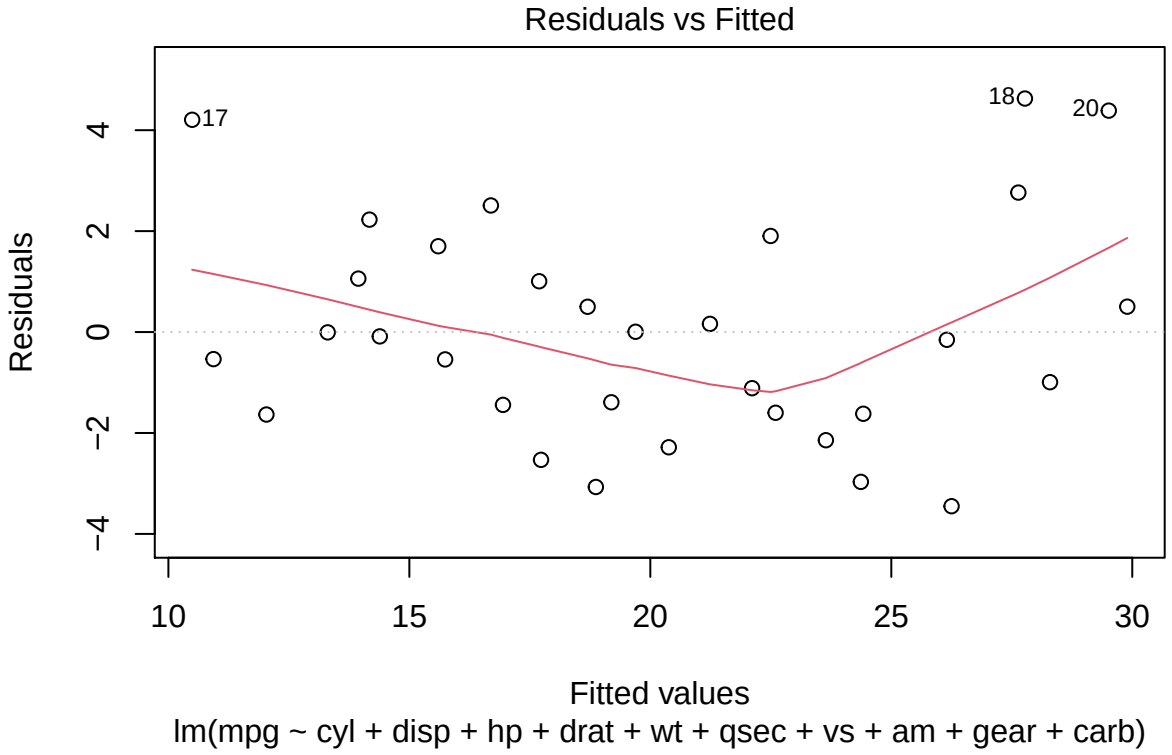
Homoscedasticity

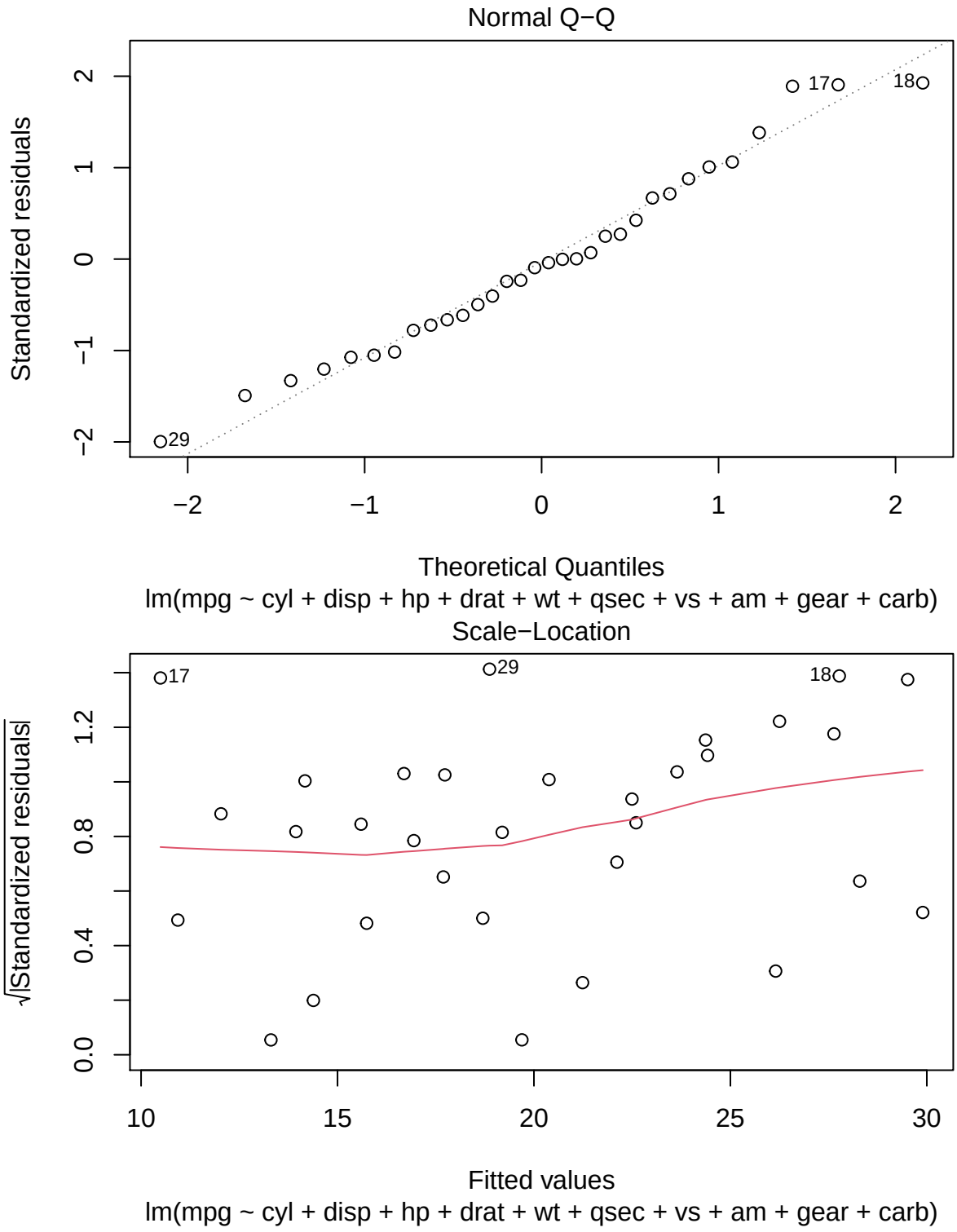
```
plot(model$residuals ~ disp)
abline(0,0)
```

