Statistica descrittiva bivariata

Massimo Aria

TABELLE DI CONTINGENZA

```
# importare il dataframe automobile (fonte UCI Machine Learning repository)

df=read.table("http://www.massimoaria.com/laboratorio/automobile.csv",header=TRUE,sep=";",dec=",")
str(df)
```

```
## 'data.frame':    205 obs. of  26 variables:
##  $ symbolig      : int  3 3 1 2 2 2 1 1 1 0 ...
##  $ normlosses    : int  NA NA NA 164 164 NA 158 NA 158 NA ...
##  $ make          : Factor w/ 22 levels "alfa-romero",..: 1 1 1 2 2 2 2 2 2 2 ...
##  $ fuel          : Factor w/ 2 levels "diesel","gas": 2 2 2 2 2 2 2 2 2 2 ...
##  $ aspiration    : Factor w/ 2 levels "std","turbo": 1 1 1 1 1 1 1 1 2 2 ...
##  $ doors         : Factor w/ 3 levels "","four","two": 3 3 3 2 2 3 2 2 2 3 ...
##  $ body          : Factor w/ 5 levels "convertible",..: 1 1 3 4 4 4 4 5 4 3 ...
##  $ drivewheels   : Factor w/ 3 levels "4wd","fwd","rwd": 3 3 3 2 1 2 2 2 2 1 ...
##  $ enginelocation: Factor w/ 2 levels "front","rear": 1 1 1 1 1 1 1 1 1 1 ...
##  $ wheelbase     : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ length        : num  169 169 171 177 177 ...
##  $ width         : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
##  $ height        : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
##  $ curbweight    : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
##  $ engine        : Factor w/ 7 levels "dohc","dohcv",..: 1 1 6 4 4 4 4 4 4 4 ...
##  $ cylinders     : Factor w/ 7 levels "eight","five",..: 3 3 4 3 2 2 2 2 2 2 ...
##  $ enginesize    : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ fuelsystem    : Factor w/ 8 levels "1bbl","2bbl",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ bore          : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.13 ...
##  $ stroke        : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
##  $ compression   : num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ hp            : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ rpm           : int  5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
##  $ cityMPG       : int  21 21 19 24 18 19 19 19 17 16 ...
##  $ highwayMPG    : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price         : int  13495 16500 16500 13950 17450 15250 17710 18920 23875 NA ...
```

```
# rendiamo il data frame df come oggetto di default
attach(df)

# creazione di una tabella di contingenza con il comando table

# tabella di frequenze assolute
T <- table(body,fuel)
T
```

```
##              fuel
## body          diesel gas
##   convertible      0   6
##   hardtop          1   7
##   hatchback        1  69
##   sedan           15  81
```

```
##   wagon            3  22
```

```
# tabella di frequenze relative
table(body,fuel)/dim(df)[1]
```

```
##              fuel
## body               diesel         gas
##   convertible 0.000000000 0.029268293
##   hardtop     0.004878049 0.034146341
##   hatchback   0.004878049 0.336585366
##   sedan       0.073170732 0.395121951
##   wagon       0.014634146 0.107317073
```

```
# oppure
prop.table(T)
```

```
##              fuel
## body               diesel         gas
##   convertible 0.000000000 0.029268293
##   hardtop     0.004878049 0.034146341
##   hatchback   0.004878049 0.336585366
##   sedan       0.073170732 0.395121951
##   wagon       0.014634146 0.107317073
```

```
# tabella di frequenze percentuali
Tp=table(body,fuel)/dim(df)[1]*100
print(Tp, digit=2)
```

```
##              fuel
## body          diesel   gas
##   convertible   0.00  2.93
##   hardtop       0.49  3.41
##   hatchback     0.49 33.66
##   sedan         7.32 39.51
##   wagon         1.46 10.73
```

```
# calcolo dei marginali di riga e colonna di una tabella doppia

rowSums(T)  # restituisce la somma per riga di una tabella
```

```
## convertible     hardtop   hatchback       sedan       wagon
##           6           8          70          96          25
```

```
colSums(T)  # restituisce la somma per colonna di un tabella
```

```
## diesel    gas
##     20    185
```

```
# identico risultato può essere ottenuto con il comando margin.table

margin.table(T,1)  # per i marignali di riga
```

```
## body
## convertible     hardtop   hatchback       sedan       wagon
##           6           8          70          96          25
```

```
margin.table(T,2)  # per i marginali di colonna
```

```
## fuel
```

```
## diesel    gas
##     20    185
```

```
sum(T) # restituisce la somma totale degli elementi di una tabella
```

```
## [1] 205
```

```
# DISTRIBUZIONI CONDIZIONATE

# per ottenere a distribuzione condizionata di un carattere ad un attributo dell'altro carattere

T[,1]/sum(T[,1]) # distribuzione della variabile body condizionata alla prima modalità (diesel) della v
```

```
## convertible     hardtop   hatchback       sedan       wagon
##        0.00        0.05        0.05        0.75        0.15
```

```
T[,2]/sum(T[,2]) # distribuzione della variabile body condizionata alla prima modalità (diesel) della v
```

```
## convertible     hardtop   hatchback       sedan       wagon
##  0.03243243  0.03783784  0.37297297  0.43783784  0.11891892
```

```
# con un ciclo for si possono ottenere tutte le distribuzioni condizionate

# per riga
for (i in 1:dim(T)[2]){
  print(T[,i]/sum(T[,i]))
}
```

```
## convertible     hardtop   hatchback       sedan       wagon
##        0.00        0.05        0.05        0.75        0.15
## convertible     hardtop   hatchback       sedan       wagon
##  0.03243243  0.03783784  0.37297297  0.43783784  0.11891892
```

```
# e per colonna
for (i in 1:dim(T)[1]){
  print(T[i,]/sum(T[i,]))
}
```

```
## diesel    gas
##      0      1
## diesel    gas
##  0.125  0.875
##     diesel        gas
## 0.01428571 0.98571429
##  diesel     gas
## 0.15625 0.84375
## diesel    gas
##   0.12   0.88
```

