# Introduzione ai modelli di regressione

Statistica Applicata
Corso di Laurea in Informatica

cristiano.varin@unive.it

## Indice



## 1 Mercato immobiliare

Il foglio elettronico `HousePrice.csv`[1] contiene dati sul prezzo delle abitazioni in una delle principali aree metropolitane americane. Le informazioni disponibili sono:

`Price` il prezzo dell'abitazione in US $;

[1] Il dataset è tratto da *Jank, W. (2011). Business Analytics for Managers. Springer.*

`SqFt` la dimensione dell'abitazione (in piedi al quadrato);

`Bedrooms` il numero di camere da letto;

`Bathrooms` il numero di bagni;

`Offers` il numero di offerte di acquisto ricevute da quando l'abitazione è sul mercato;

`Brick` se la casa ha muri di mattone o no;

`Neighborhood` il distretto dove si trova la casa (in questi dati i possibili distretti sono `est`, `west` e `north`).

**Obiettivi**: comprendere quali siano i *driver* del prezzo di un immobile, valutare quanto e come aumenti il prezzo dell'abitazione al variare della dimensione, del numero di camere da letto, ...

## 2 Analisi esplorative

Lettura dati `HousePrices`

```
## [1] 128   8
##   HomeID  Price SqFt Bedrooms Bathrooms Offers Brick Neighborhood
## 1      1 114300 1790        2         2      2    No         East
## 2      2 114200 2030        4         2      3    No         East
## 3      3 114800 1740        3         2      1    No         East
## 4      4  94700 1980        3         2      3    No         East
## 5      5 119800 2130        3         3      3    No         East
## 6      6 114600 1780        3         2      2    No        North
## [1] "HomeID"       "Price"        "SqFt"         "Bedrooms"
## [5] "Bathrooms"    "Offers"       "Brick"        "Neighborhood"
```

```
house <- read.csv( "HousePrices.csv" )
dim(house)
head(house)
names(house)
```

Riassunto dei dati

```
summary(house)
```

```
##      HomeID          Price             SqFt         Bedrooms
##  Min.   :  1.0   Min.   : 69100   Min.   :1450   Min.   :2.00
##  1st Qu.: 32.8   1st Qu.:111325   1st Qu.:1880   1st Qu.:3.00
##  Median : 64.5   Median :125950   Median :2000   Median :3.00
```

```
##  Mean   : 64.5   Mean   :130427   Mean   :2001   Mean   :3.02
##  3rd Qu.: 96.2   3rd Qu.:148250   3rd Qu.:2140   3rd Qu.:3.00
##  Max.   :128.0   Max.   :211200   Max.   :2590   Max.   :5.00
##    Bathrooms        Offers     Brick    Neighborhood
##  Min.   :2.00   Min.   :1.00   No :86   East :45
##  1st Qu.:2.00   1st Qu.:2.00   Yes:42   North:44
##  Median :2.00   Median :3.00            West :39
##  Mean   :2.44   Mean   :2.58
##  3rd Qu.:3.00   3rd Qu.:3.00
##  Max.   :4.00   Max.   :6.00
```

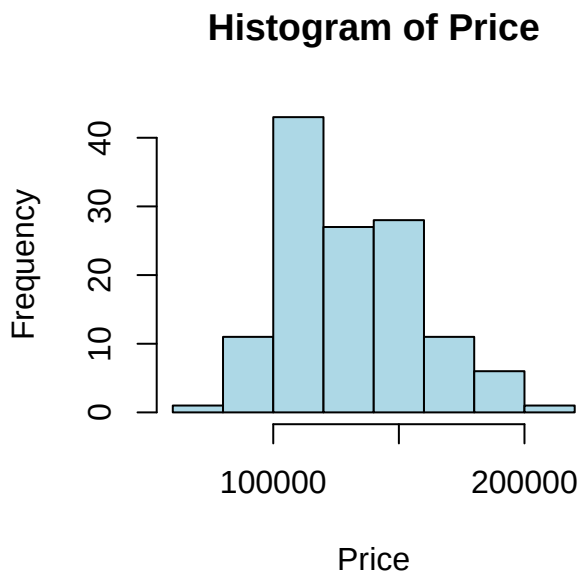
## 2.1 Prezzo

Riassunto del prezzo

```
attach(house)
summary(Price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   69100  111000  126000  130000  148000  211000
```