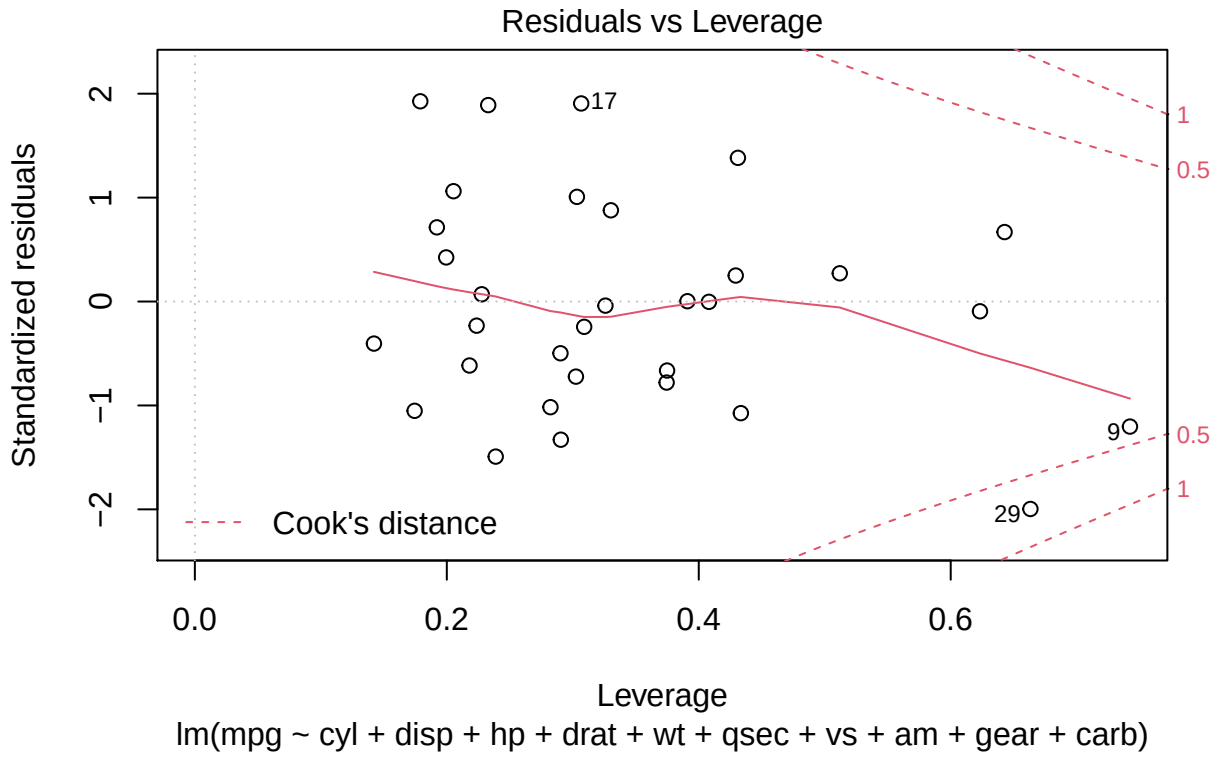


Residual analysis

```
plot(model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb)
Scale−Location
√|Standardized residuals|
Fitted values
lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb)



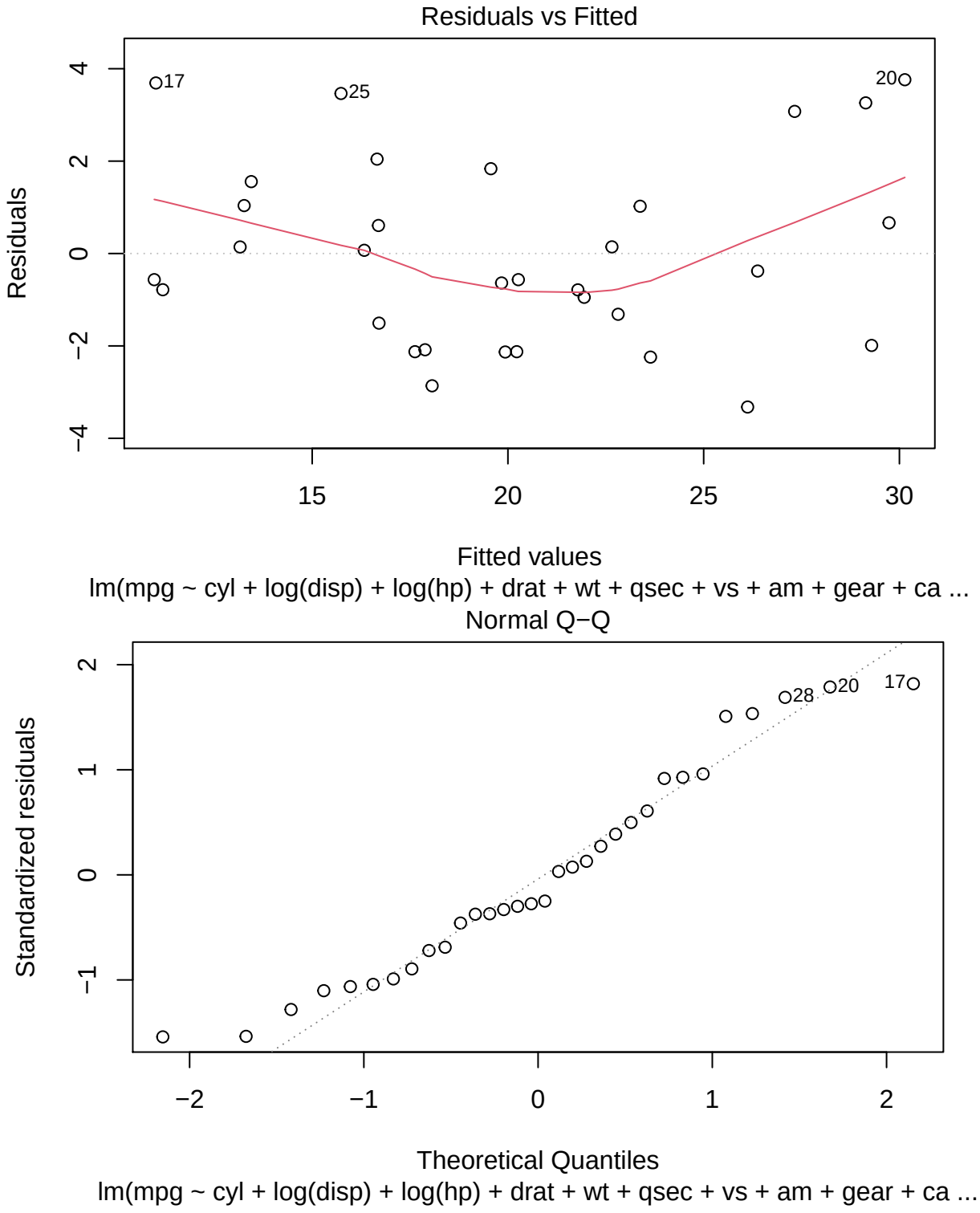
Transformation

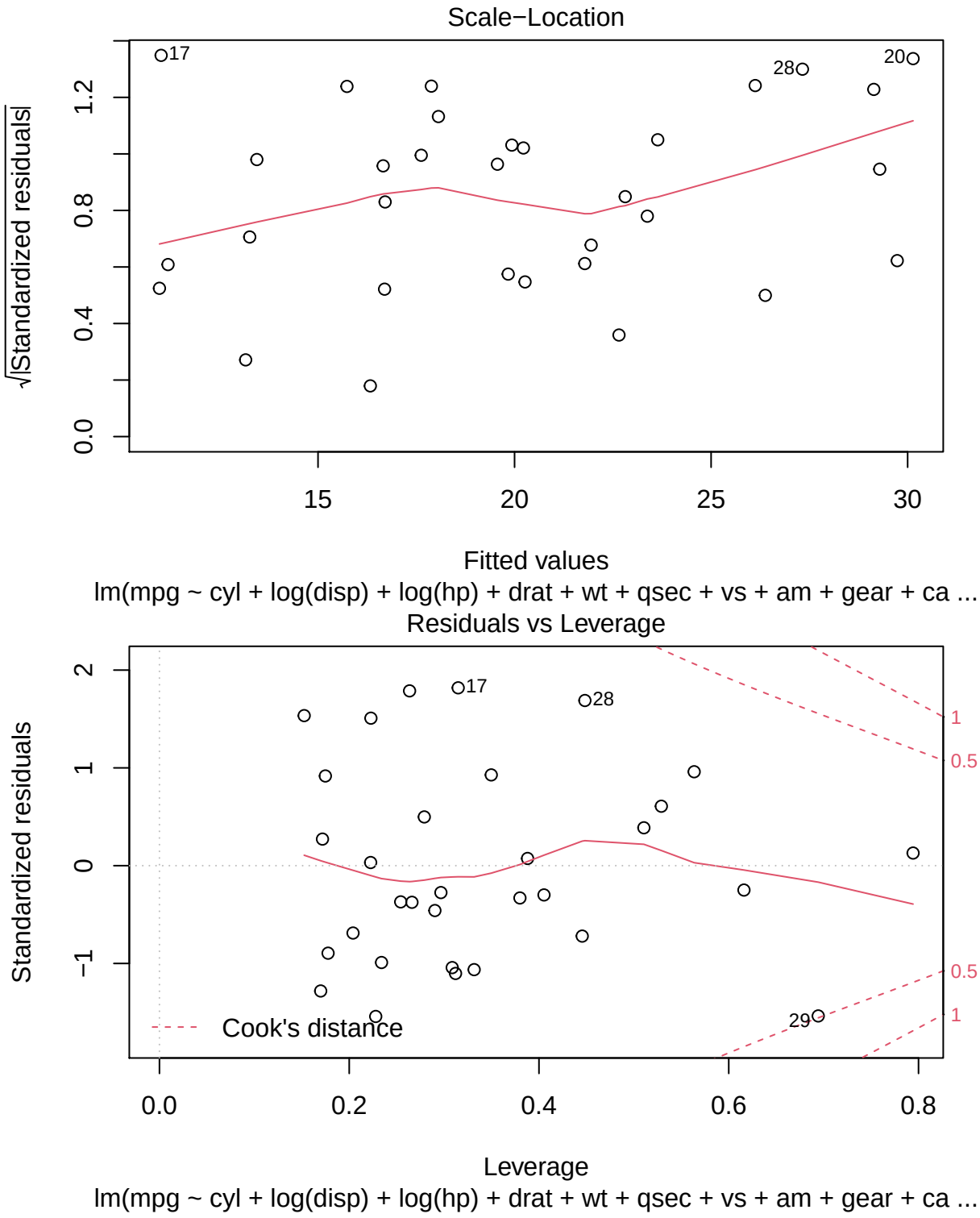
```
model1 <- lm(mpg ~cyl+log(disp)+log(hp)+drat+wt+qsec+vs+am+gear+carb)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + log(disp) + log(hp) + drat + wt + qsec +
##     vs + am + gear + carb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3225 -1.6278 -0.4725  1.1672  3.7616
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.749992  25.678622   2.054   0.0526 .
## cyl          0.934117   1.010657   0.924   0.3658
## log(disp)   -4.923860   3.852996  -1.278   0.2152
## log(hp)     -3.406400   3.003988  -1.134   0.2696
## drat         0.169684   1.549901   0.109   0.9139
## wt          -0.975286   1.506152  -0.648   0.5243
## qsec         0.156231   0.686120   0.228   0.8221
## vs          -0.005731   1.904313  -0.003   0.9976
## am           0.986507   2.084110   0.473   0.6408
## gear         1.706226   1.463070   1.166   0.2566
## carb        -0.973755   0.653397  -1.490   0.1510
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.452 on 21 degrees of freedom
## Multiple R-squared:  0.8879, Adjusted R-squared:  0.8345
## F-statistic: 16.63 on 10 and 21 DF,  p-value: 8.023e-08
```