MISURARE IL GRADO DI ASSOCIAZIONE IN UNA TABELLA DI CONTINGENZA

Il grado di associazione si misura con l'indice di associazione χ^2 di Pearson

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

dove

$$n_{ij}^* = \frac{n_{i+} * n_{+j}}{N}$$

rappresenta la frequenza teorica per la generica cella (i,j) in caso di indipendenza assoluta tra le due variabili.

```
# L'associazione può essere misurata con il comando summary applicato ad un oggetto della classe table

summary(T)
```

```
## Number of cases in table: 205
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 10.129, df = 4, p-value = 0.0383
##  Chi-squared approximation may be incorrect
```

```
# ESERCIZIO:
# Scrivere una funzione che calcola il Chi quadrato senza utilizzare il comando summary
```

ANALISI DI REGRESSIONE

Grafici di dispersione

```
# installiamo il pacchetto gdata per poter importare file excel

# install.packages("rio") #l'installazione va effettuata solo la prima volta

# con il comando library carichiamo il pacchetto in memoria
library(rio)

# usiamo il comando import per leggere il file excel
df <- import("http://www.massimoaria.com/laboratorio/impiegati.xlsx")
str(df)
```

```
## 'data.frame':    474 obs. of  9 variables:
##  $ gender  : chr  "Male" "Male" "Female" "Female" ...
##  $ educ    : num  15 16 12 8 15 15 15 12 15 12 ...
##  $ jobcat  : chr  "Manager" "Clerical" "Clerical" "Clerical" ...
##  $ salary  : num  57000 40200 21450 21900 45000 ...
##  $ salbegin: num  27000 18750 12000 13200 21000 ...
##  $ jobtime : num  98 98 98 98 98 98 98 98 98 98 ...
##  $ prevexp : num  144 36 381 190 138 67 114 0 115 244 ...
##  $ minority: chr  "No" "No" "No" "No" ...
##  $ age     : num  49 43 72 54 46 43 45 35 55 55 ...
```

```
attach(df)
head(df)
```

```
##   gender educ   jobcat salary salbegin jobtime prevexp minority age
## 1   Male   15  Manager  57000    27000      98     144       No  49
## 2   Male   16 Clerical  40200    18750      98      36       No  43
## 3 Female   12 Clerical  21450    12000      98     381       No  72
## 4 Female    8 Clerical  21900    13200      98     190       No  54
## 5   Male   15 Clerical  45000    21000      98     138       No  46
## 6   Male   15 Clerical  32100    13500      98      67       No  43
```

