Ordinary Least Squares (OLS)

Consider a linear regression of the form $y = X\beta + e$. The OLS estimates of the vector of regression coefficients is given by

$$\hat{\beta} = \text{argmin } RSS(y, X, \beta)$$

Where $RSS(y, X, \beta) = (y - X\beta)'(y - X\beta)$ is the residual sum of squares.

The solution of the above problem can be obtained from the following systems of equatrions

$$(X'X)\beta = X'y$$

Estimation (full-rank case)

Computation of OLS estimates using `lm`, `lsfit` and with matrix operations.

```
   fm=lm(y~X) ;coef(fm) ; summary(fm)

   fm=lsfit(y=y,x=X); coef(fm)    # a bit faster than lm

   XtX=crossprod(X)     # matrix of coefficients, X'X
   Xty=crossprod(X,y)   # right-hand side X'y
   bHat=solve(XtX,Xty)  # solution
```

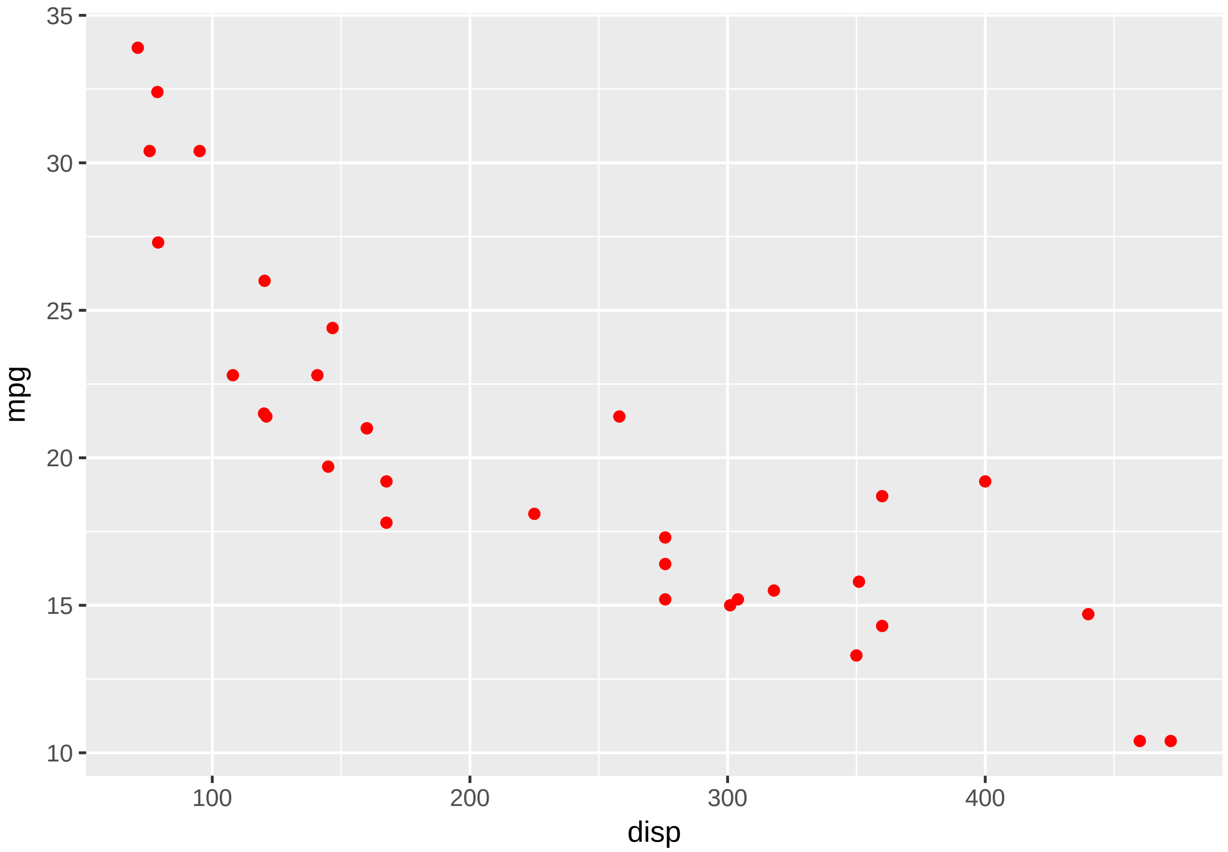
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