```
plot(model1)
```





AIC analysis on the original model and AIC analysis on the transition model

```
stepAIC(model)
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
```

```
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
```

```
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
##
## Call:
## lm(formula = mpg ~ wt + qsec + am)
##
## Coefficients:
## (Intercept)           wt         qsec           am
##       9.618       -3.917        1.226        2.936
```

```
stepAIC(model1)
```

```
## Start:  AIC=65.93
## mpg ~ cyl + log(disp) + log(hp) + drat + wt + qsec + vs + am +
```

```
##     gear + carb
##
##             Df Sum of Sq    RSS    AIC
## - vs         1    0.0001 126.28 63.928
## - drat       1    0.0721 126.35 63.947
## - qsec       1    0.3118 126.59 64.007
## - am         1    1.3473 127.63 64.268
## - wt         1    2.5214 128.80 64.561
## - cyl        1    5.1370 131.42 65.204
## - log(hp)    1    7.7323 134.01 65.830
## <none>                   126.28 65.928
## - gear       1    8.1782 134.46 65.936
## - log(disp)  1    9.8204 136.10 66.325
## - carb       1   13.3555 139.63 67.146
##
## Step:  AIC=63.93
## mpg ~ cyl + log(disp) + log(hp) + drat + wt + qsec + am + gear +
##     carb
##
##             Df Sum of Sq    RSS    AIC
## - drat       1    0.0720 126.35 61.947
## - qsec       1    0.3497 126.63 62.017
## - am         1    1.4160 127.70 62.285
## - wt         1    2.5408 128.82 62.566
## - cyl        1    5.4640 131.74 63.284
## - log(hp)    1    7.9966 134.28 63.893
## <none>                   126.28 63.928
## - gear       1    8.2188 134.50 63.946
## - log(disp)  1    9.9454 136.22 64.354
## - carb       1   13.3899 139.67 65.153
##
## Step:  AIC=61.95
## mpg ~ cyl + log(disp) + log(hp) + wt + qsec + am + gear + carb
##
##             Df Sum of Sq    RSS    AIC
## - qsec       1    0.3086 126.66 60.025
## - am         1    1.4720 127.82 60.317
## - wt         1    2.5345 128.89 60.582
## - cyl        1    5.3971 131.75 61.285
## <none>                   126.35 61.947
## - log(hp)    1    8.5533 134.91 62.043
## - gear       1    8.6525 135.00 62.066
## - log(disp)  1   10.1591 136.51 62.421
## - carb       1   13.3729 139.72 63.166
##
## Step:  AIC=60.02
## mpg ~ cyl + log(disp) + log(hp) + wt + am + gear + carb
##
##             Df Sum of Sq    RSS    AIC
## - am         1    1.1724 127.83 58.320
## - wt         1    2.3372 129.00 58.610
## - cyl        1    5.1185 131.78 59.292
## <none>                   126.66 60.025
## - gear       1    8.7232 135.38 60.156
```