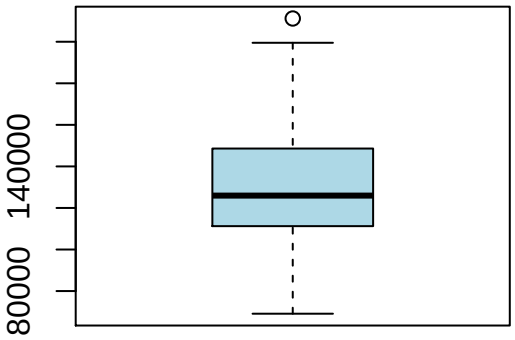
Istogramma

```
hist( Price, col = "lightblue" )
```



Boxplot

```
boxplot( Price, col = "lightblue" )
```




## 2.2 Prezzo e dimensione

Riassunto della dimensione

```
summary(SqFt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1450    1880    2000    2000    2140    2590
```

Boxplot

```r
boxplot( SqFt, col = "lightblue" )
```

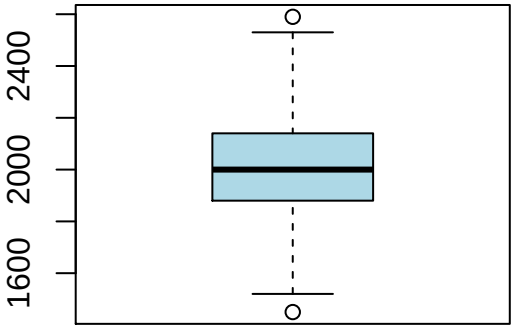
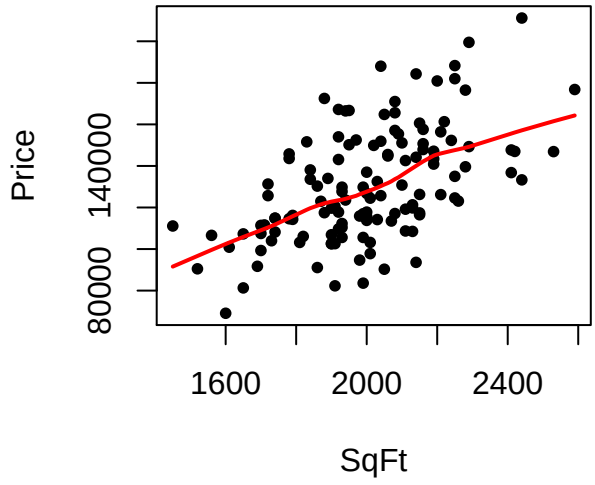



Grafico a dispersione del prezzo rispetto alla dimensione

```r
plot( Price ~ SqFt, pch = 20 )
## aggiungiamo una linea che indica "la relazione" media
lines( lowess( Price ~ SqFt ), col = "red", lwd = 2 )
```




Correlazione fra prezzo e dimensione

```r
cor( Price, SqFt )

## [1] 0.553
```

## 2.3 Prezzo e stanze da letto

Riassunto del numero di stanze da letto

```r
summary(Bedrooms)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    3.00    3.00    3.02    3.00    5.00
```

Boxplot (non molto informativo...)

```r
boxplot( Bedrooms, col = "lightblue" )
```