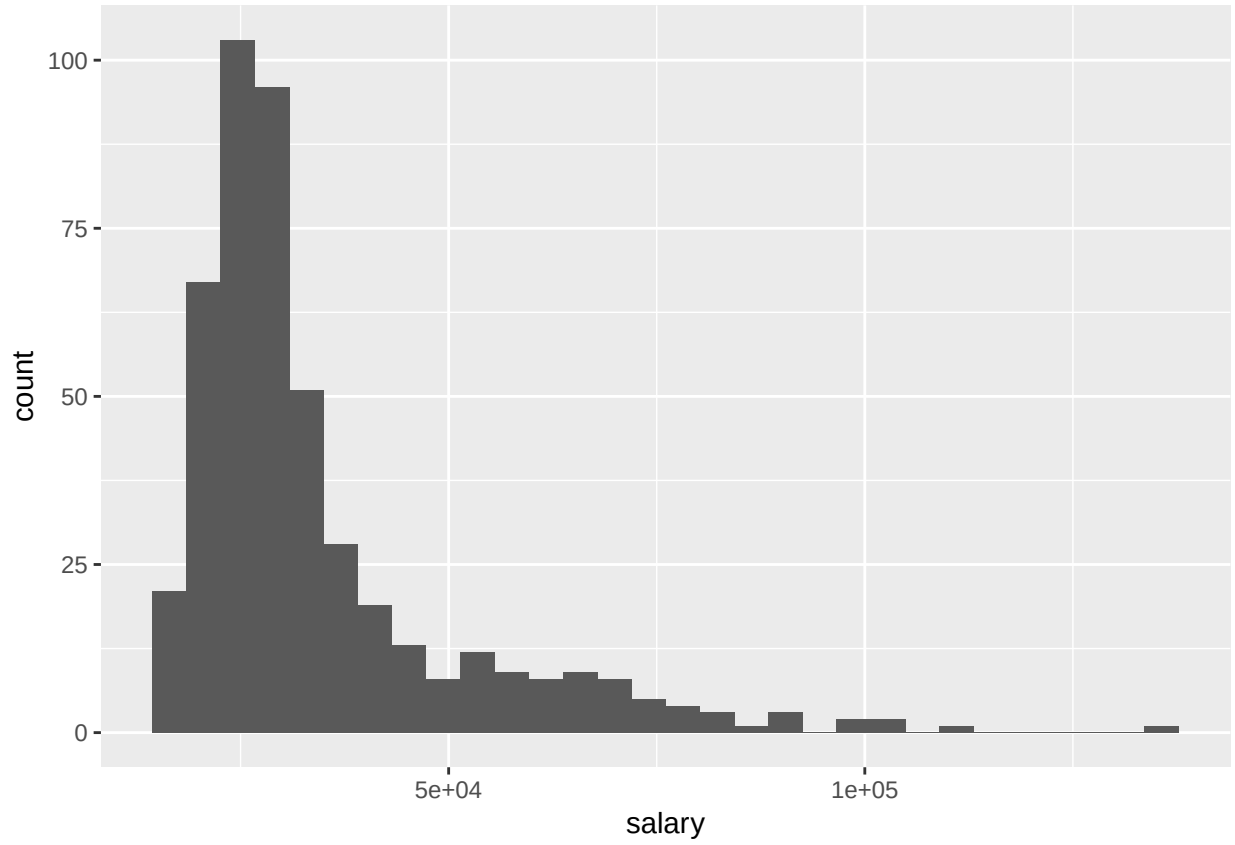
```
# installiamo il pacchetto ggplot2 per creare grafici con R
# install.packages("ggplot2")
# e carichiamo in memoria la libreria
library(ggplot2)

# usiamo la funzione qplot per disegnare istogrammi e scatterplot
```

```
?qplot
```

```
## starting httpd help server ... done
```

