Sometimes we are interested in obtaining a simple model that explains the relationship between two or more variables.

For example, suppose that we are interested in studying the relationship between the income of parents and the income of their children in a certain country: we suspect that children from wealthier families generally tend to become wealthier when they grow up. Here, we can consider two variables:

- The family income can be defined as the average income of parents per year
- The child income can be defined as his/her average income per year

To examine the relationship between the two variables, we collect some data

$$(x_i, y_i), \qquad \text{for } i = 1, 2, \cdots, n, \tag{1}$$

where $y_i$ is the average income of a child and $x_i$ is the average income of his/her parents.

## Simple linear regression

- **Predictor**, **explanatory**, **independent** variable $X$
- **Response**, **outcome**, **dependent** variable $Y$

| $i$ | $Y$ | $X$ |
|---|---|---|
| 1 | $y_1$ | $x_1$ |
| 2 | $y_2$ | $x_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $y_n$ | $x_n$ |

*Multiple linear regression* concers two or more predictor variables

We are not considering **deterministic** (or **functional**) **relationships**, like

$$\mathsf{F} = \frac{9}{5}\mathsf{C} + 32$$

We are interested in **statistical relationships** in which the relationship is not perfect.

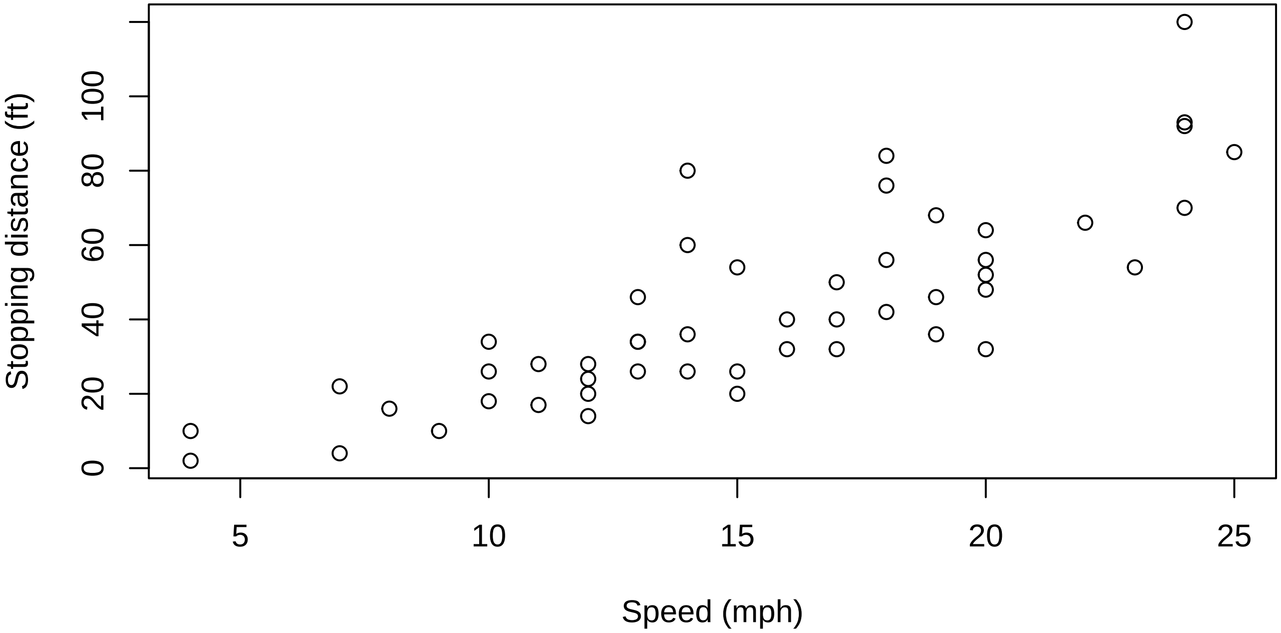
`cars` dataset in R contains 50 observations on 2 variables:

`speed`: Speed (mph)

`dist`: Stopping distance (ft)

[Mordecai Ezekiel, Methods of Correlation Analysis, Wiley (1930)]

```
plot(cars, xlab = "Speed (mph)", ylab = "Stopping distance (ft)")
```



The **Pearson correlation coefficient** is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between $-1$ and 1.