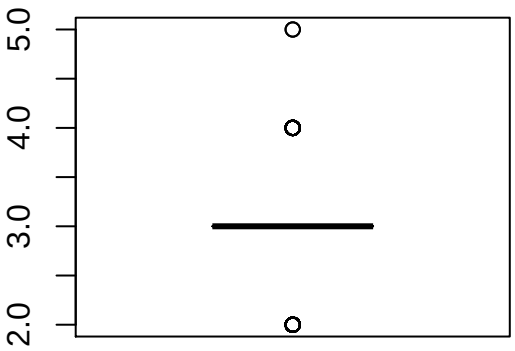



Tabella di frequenza

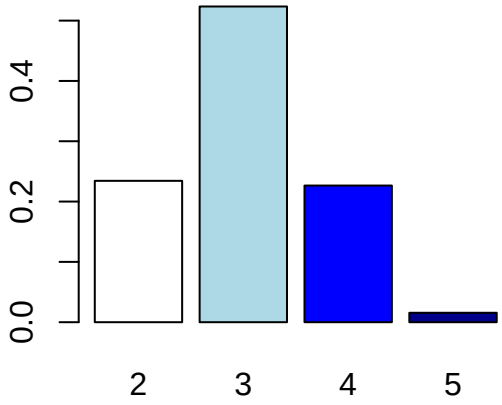
```r
table(Bedrooms)

## Bedrooms
##  2  3  4  5
## 30 67 29  2
```

Tabella di frequenza relativa

```
tab.bed <- prop.table( table(Bedrooms) )
tab.bed

## Bedrooms
##       2       3       4       5
## 0.23438 0.52344 0.22656 0.01562
```

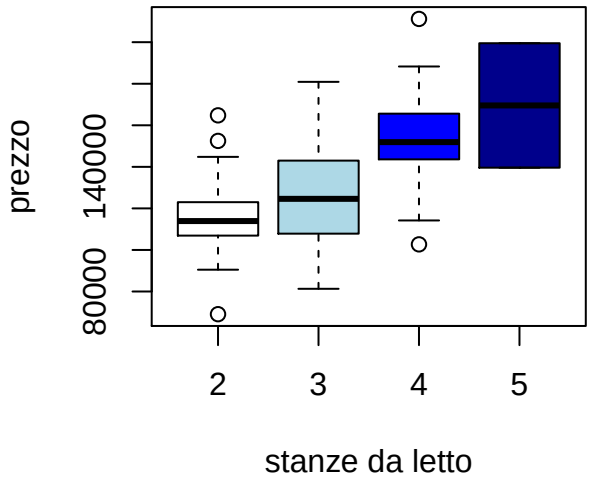
Grafico a barre della tabella di frequenza relativa

```
barplot( tab.bed, col = c("white", "lightblue", "blue", "darkblue") )
```




Boxplot del prezzo condizionatamente al numero di stanze da letto

```
boxplot( Price ~ Bedrooms, col = c("white", "lightblue", "blue",
"darkblue"), xlab = "stanze da letto", ylab = "prezzo" )
```




Correlazione fra prezzo e numero di stanze

```
cor( Price, Bedrooms )

## [1] 0.5259
```

## 2.4 Prezzo e bagni

Riassunto del numero di stanze da bagno

```
summary(Bathrooms)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    2.00    2.00    2.45    3.00    4.00
```

Tabella di frequenza

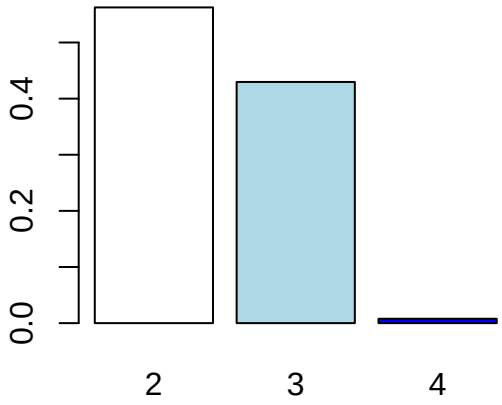
```
table(Bathrooms)

## Bathrooms
##  2  3  4
## 72 55  1
```