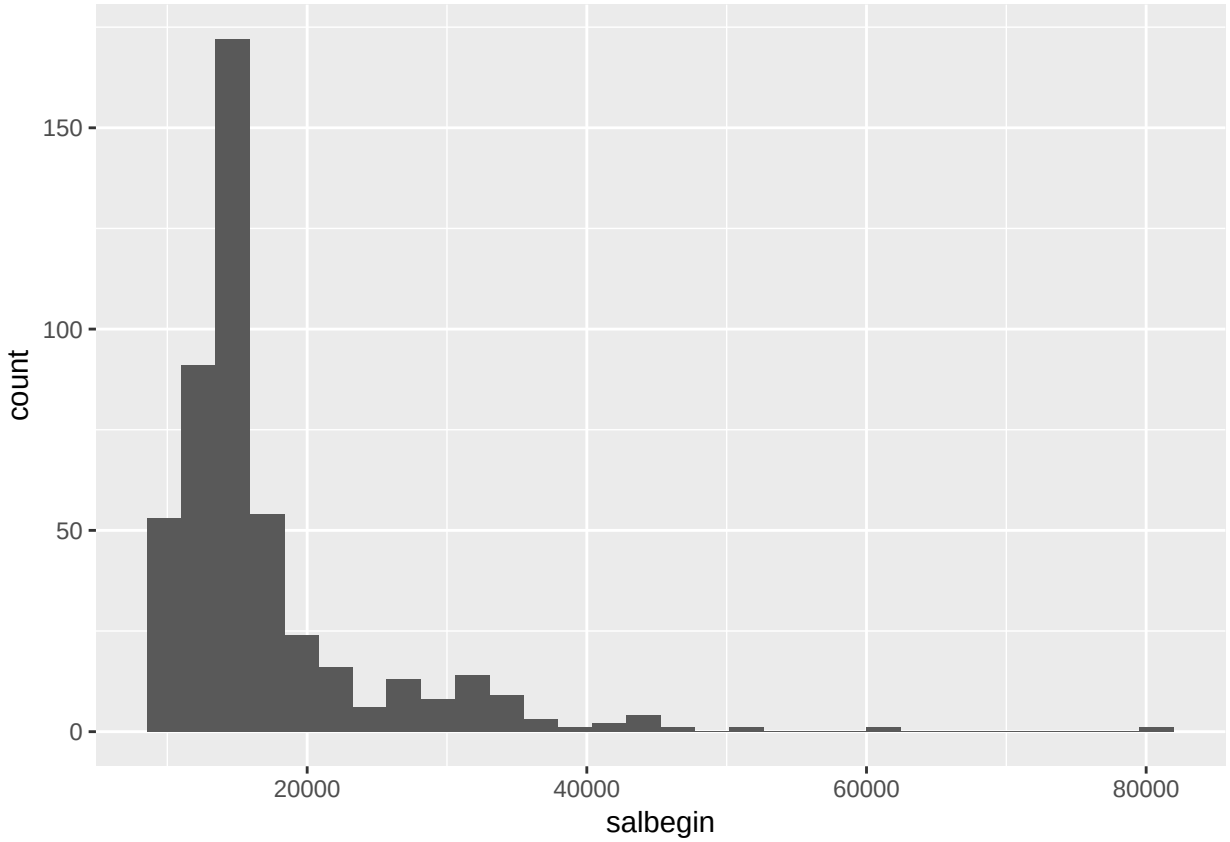
```
qplot(salary, geom="histogram")   # istogramma del salario attuale
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

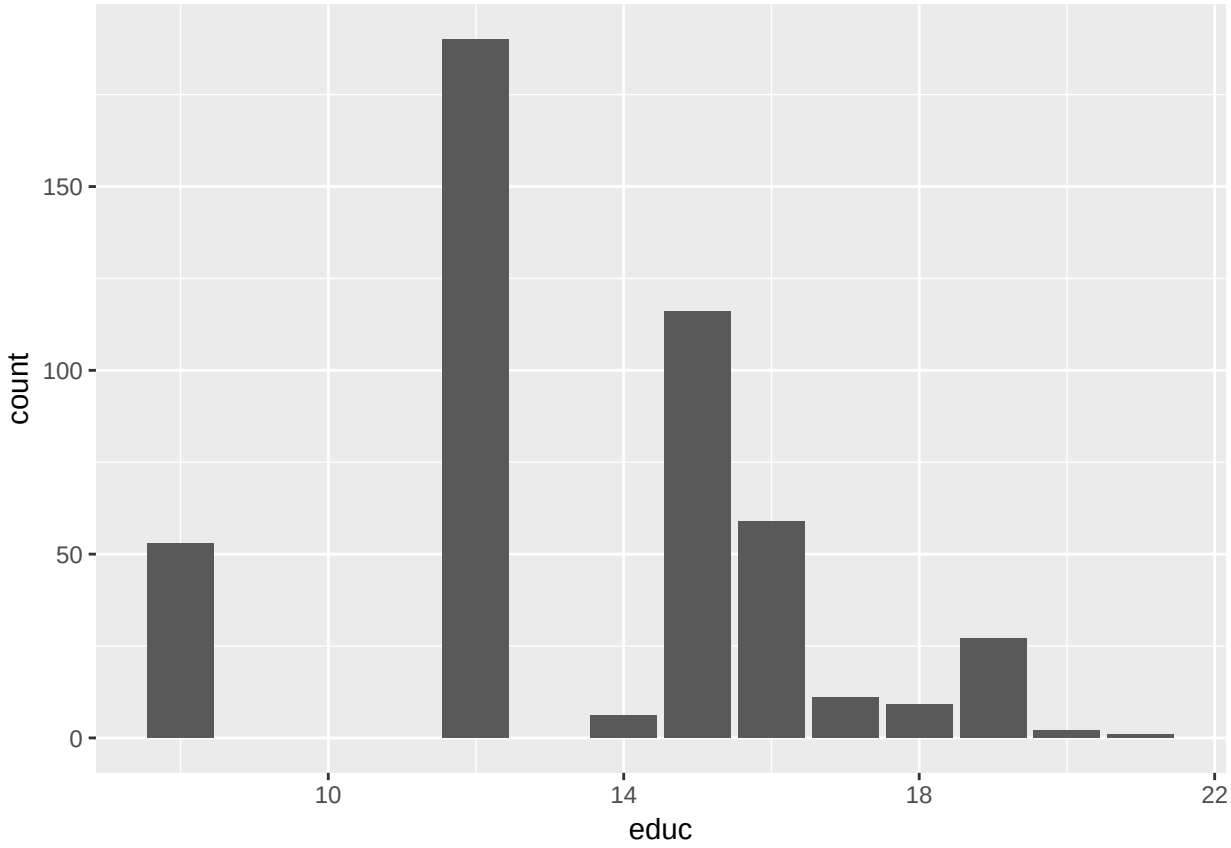