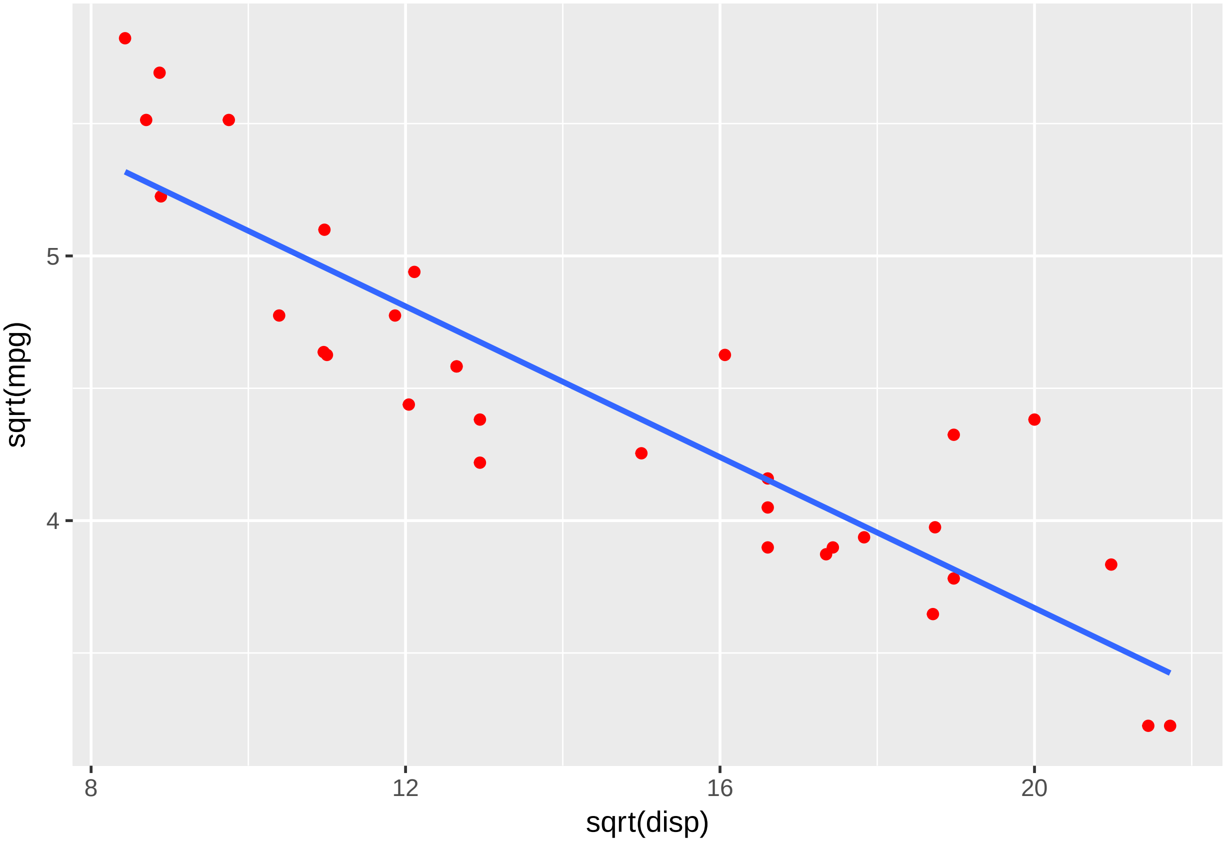
```
library(ggplot2)
mtcars %>%
 ggplot(aes(x = disp, y = mpg)) +
 geom_point(colour = "red")
```



Transformation

If the relationship is non-linear, a common approach in linear regression modelling is to transform the response and predictor variable in order to coerce the relationship to one that is more linear.

```
mtcars %>%
  ggplot(aes(x = sqrt(disp), y = sqrt(mpg))) +
  geom_point(colour = "red") +
  geom_smooth(method = "lm", fill = NA)
```



```r
 lmodel <- lm(sqrt(mpg) ~ sqrt(disp), data = mtcars)

lmodel$coefficients
```

```
## (Intercept)  sqrt(disp) 
##   6.5192052  -0.1424601
```

```r
summary(lmodel)
```

```
## 
## Call:
## lm(formula = sqrt(mpg) ~ sqrt(disp), data = mtcars)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.45591 -0.21505 -0.07875  0.16790  0.71178 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  6.51921    0.19921   32.73  < 2e-16 ***
## sqrt(disp)  -0.14246    0.01312  -10.86 6.44e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3026 on 30 degrees of freedom
## Multiple R-squared:  0.7973, Adjusted R-squared:  0.7905 
## F-statistic:   118 on 1 and 30 DF,  p-value: 6.443e-12
```

The p-value of 6.443e-12 indicates a statistically significant relationship at the p<0.001 cut-off level.

The multiple R-squared value (R-squared) of 0.7973 gives the variance explained and can be used as a measure of predictive power (in the absence of overfitting).

The RMSE is included in the output (Residual standard error) where it has a value of 0.3026.

Interpretation: for every unit increase in the square root of engine displacement there is a -0.14246 decrease in the square root of fuel efficiency (mpg). Therefore, fuel efficiency decreases with increasing engine displacement.