Tabella di frequenza relativa

```
tab.bath <- prop.table( table(Bathrooms) )
tab.bath

## Bathrooms
##        2        3        4
## 0.562500 0.429688 0.007812
```

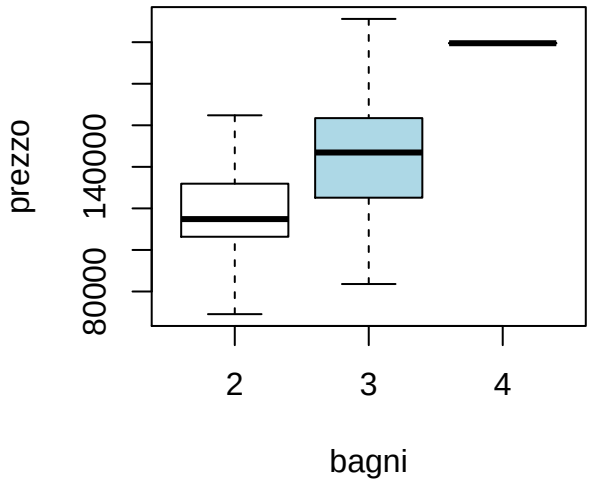
Grafico a barre

```
barplot( tab.bath, col = c("white", "lightblue", "blue") )
```




Boxplot del prezzo condizionatamente al numero di stanze da bagno

```
boxplot( Price ~ Bathrooms, , col = c("white", "lightblue", "blue"),
xlab = "bagni", ylab = "prezzo" )
```




Correlazione fra prezzo e stanze da bagno

```
cor( Price, Bathrooms )

## [1] 0.5233
```

## 2.5 Prezzo e offerte

Riassunto del numero di offerte

```
summary(Offers)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    3.00    2.58    3.00    6.00
```

Tabella di frequenze

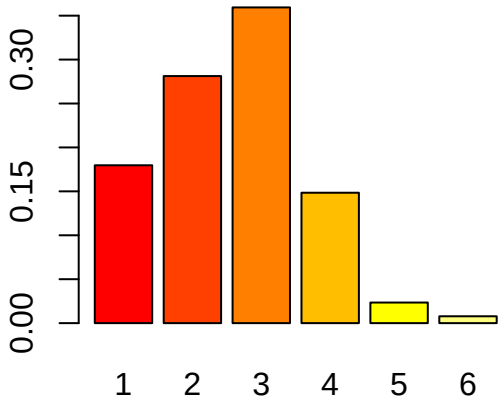
```
table(Offers)

## Offers
##  1  2  3  4  5  6
## 23 36 46 19  3  1
```

Tabella di frequenze relative

```
tab.offers <- prop.table( table(Offers) )
tab.offers

## Offers
##        1        2        3        4        5        6
## 0.179688 0.281250 0.359375 0.148438 0.023438 0.007812
```

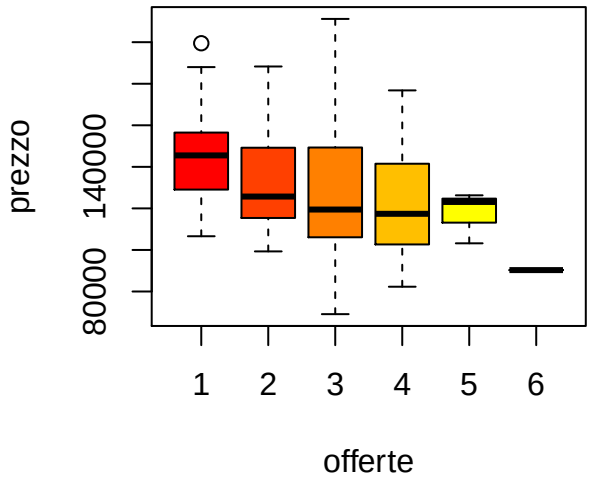
Grafico a barre

```
barplot( tab.offers, col = heat.colors(6) )
```