The formula is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

In R, there is a function computing this:

```
cor(df$speed, df$dist, method = "pearson")
[1] 0.8068949
```

The idea is in Exercise 2, you do the computations manually to get the idea.

## The simple linear model

The relationship between a response variable $Y$ and a predictor variable $X$ is postulated as a **linear model**

$$Y = a + bX + \varepsilon,$$

where $a$ and $b$ are constants called **intercept** and **slope**.

The linear regression model contains an **error term** $\varepsilon$ accounting the variability in $Y$ that cannot be explained by the linear relationship between $X$ and $Y$.

Each observation can be written as

$$y_i = a + by_i + \varepsilon_i$$

the quantities $\varepsilon_i$ are random variables representing errors in the relationship.

The errors $\{\varepsilon_i\}$ are independent and zero-mean normal random variables, that is,

$$\varepsilon_i \sim N(0, \sigma^2). \tag{2}$$

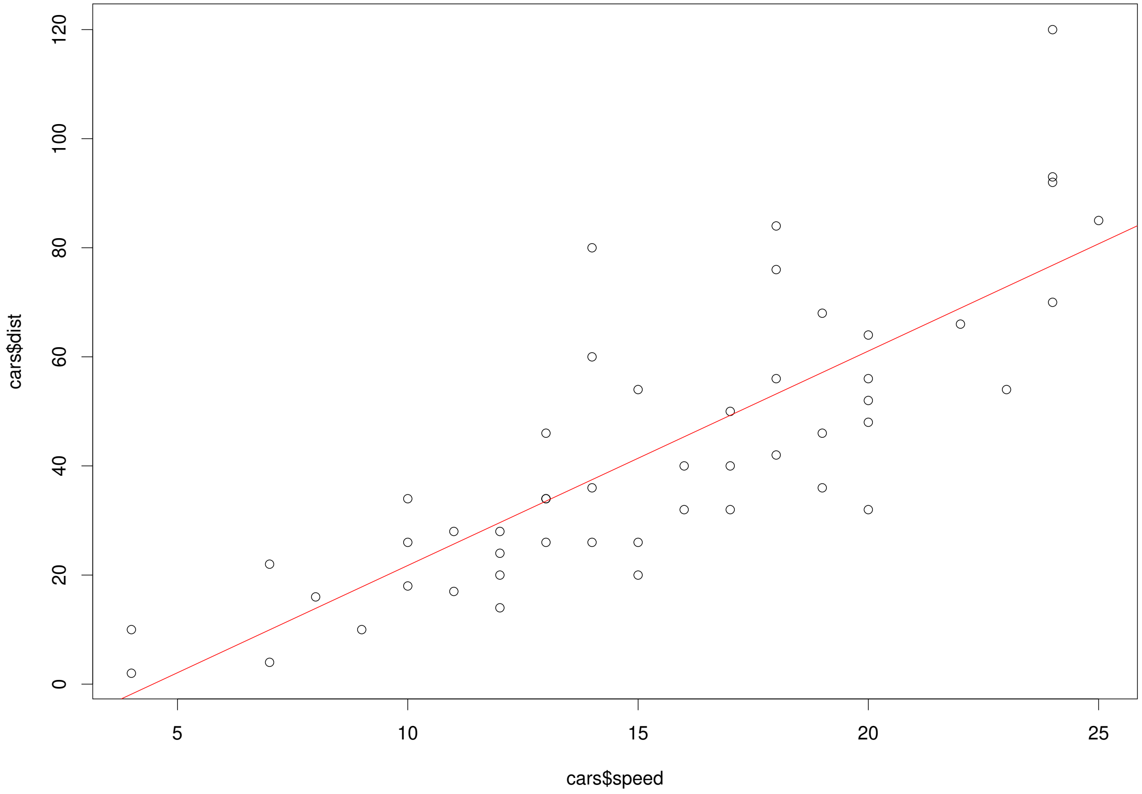
A linear model

$$\text{dist} = a + b \cdot \text{speed} + \varepsilon$$

represents relationship between speed and stopping distance.

In R, the `lm()` (*linear model*) function can be used to create a simple regression model.