```
## - log(hp)    1    9.3330 135.99 60.300
## - log(disp)  1   12.4852 139.15 61.033
## - carb       1   15.9928 142.65 61.830
##
## Step:  AIC=58.32
## mpg ~ cyl + log(disp) + log(hp) + wt + gear + carb
##
##             Df Sum of Sq    RSS    AIC
## - wt         1    2.7307 130.56 56.996
## - cyl        1    6.6592 134.49 57.945
## <none>                   127.83 58.320
## - log(hp)    1    8.7242 136.56 58.432
## - gear       1   15.8483 143.68 60.060
## - log(disp)  1   16.0475 143.88 60.104
## - carb       1   16.7435 144.58 60.258
##
## Step:  AIC=57
## mpg ~ cyl + log(disp) + log(hp) + gear + carb
##
##             Df Sum of Sq    RSS    AIC
## - log(hp)    1     7.249 137.81 56.725
## <none>                   130.56 56.996
## - cyl        1    13.526 144.09 58.150
## - gear       1    32.779 163.34 62.164
## - carb       1    38.572 169.13 63.279
## - log(disp)  1    53.640 184.20 66.010
##
## Step:  AIC=56.73
## mpg ~ cyl + log(disp) + gear + carb
##
##             Df Sum of Sq    RSS    AIC
## - cyl        1     8.707 146.52 56.686
## <none>                   137.81 56.725
## - gear       1    26.481 164.29 60.349
## - carb       1    60.918 198.73 66.439
## - log(disp)  1    97.216 235.03 71.807
##
## Step:  AIC=56.69
## mpg ~ log(disp) + gear + carb
##
##             Df Sum of Sq    RSS    AIC
## <none>                   146.52 56.686
## - gear       1    20.189 166.71 58.817
## - carb       1    52.570 199.09 64.497
## - log(disp)  1   152.562 299.08 77.519
##
## Call:
## lm(formula = mpg ~ log(disp) + gear + carb)
##
## Coefficients:
## (Intercept)    log(disp)         gear         carb
##      51.789       -6.592        1.787       -1.227
```

Run a model for each using the recommended variables.

```r
model2<-lm(mpg~qsec+wt+am)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ qsec + wt + am)
##
## 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
## qsec          1.2259     0.2887   4.247 0.000216 ***
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## am            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
```

```r
model3<-lm(mpg~log(disp)+gear+carb)
summary(model3)
```

```
##
## Call:
## lm(formula = mpg ~ log(disp) + gear + carb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0461 -1.3931 -0.5111  1.8053  4.2983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.7887     8.5069   6.088 1.45e-06 ***
## log(disp)    -6.5917     1.2208  -5.399 9.31e-06 ***
## gear          1.7869     0.9097   1.964  0.05950 .
## carb         -1.2271     0.3872  -3.170  0.00368 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.288 on 28 degrees of freedom
## Multiple R-squared:  0.8699, Adjusted R-squared:  0.8559
## F-statistic:  62.4 on 3 and 28 DF,  p-value: 1.62e-12
```

Multi-colinearity check

plot(mtcars) may not be clear from this plot so check pairwise correlations

```r
cor(qsec, wt)
```

```
## [1] -0.1747159
```

```
cor(am, wt)
```

```
## [1] -0.6924953
```

Session 8: Logistic Regression in R

Dr. Dhaval Maheta, VNSGU Surat

What decision should the manufacturer choose to make an automatic or manual car?

Set the dataset mtcars

Use the null/base model