Boxplot del prezzo condizionatamente al numero di offerte

```
boxplot( Price ~ Offers, col = heat.colors(6), xlab = "offerte",
ylab = "prezzo" )
```




Correlazione fra prezzo e offerte

```
cor( Price, Offers )

## [1] -0.3136
```

## 2.6 Prezzo e mattoni

Riassunto delle variabile `Brick`

```
summary(Brick)

##  No Yes
##  86  42
```

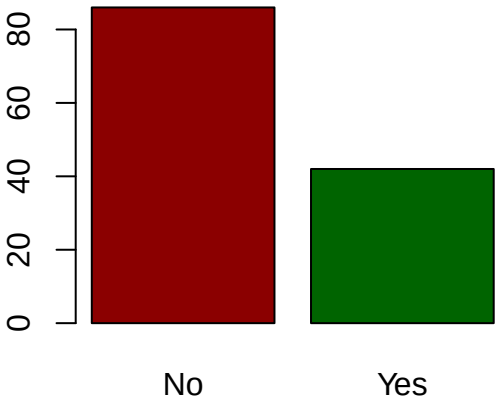
La variabile `Brick` è una variabile categoriale, ovvero un '`factor`'

```
class(Brick)

## [1] "factor"

class(Price)

## [1] "integer"

class(Bedrooms)
```

```
## [1] "integer"

class(Neighborhood)

## [1] "factor"
```

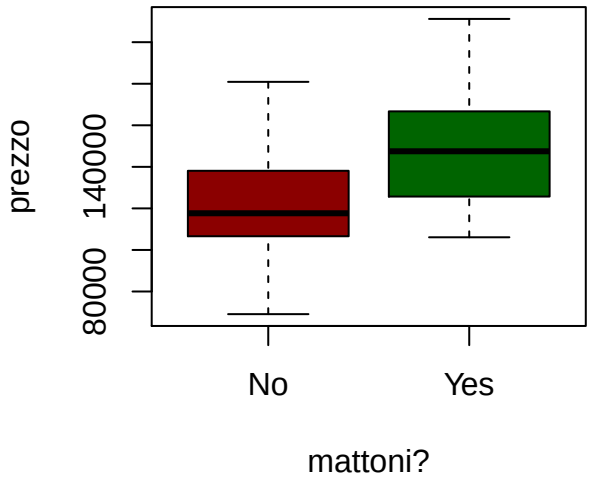
Grafico a barre

```
plot( Brick, col = c("darkred", "darkgreen") )
```




Boxplot del prezzo condizionatamente alla presenza o meno di mattoni

```
boxplot( Price ~ Brick, col = c("darkred", "darkgreen"),
ylab = "prezzo", xlab = "mattoni?" )
```




Sintesi numerica del prezzo a seconda della presenza o meno di mattoni

```
by( Price, Brick, summary )

## Brick: No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   69100  107000  118000  122000  138000  181000
## ------------------------------------------------------------
## Brick: Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  106000  126000  148000  148000  167000  211000
```

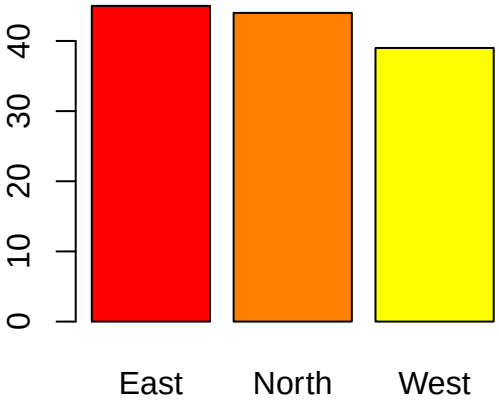
## 2.7 Prezzo e quartiere

Riassunto di `Neighborhood`

```
summary( Neighborhood )

##  East North  West
##    45    44    39
```