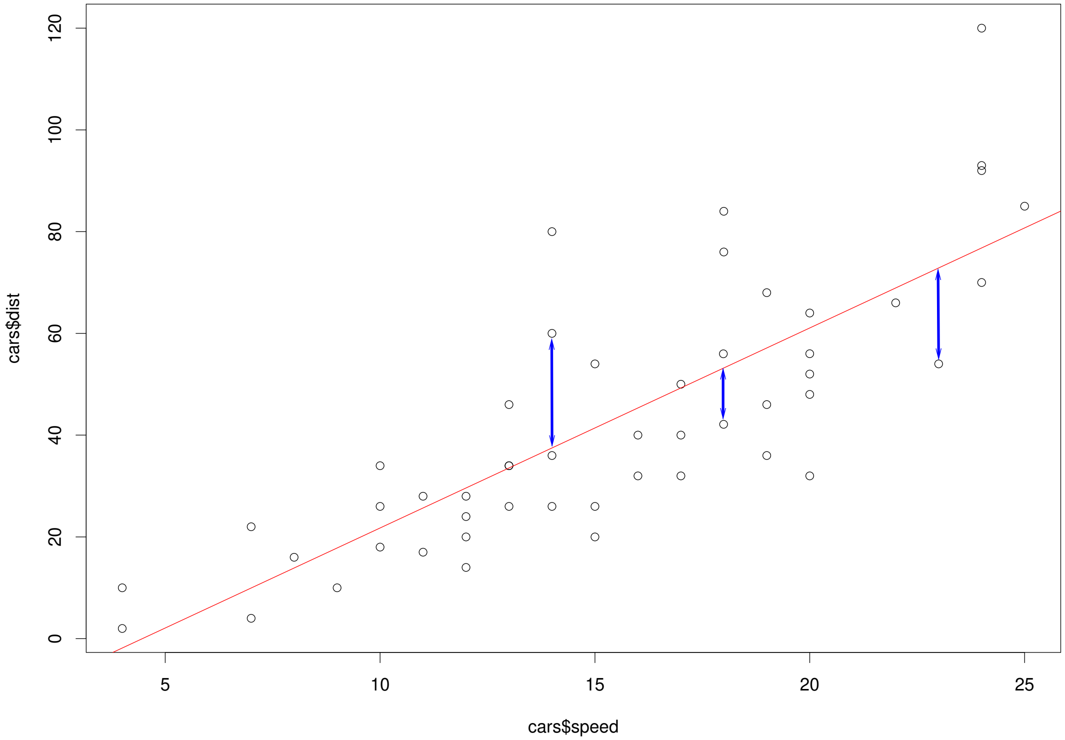
```
model1 <- lm(dist ∼ speed, data=cars)  # builds the model
abline(model1, col="red")     # plots a red straight line
```



## Parameter estimation

Based on the available data, we estimate the parameters $a$ and $b$ to find the straight line that gives the *best fit*.

We estimate the parameters using the **least squares method** that minimizes the sum of squares of *vertical distances* from each point to the line.



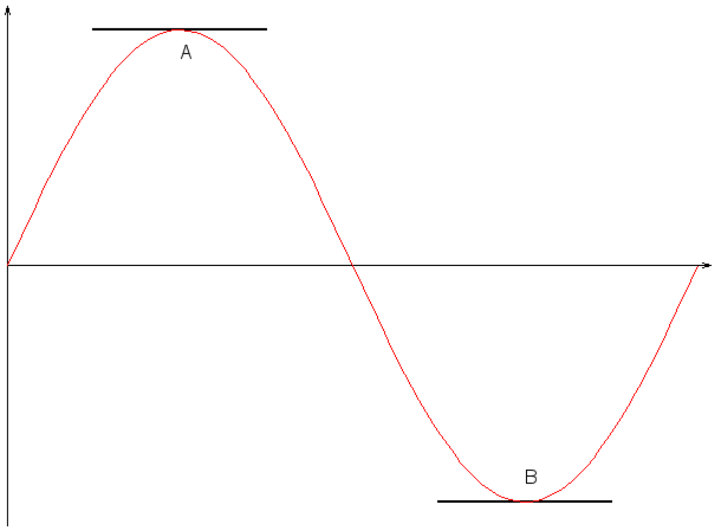
The sum of squares of these distances is