```
qplot(salbegin, geom="histogram") # istogramma del salario iniziale
```

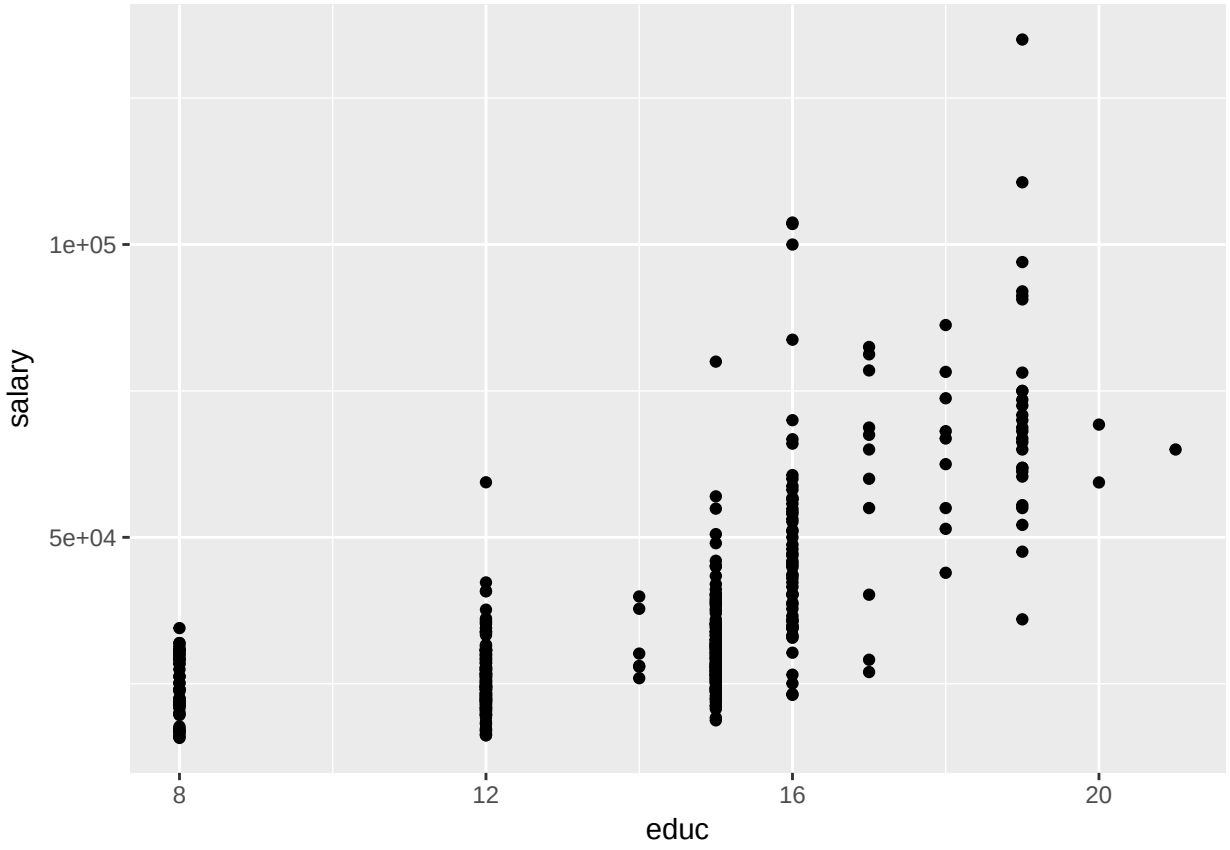
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



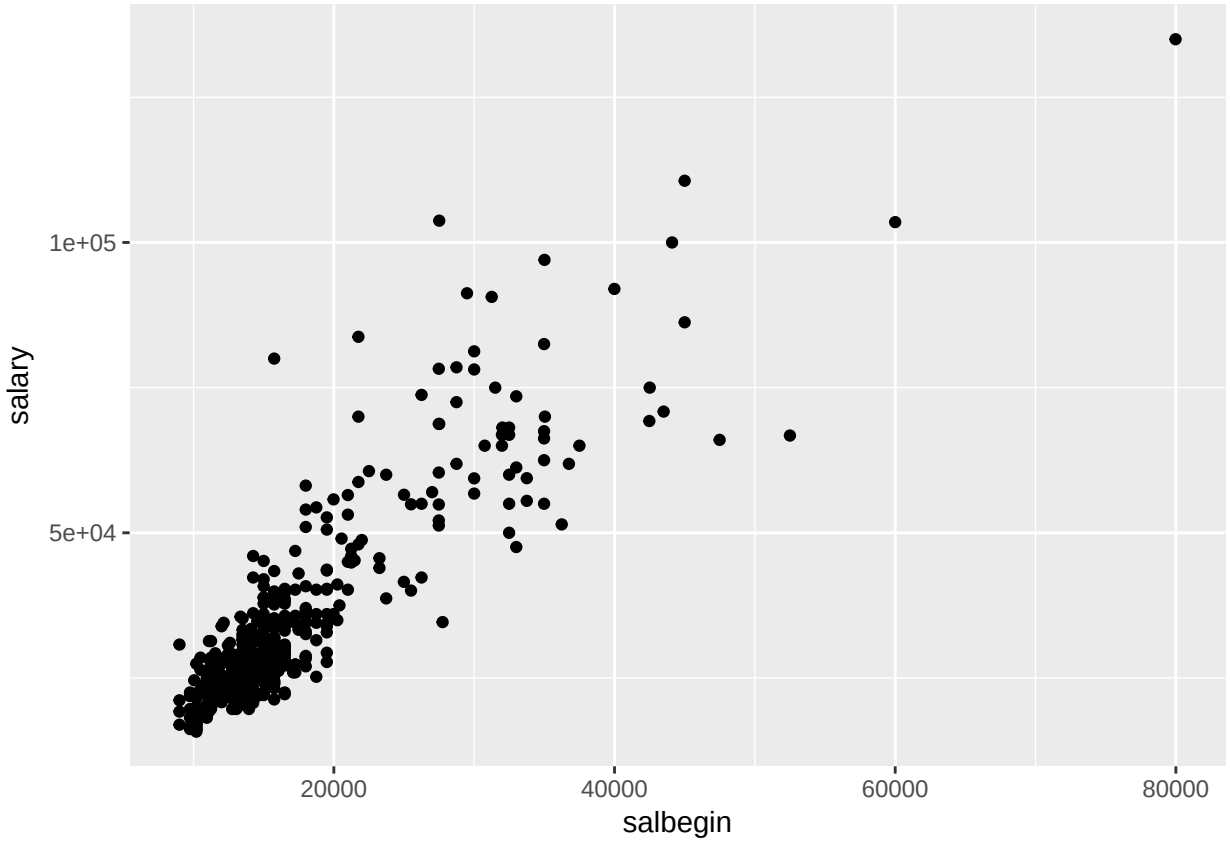