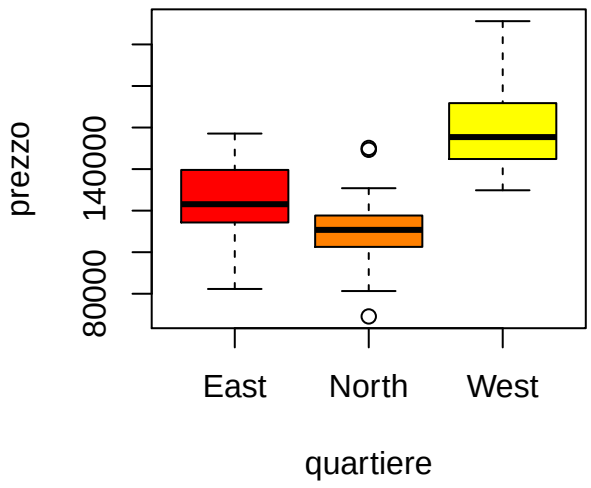
Grafico a barre

```
plot( Neighborhood, col = heat.colors(3) )
```




Boxplot del prezzo condizionatamente al quartiere

```
boxplot( Price ~ Neighborhood, col = heat.colors(3),
ylab = "prezzo", xlab = "quartiere" )
```



Sintesi del prezzo al variare del quartiere

```
by( Price, Neighborhood, summary )

## Neighborhood: East
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   82300  114000  123000  125000  140000  157000
## ------------------------------------------------------------
## Neighborhood: North
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   69100  103000  111000  110000  118000  150000
## ------------------------------------------------------------
## Neighborhood: West
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  130000  145000  155000  159000  172000  211000
```

# 3 Modelli di regressione

## 3.1 Matrici di correlazione

Matrice di correlazione delle variabili numeriche contenute in `house`

```
cor.matrix <- cor( cbind( Price, SqFt, Bedrooms, Bathrooms, Offers ) )
round(cor.matrix, 2)

##            Price SqFt Bedrooms Bathrooms Offers
## Price       1.00 0.55     0.53      0.52  -0.31
## SqFt        0.55 1.00     0.48      0.52   0.34
## Bedrooms    0.53 0.48     1.00      0.41   0.11
## Bathrooms   0.52 0.52     0.41      1.00   0.14
## Offers     -0.31 0.34     0.11      0.14   1.00
```

## 3.2 Regressione lineare semplice

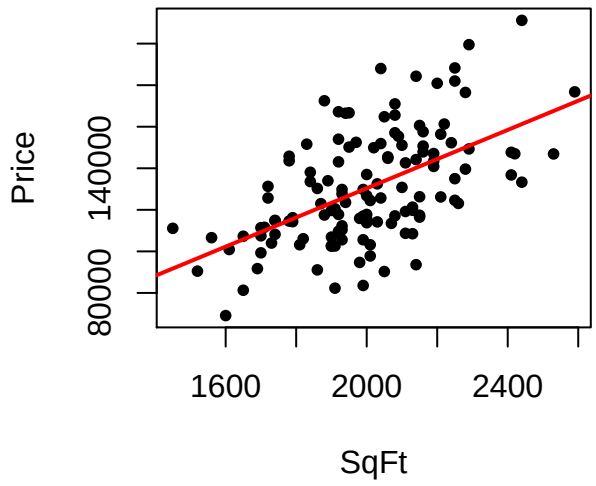
Retta di regressione fra prezzo e dimensione

```
mod0 <- lm( Price ~ SqFt, data = house )
mod0

##
## Call:
## lm(formula = Price ~ SqFt, data = house)
##
```

```
## Coefficients:
## (Intercept)         SqFt
##    -10091.1         70.2
```

Rappresentazione grafica

```
plot( Price ~ SqFt, pch = 20 )
abline( mod0, col = "red", lwd = 2 )
```