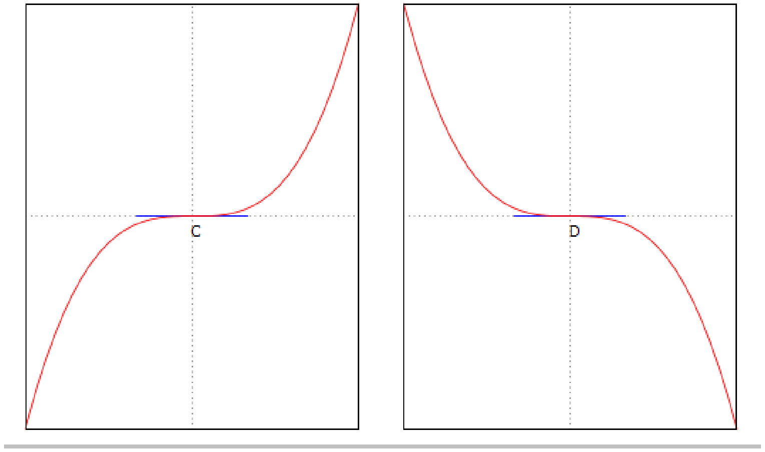
$$S(a,b) = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

To minimize $S(a,b)$, we need to find **stationary points**:

$$\frac{\partial}{\partial a} S(a,b) = 0 \qquad \text{and} \qquad \frac{\partial}{\partial b} S(a,b) = 0$$

A **stationary point** of a function is a point where the derivative is equal to 0. These points are called “stationary” because at these points the function is neither increasing nor decreasing.





There are two types of **turning points**:

- **local maximum**, the largest value in the local region
- **local minimum**, the smallest value in the local region

All turning points are stationary points, but not all stationary points are turning points.

From the first equation, we get

$$\begin{aligned}\frac{\partial}{\partial a} S(a,b) &= \frac{\partial}{\partial a} \sum_{i=1}^{n} (a^2 + 2abx_i - 2ay_i + b^2 {x_i}^2 - 2bx_i y_i + {y_i}^2) \\ &= 2 \sum_{i=1}^{n} (a + bx_i - y_i)\end{aligned}$$

Setting this to zero we get:

$$na = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} bx_i$$

Let us denote the mean by

$$\bar{y} = \left( \sum_{i=1}^{n} y_i \right) \Big/ n$$

We obtain the solution

$$a = \bar{y} - b\bar{x}$$

From the other equation we get

$$\begin{aligned}\frac{\partial}{\partial b}S(a,b) &= \frac{\partial}{\partial b}\sum_{i=1}^{n}(a^2+2abx_i-2ay_i+b^2{x_i}^2-2bx_iy_i+{y_i}^2)\\ &= -2\sum_{i=1}^{n}(x_iy_i-ax_i-b{x_i}^2)\end{aligned}$$

Setting this to zero we get:

$$0=\sum_{i=1}^{n}(x_iy_i-ax_i-b{x_i}^2)$$

Because $a=\bar{y}-b\bar{x}$, we have

$$\begin{aligned}0 &= \sum(x_iy_i-(\bar{y}-b\bar{x})x_i-b{x_i}^2)=\sum(x_iy_i-\bar{y}x_i+b\bar{x}x_i-b{x_i}^2)\\ &= \sum(x_iy_i-\bar{y}x_i)-b\sum({x_i}^2-\bar{x}x_i)\end{aligned}$$

We obtain

$$b=\frac{\sum(x_iy_i-\bar{y}x_i)}{\sum({x_i}^2-\bar{x}x_i)}$$

## Correctness

A **partial derivative** of a function of several variables is its derivative with respect to one of those variables.

The partial derivative of a function $f(x, y, \ldots)$ with respect to the variable $x$ is denoted by

$$\frac{\partial f(x, y, \ldots)}{\partial x}$$

**Example.** Let $f(x, y) = y^3 x^2$. Calculate $\frac{\partial f(x,y)}{\partial x}$

We simply view $y$ as being a fixed number and calculate the ordinary derivative with respect to $x$. Therefore, the term $y^3$ is viewed as a constant $a$. The derivative of $ax^2$ (with respect to $x$) is $2ax$. This means that the solution is $2y^3x$.