```
mtcars
```

```
##                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4          21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag      21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710         22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive     21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant            18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360         14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D          24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230           22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280           19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C          17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE         16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL         17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC        15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4  8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128           32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic        30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla     33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona      21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger   15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin        15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28         13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird   19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9          27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2      26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa       30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L     15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino       19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora      15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E         21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

```
attach(mtcars)
```

```
## The following objects are masked from mtcars (pos = 3):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
## The following objects are masked from mtcars (pos = 4):
##
```

```
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## The following objects are masked from mtcars (pos = 16):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## The following objects are masked from mtcars (pos = 20):
##
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
base = glm(am~1,data = mtcars,family = binomial)
base
```

```
##
## Call:  glm(formula = am ~ 1, family = binomial, data = mtcars)
##
## Coefficients:
## (Intercept)
##     -0.3795
##
## Degrees of Freedom: 31 Total (i.e. Null);  31 Residual
## Null Deviance:       43.23
## Residual Deviance: 43.23     AIC: 45.23
```

```
summary(base)
```

```
##
## Call:
## glm(formula = am ~ 1, family = binomial, data = mtcars)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.021  -1.021  -1.021   1.342   1.342
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3795     0.3599  -1.054    0.292
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.23  on 31  degrees of freedom
## Residual deviance: 43.23  on 31  degrees of freedom
## AIC: 45.23
##
## Number of Fisher Scoring iterations: 4
```

The result of line "base" gives the intercept -0.379 which is the log of car being manual.

NOTE: Which ever category is coded as 1 is considered to be as the reference category.

Null deviance = 43.23

Residual deviance = 43.23

Let us check a model with some random varibles

```
fit01 = glm(am~mpg+disp+hp+wt, family = binomial)
fit01
```

```
##
## Call:  glm(formula = am ~ mpg + disp + hp + wt, family = binomial)
##
## Coefficients:
## (Intercept)          mpg         disp           hp           wt
##   -18.48207      1.13503     -0.02588      0.10871     -4.80560
##
## Degrees of Freedom: 31 Total (i.e. Null);  27 Residual
## Null Deviance:       43.23
## Residual Deviance: 8.162     AIC: 18.16
```

```
summary(fit01)
```

```
##
## Call:
## glm(formula = am ~ mpg + disp + hp + wt, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.84992  -0.15966  -0.00615   0.01257   1.46081
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -18.48207   40.90451  -0.452    0.651
## mpg           1.13503    1.55720   0.729    0.466
## disp         -0.02588    0.04087  -0.633    0.527
## hp            0.10871    0.09837   1.105    0.269
## wt           -4.80560    3.97978  -1.208    0.227
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance:  8.162  on 27  degrees of freedom
## AIC: 18.162
##
## Number of Fisher Scoring iterations: 9
```

Use ANOVA and check for deviance if deviation is very high then they tend to have higher explanatory power. Using the variables in the next model and check for significance of varibles using pvalue. They are contributing in classification of 'am' our dependent Y variable.

With each increase in

```
anova(fit01)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: am
##
## Terms added sequentially (first to last)
##
```

```
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                    31     43.230
## mpg   1  13.5546        30     29.675
## disp  1   1.0693        29     28.606
## hp    1  18.4577        28     10.148
## wt    1   1.9862        27      8.162
```

```
fit02 = glm(am~mpg+hp+wt,family = binomial)
summary(fit02)
```

```
##
## Call:
## glm(formula = am ~ mpg + hp + wt, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.93381  -0.09191  -0.00913   0.01139   1.47331
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.72137   40.00281  -0.393   0.6943
## mpg           1.22930    1.58109   0.778   0.4369
## hp            0.08389    0.08228   1.020   0.3079
## wt           -6.95492    3.35297  -2.074   0.0381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.2297  on 31  degrees of freedom
## Residual deviance:  8.7661  on 28  degrees of freedom
## AIC: 16.766
##
## Number of Fisher Scoring iterations: 10
```

Now lets drop mpg and make the third model

```
fit03 = glm(am~hp+wt, family = binomial)
summary(fit03)
```

```
##
## Call:
## glm(formula = am ~ hp + wt, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2537  -0.1568  -0.0168   0.1543   1.3449
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 18.86630    7.44356   2.535  0.01126 *
## hp           0.03626    0.01773   2.044  0.04091 *
## wt          -8.08348    3.06868  -2.634  0.00843 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance: 10.059  on 29  degrees of freedom
## AIC: 16.059
##
## Number of Fisher Scoring iterations: 8
```

(exp(coeff hp) - 1) and then multiply by hundred

= (exp(0.036) - 1)*100

= (1.0366 - 1)*100

= 3.6%

Interpretation: With every increase in hp by unit, 3.6% increase in chances of car being manual

Now practice the same thing in "Rcmdr"

Run library(Rcmdr)