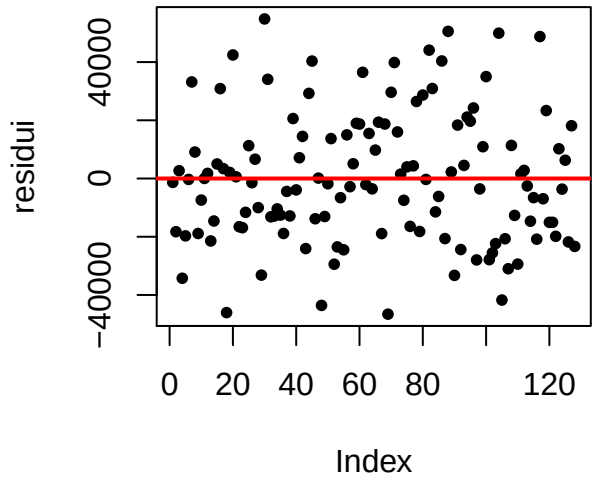



Residui

```
residui <- residuals(mod0)
summary(residui)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -46600  -16600   -1610       0   15100   54800
```

Grafico dei residui

```
plot( residui, pch = 20 )
abline( h = 0, col = "red", lwd = 2 )
```




Previsione del prezzo per un'abitazione di dimensione 2000 piedi quadri

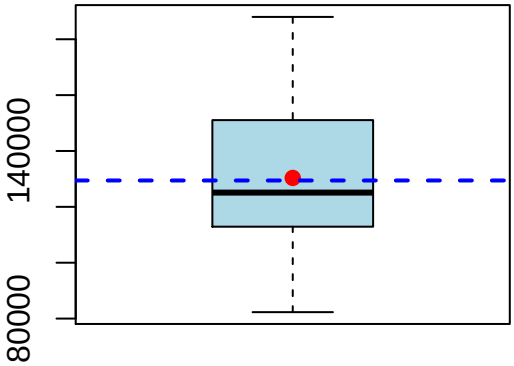
```
pred <- predict( mod0, newdata=data.frame( SqFt=2000 ) )
pred

##      1
## 130362
```

Prezzi delle case di 2000 piedi quadri

```
Price[ SqFt==2000 ]

## [1] 113800 137000 117800 126800 115700
```

Distribuzione del prezzo delle case con dimensione "vicina" a 2000 piedi quadri

```
boxplot( Price[ SqFt > 1900 & SqFt < 2100 ], col = "lightblue" )
points( 1, pred, col = "red", pch = 20, cex = 1.5 )
abline( h = mean( Price[ SqFt > 1900 & SqFt < 2100 ] ), col = "blue",
lwd = 2, lty = "dashed" )
```




## 3.3 Regressione lineare multivariata

Modelli di regressione multivariati

```
mod1 <- update( mod0, . ~ . + Bedrooms )
mod1

##
## Call:
## lm(formula = Price ~ SqFt + Bedrooms, data = house)
##
## Coefficients:
## (Intercept)         SqFt     Bedrooms
##     -6367.6         49.5      12486.1

mod2 <- update( mod0, . ~ . + Bathrooms )
mod2

##
## Call:
## lm(formula = Price ~ SqFt + Bathrooms, data = house)
##
## Coefficients:
## (Intercept)         SqFt    Bathrooms
##     -8437.6         48.8      16829.0
```

```
mod3 <- update( mod0, . ~ . + Offers )
mod3

##
## Call:
## lm(formula = Price ~ SqFt + Offers, data = house)
##
## Coefficients:
## (Intercept)         SqFt       Offers
##    -21841.7         94.4     -14170.8
```

### 3.3.1 Somme dei quadrati dei residui

Confronto fra i modelli in termini della varianza dei residui

```
mean( residuals(mod0)^2 )

## [1] 497256650

mean( residuals(mod1)^2 )

## [1] 434818652

mean( residuals(mod2)^2 )

## [1] 443200585

mean( residuals(mod3)^2 )

## [1] 295294654
```