**Second order** partial derivatives are denoted by

$$\frac{\partial^2 f}{\partial x^2}, \quad \frac{\partial^2 f}{\partial y \partial x}, \quad \frac{\partial^2 f}{\partial x \partial y}, \quad \frac{\partial^2 f}{\partial y^2}$$

Note we will derivate with respect to the "inner" first. It is also known that if the partial derivatives are *continuous*, the order of differentiation can be interchanged (Clairaut's theorem) so the Hessian matrix will be symmetric.

Let $f(x, y)$ be a function. The **Hessian matrix** $H(x, y)$ is a square matrix of second-order partial derivatives:

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

## Second partial derivative test

$$\frac{\partial^2 S(a,b)}{\partial a^2} = \sum_{i=1}^{n} 2 = 2n, \qquad \frac{\partial^2 S(a,b)}{\partial b^2} = 2\sum_{i=1}^{n} {x_i}^2$$

$$\frac{\partial^2 S(a,b)}{\partial a \partial b} = \frac{\partial^2 S(a,b)}{\partial b \partial a} = 2\sum_{i=1}^{n} x_i = 2n\bar{x}$$

The **Hessian matrix**

$$H(a,b) = 2\begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum {x_i}^2 \end{pmatrix}$$

Its determinant is

$$\det H(a,b) = 4\left(n\sum_{i=1}^{n} {x_i}^2 - n^2\bar{x}^2\right) = 4n\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq 0$$

## Second partial derivative test

Suppose that $(\alpha, \beta)$ is a stationary point of $S(a, b)$. Then the second partial derivative test asserts that if

$$\det H(\alpha, \beta) > 0 \qquad \text{and} \qquad \frac{\partial^2 S(\alpha, \beta)}{\partial a^2} > 0,$$

then $S$ has a minimum at $(\alpha, \beta)$

Because $\det H(a, b) > 0$ and $2n > 0$ for any $(a, b)$, function $S(a, b)$ has a minimum at

$$a = \bar{y} - b\bar{x} \qquad \text{and} \qquad b = \frac{\sum(x_i y_i - \bar{y} x_i)}{\sum({x_i}^2 - \bar{x} x_i)}$$

## Using our formula:

```
> x_bar <- mean(cars$speed)
> y_bar <- mean(cars$dist)

> b <- sum(cars$speed * cars$dist - y_bar * cars$speed) /
       sum(cars$speed^2 - x_bar*cars$speed)
> a <- y_bar - b * x_bar

> a;b
[1] -17.5790948905
[1] 3.93240875912
```

## Using `lm()`:

```
> model1 <- lm(dist ∼ speed, data=cars)
> coef(model1)
(Intercept)           speed
-17.57909489051   3.93240875912
```

We have the model: $\texttt{dist} = 3.93240875912 * \texttt{speed} - 17.57909489051$

We can add predicted values to the frame

```
cars$dist_hat <- predict(model1)
```

In general '$y$ hat' (written $\widehat{y}$) is the predicted value of $y$ (the dependent variable) in a regression equation.

The value $\varepsilon_i = y_i - (a + bx_i)$ is the difference between the observed value and the value predicted by model. It is called commonly as **residual** (or **error**).

