

# Subset dei dati

Esistono diversi modi per ottenere un sottoinsieme (\*subset\*) del dataset con cui stiamo lavorando: utilizzando l'[indicizzazione](#), utilizzando l'argomento subset dove previsto, con la funzione subset().

## Con l'indice degli elementi

```
# dati  
library(carData)  
data(SLID)
```

In generale, quando ci si riferisce ad elemento di un vettore o altro oggetto indicizzato con la parentesi quadra singola ([r,c]), si estrae un subset di dati.

```
str(SLID[1])
```

```
## 'data.frame':  7425 obs. of  1 variable:  
## $ wages: num  10.6 11 NA 17.8 NA ...
```

La variabile wages viene restituita come [dataframe](#) composto da una sola variabile: di fatto, è un subset.

## Selezionare solo alcune variabili

Quindi, ad esempio, per selezionare solo alcune variabili, possiamo scrivere:

```
SLID[,1:3]
```

```
   wages education age  
1  10.56      15.0  40  
2  11.00      13.2  19  
3    NA      16.0  49  
...
```

Vedi anche: [select](#), [rename](#) (dplyr)

## Selezionare solo alcune righe

Lo stesso accade per le righe:

```
str(SLID[1:10,])
```

```
## 'data.frame':  10 obs. of  5 variables:
## $ wages      : num  10.6 11 NA 17.8 NA ...
## $ education: num  15 13.2 16 14 8 16 12 14.5 15 10
## $ age        : int  40 19 49 46 71 50 70 42 31 56
## $ sex        : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 1 1 1
2 1
## $ language   : Factor w/ 3 levels "English","French",...: 1 1 3 3 1
1 1 1 1 1
```

Vedi anche: [filter \(dplyr\)](#)

## Esempi

Oppure, dal momento che la [funzione "table\(\)"](#) non accetta come argomento subset, possiamo usare questo sistema:

```
# table(SLID["language"])
# table(SLID[, "language"])
table(SLID$language)
```

```
##
## English   French   Other
##    5716      497    1091
```

Tabella della variabile *language* per i casi con età < 66:

```
table(SLID[SLID$age<66, "language"])
```

```
## English   French   Other
##    4911      426    867
```

Tabella *language* x *sex* (c(5,4)) per i casi con age < 66, ed esclusa la modalità *Other*:

```
table(SLID[SLID$age<66 & SLID$language!="Other", c(5,4)])
```

```
##           sex
## language Female Male
##   English   2535 2376
##   French     214  212
##   Other        0    0
```

```
# oppure
```

```
table(SLID[SLID$age<66 & SLID$language!="Other", c("language",  
"sex")])
```

## Argomento subset

L'argomento subset è utilizzato dalle funzioni che accettano l'argomento data. Ad esempio, t.test per i casi con age < 66:

```
t.test(wages~sex, alternative='two.sided',  
       conf.level=.95, data=SLID, subset = SLID$age<66)
```

```
##  
##  Welch Two Sample t-test  
##  
## data:  wages by sex  
## t = -14.029, df = 4021.2, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
##   -3.826934 -2.888429  
## sample estimates:  
## mean in group Female    mean in group Male  
##           13.88273           17.24041
```

## Funzione subset

La funzione subset ( ) opera su colonne o su righe:

```
# subset con tutte le variabili  
newdata <- subset(SLID, language == "English")
```

Per i dataframe è disponibile l'argomento select che permette di selezionare le colonne, avendo indicato le righe da selezionare. In questo modo, è possibile selezionare sia variabili che casi:

```
# subset con alcune variabili (ID e Weight)  
newdata <- subset(SLID, age > 24 | age < 65,  
                  select=c(wages, sex))
```

## Con le funzioni del tidyverse

vedi in particolare:

- [filter](#)
- [select](#)

# Script di esempio

E' possibile scaricare ed eseguire lo script dell'esempio:

[Subset\\_dati.R](#)

```
# dati
library(carData)
data(SLID)

# INDICIZZAZIONE

str(SLID[1])
str(SLID[1:10,])

# table(SLID["language"])
# table(SLID[, "language"])
table(SLID$language)

table(SLID[SLID$age<66, "language"])

table(SLID[SLID$age<66 & SLID$language!="Other", c(5,4)])

# ARGOMENTO SUBSET

t.test(wages~sex, alternative='two.sided',
       conf.level=.95, data=SLID, subset = SLID$age<66)

# FUNZIONE SUBSET

# subset con tutte le variabili
newdata <- subset(SLID, language == "English")

# subset con alcune variabili (ID e Weight)
newdata <- subset(SLID, age > 24 | age < 65,
                  select=c(ID, Weight))
```

[Gestione dei dati](#)

From:

<https://www.agnesevardanega.eu/wiki/> - Ricerca Sociale con R

Permanent link:

[https://www.agnesevardanega.eu/wiki/r/gestione\\_dei\\_dati/subset\\_dati](https://www.agnesevardanega.eu/wiki/r/gestione_dei_dati/subset_dati)

Last update: **11/08/2025 14:37**