```
qplot(educ, geom="bar")      # diagramma a barre del livello di educazione
```



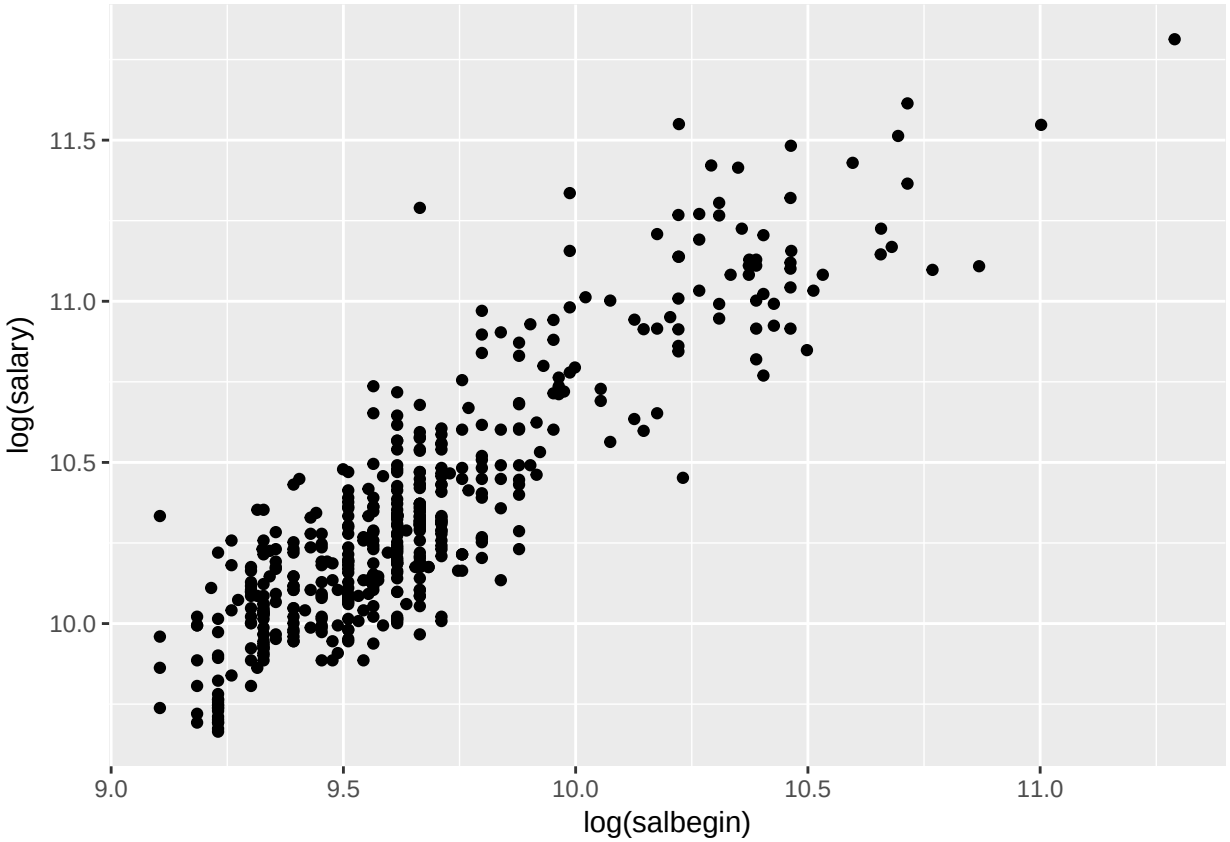