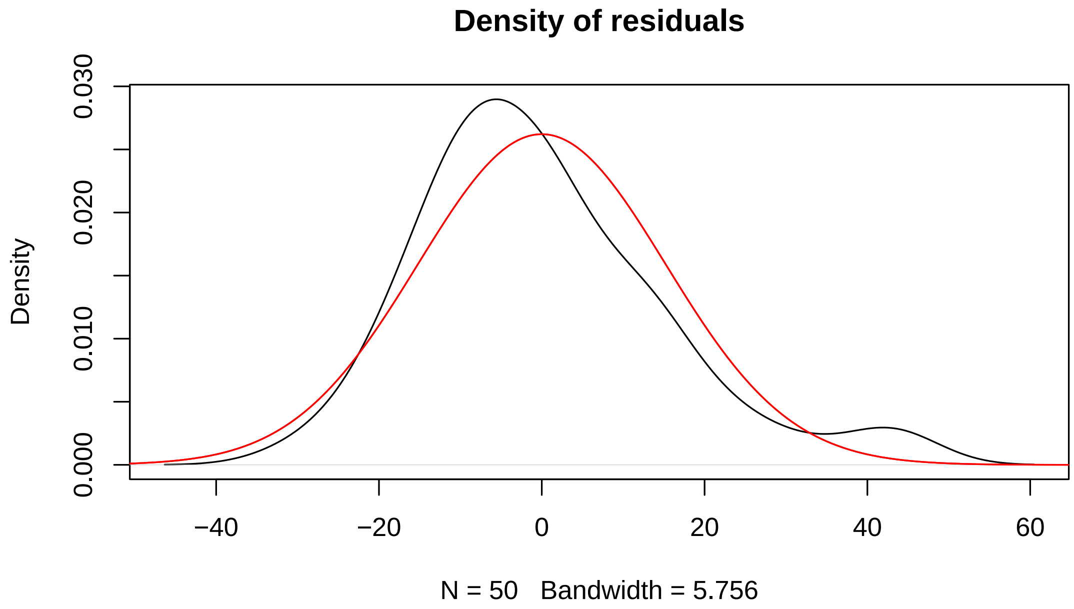
```
cars$res <- cars$dist - cars$dist_hat
```

**Residual density plot** detects the incorrect behavior of residuals.

```
d <- density(cars$res)
plot(d, main = "Density of residuals")
```

As we noted, residuals should be *normally distributed*. Let us compare this to normal distrituion.

```
SD <- sd(cars$res)
x <- seq(-70, 70 , length=500)
y <- dnorm(x, mean = 0, sd = SD)
points(x,y, type = 'l', col = "red")
```



## Some details of `lm()` summary

```
> summary(model1)

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
```

**Residuals** statistics give information about the symmetry of the residual distribution. The median should be close to 0 as the mean of the residuals is 0, and symmetric distributions have median=mean. Further, the 3Q and 1Q should be close to each other in magnitude. They would be equal under a symmetric 0 mean distribution. The max and min should also have similar magnitude.

**Significance stars** map p-values to symbols. They tell how statistically significant predictors the variables are. We will consider p-values later in this course.

**Residual standard error** is like standard deviation.

**The R-squared** statistic provides a measure of how well the model is fitting the actual data. It always lies between 0 and 1 (a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable).

In our example, 'multiple R-squared' is 0.6511. This means that 65% of the variance found in the response variable (dist) can be explained by the predictor variable (speed).

## Regression sum of squares

$$\text{SSR} = \sum_{n=1}^{n} (\widehat{y}_i - \bar{y})^2$$

How far the regression line $\widehat{Y}$ is from horisontal $\bar{Y}$ ( “no relationship” )

## Total sum of squares

$$\text{SSTO} = \sum_{n=1}^{n} (y_i - \bar{y})^2$$

How far the datapoints $Y$ are from horisontal $\bar{Y}$ ( “no relationship” )

**Multiple R-squared** is used for evaluating how well the model fits the data (between 0 and 1) – how close the data points are to the fitted regression line

$$R^2 = \frac{\text{SSR}}{\text{SSTO}}$$

“explained variation” / “total variation”

```
SSR <- sum((cars$dist_hat - y_bar)^2) # regression sum of squares
SSTO <- sum((cars$dist - y_bar)^2) # total sum of squares
[1] 0.6510794
```

- If $R^2 = 1$, all data points lie perfectly on the regression line
- $R^2 = 0$ means that the regression line is horizontal