```
# scatterplot
qplot(educ,salary)
```



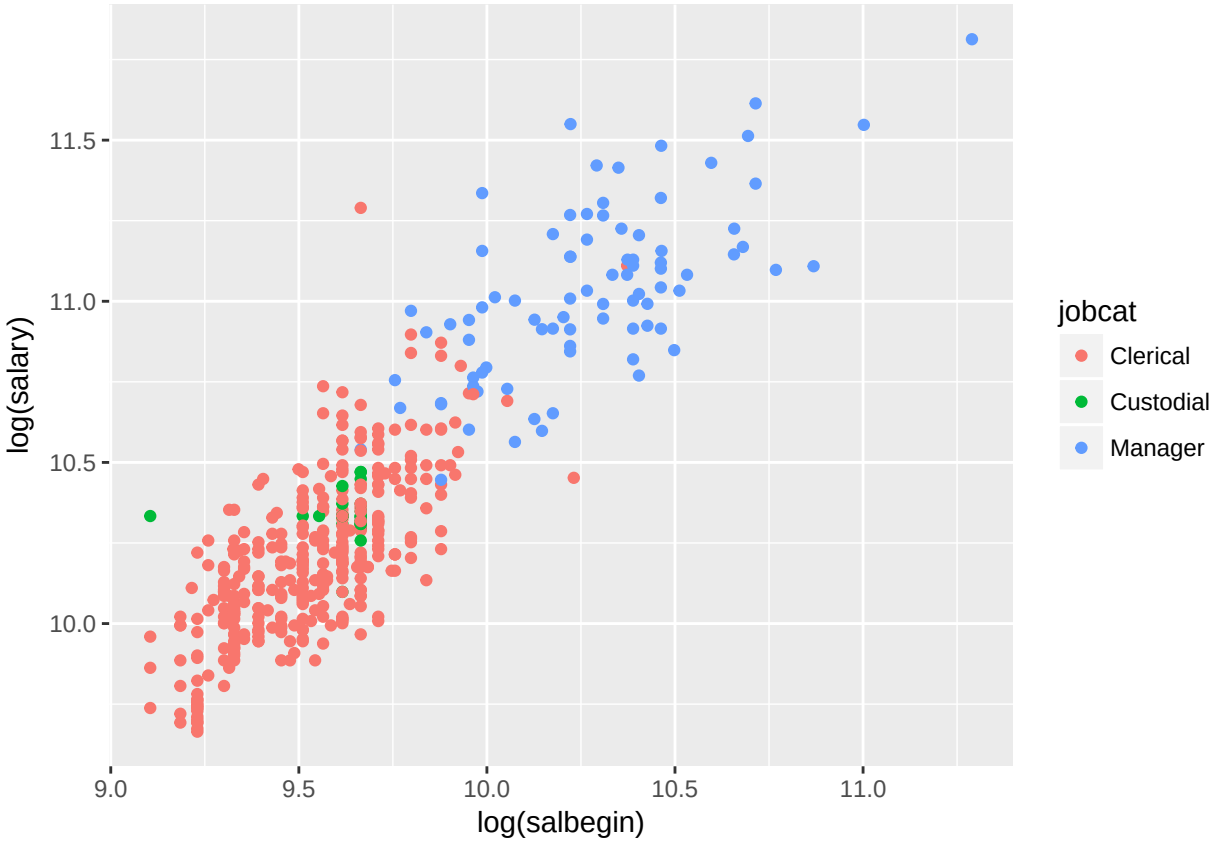

```
qplot(salbegin,salary)
```



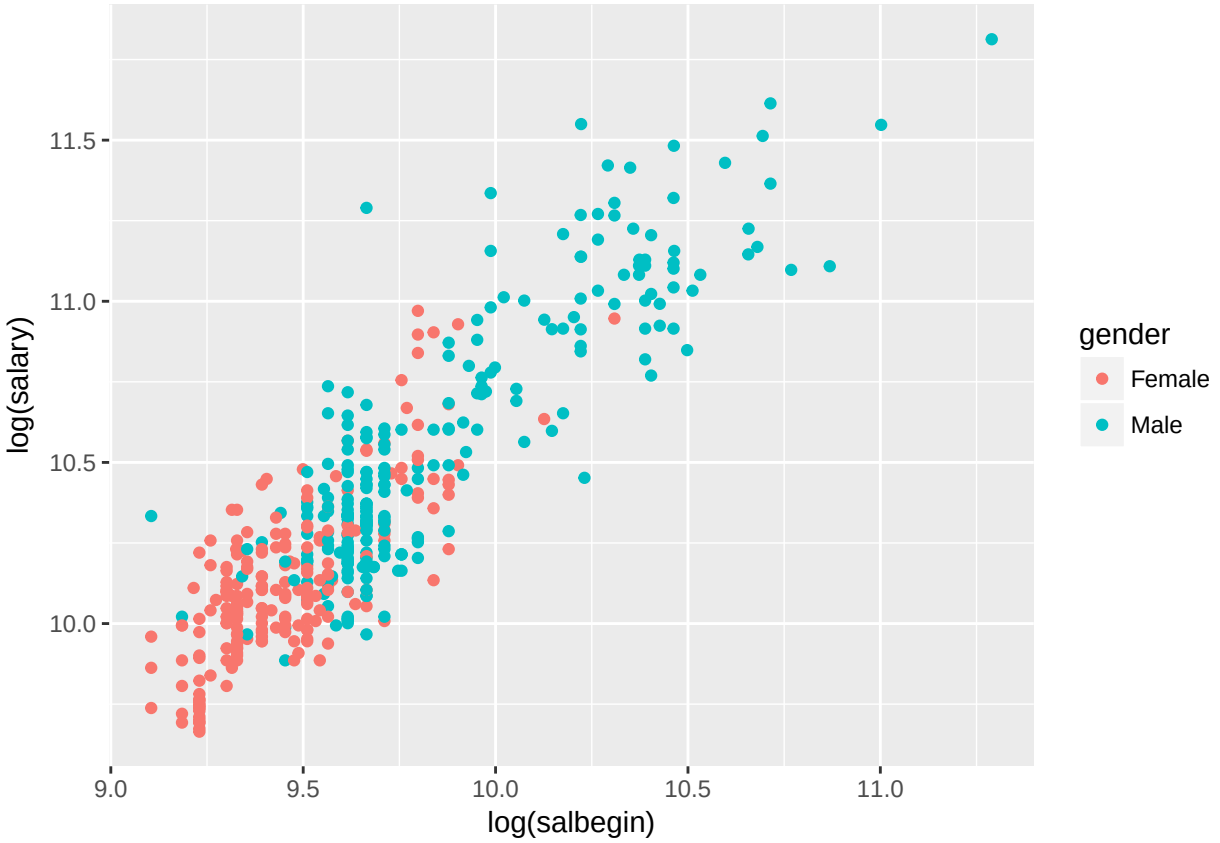

```
# scatterplot con trasforazione logaritmica
qplot(log(salbegin),log(salary))
```