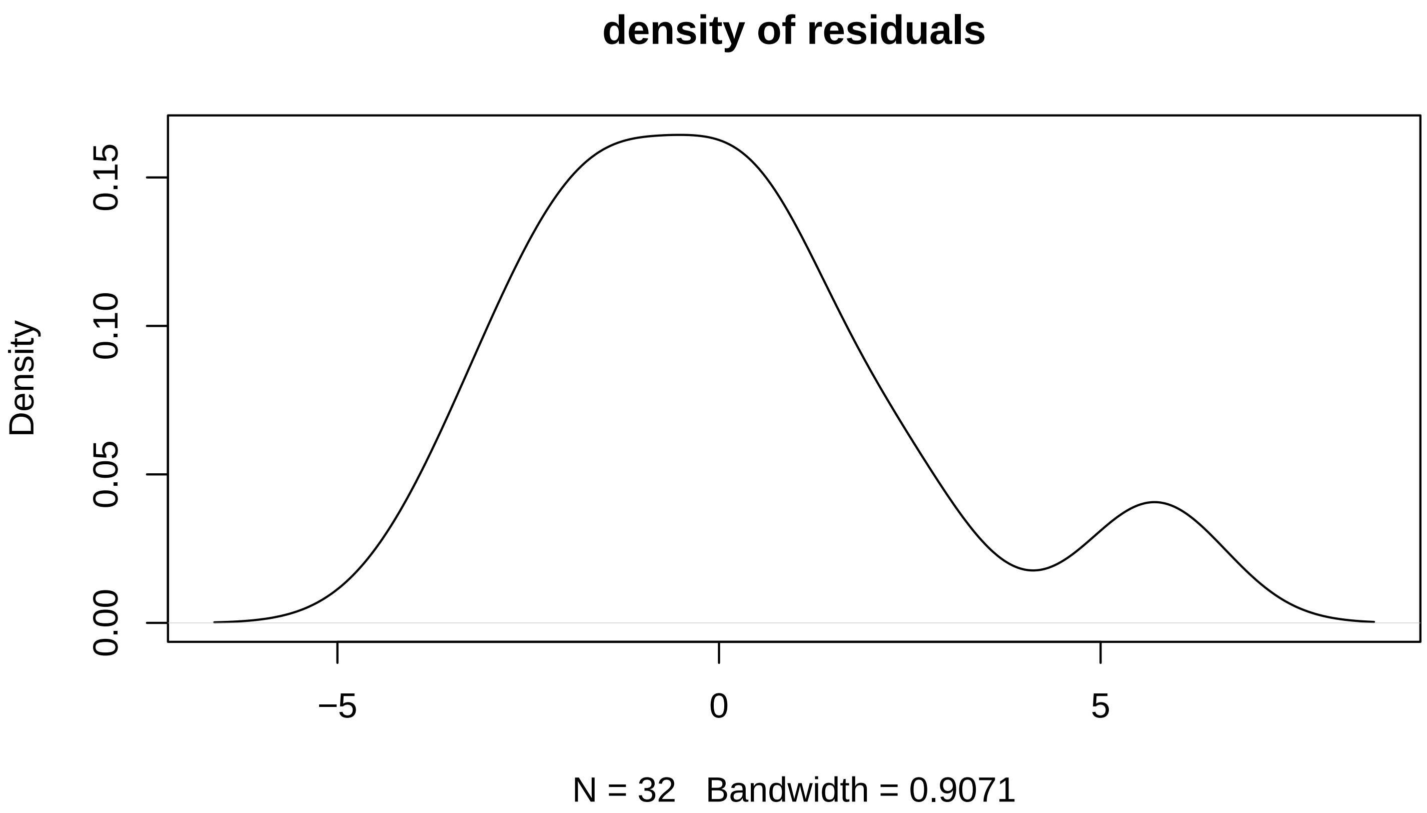
The **adjusted R-squared** is a modified version of R-squared taking into account the number of predictors in the model mtcars data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

| variable | description |
|---|---|
| mpg | Miles/(US) gallon |
| cyl | Number of cylinders |
| disp | Displacement (cu.in.) |
| hp | Gross horsepower |
| drat | Rear axle ratio (rel. to transimission) |
| wt | Weight (lb/1000) |
| qsec | 1/4 mile time |
| vs | V-engine 0 / Straight 1 |
| am | Transmission (0 = automatic, 1 = manual) |
| gear | Number of forward gears |
| carb | Number of carburetors |

Let us try with this kind of model:

```
df <- mtcars
m2 <- lm(mpg ~ wt + hp + disp, data = df)

df$mpg_hat <- predict(m2)
df$res <- df$mpg - df$mpg_hat
d2 <- density(df$res)
plot(d2, main="density of residuals"))
```


density of residuals
Density
0.00
0.05
0.10
0.15
−5
0
5
N = 32   Bandwidth = 0.9071