```
# con punti colorati per tipologia di impiego
qplot(log(salbegin),log(salary),colour=jobcat)
```




```
# e per genere
qplot(log(salbegin),log(salary),colour=gender)
```



CORRELAZIONE LINEARE

si misura con il coefficiente di correlazione lineare ρ di Bravais-Pearson

$$\rho_{x,y} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{\sigma_x * \sigma_y}$$

dove x e y sono due generiche variabili numeriche, il numeratore rappresenta la misura della loro covarianza e al denominatore σ_x e σ_y rappresentano rispettivamente la deviazione standard di x e y

in R è possibile calcolarlo attraverso la funzione cor

```
?cor

cor(salbegin,salary)
```

```
## [1] 0.8801175
```

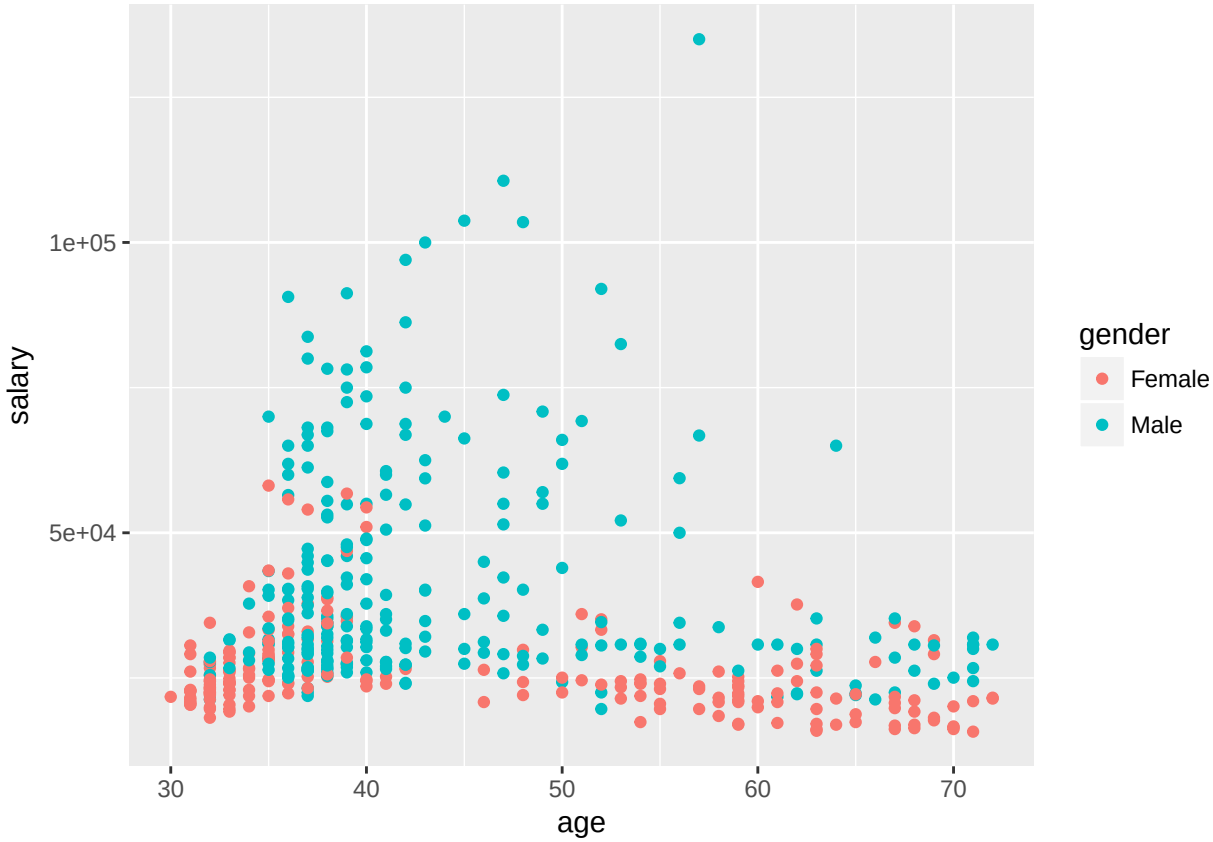
```
cor(age,salary)
```

```
## [1] -0.145844
```

```
cor(log(salbegin),log(salary))
```

```
## [1] 0.8863679
```

```
qplot(age, salary,colour=gender)
```


salary
1e+05
5e+04
30
40
50
60
70
age
gender
Female
Male