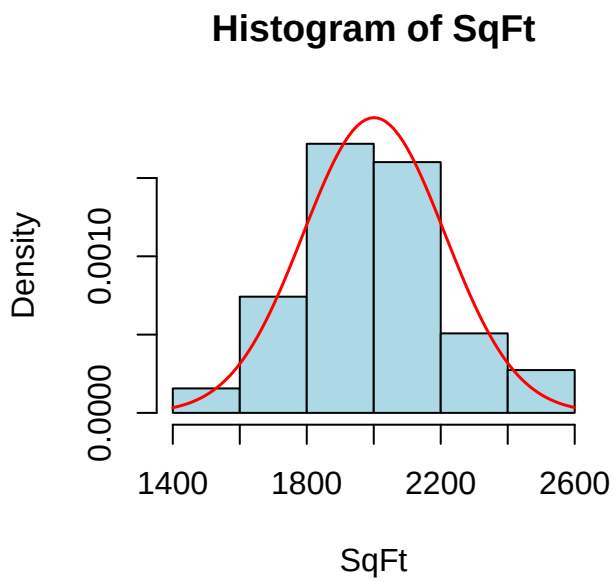
# 4 Variabilità campionaria

## 4.1 Simulazione dal modello

Controlliamo se la distrubuzione della dimensione degli appartamenti segue una legge normale

```
mu <- mean(SqFt)
mu

## [1] 2001
```

```
sigma2 <- var(SqFt)
sigma2

## [1] 44763

hist( SqFt, freq = FALSE, ylim = c(0, 1/sqrt( 2*pi*sigma2) ),
col = "lightblue" )
curve( dnorm(x, mean = mu, sd = sqrt(sigma2) ),
col = "red", lwd = 1.5, add = TRUE )
```




Simuliamo la dimensione di 1000 abitazioni dalla distribuzione normale "stimata", quindi simuliamo il loro prezzo dal modello di regressione calcolato precedentemente

```
set.seed(123)
N <- 1000
SqFt.sim <- rnorm( N, mean = mu, sd = sqrt( sigma2 ) )
errori <- rnorm( N, mean = 0, sd = sd( residui ) )
coef( mod0 )

## (Intercept)        SqFt
##   -10091.13       70.23

Price.sim <- coef( mod0 )[1] + coef( mod0 )[2] * SqFt.sim + errori
summary( Price.sim )
```

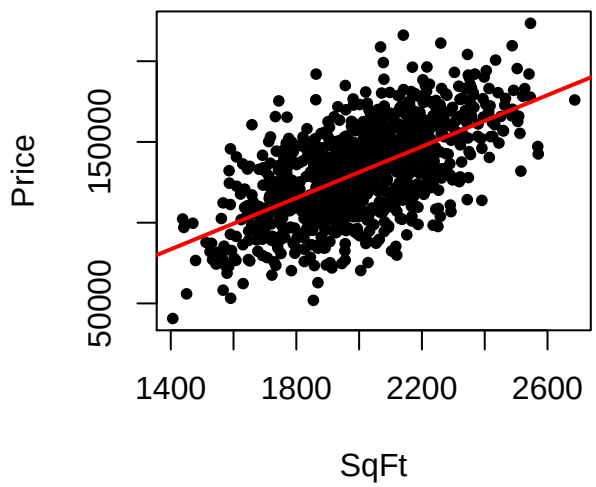
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40700  111000  133000  132000  150000  223000

data.sim <- data.frame( Price = Price.sim, SqFt = SqFt.sim )
```

Calcoliamo la stima ai minimi quadrati della retta di regressione usando i dati simulati

```
mod.sim <- lm( Price ~ SqFt, data = data.sim )
mod.sim

##
## Call:
## lm(formula = Price ~ SqFt, data = data.sim)
##
## Coefficients:
## (Intercept)         SqFt
##    -27813.9         79.5
```

Rappresentazione grfica

```
plot( Price ~ SqFt, data = data.sim, pch = 20 )
abline( mod.sim, col = "red", lwd = 2 )
```



## 4.2 Regressione con dati campionari

Estraiamo un campione di dimensione 150 dai dati simulati

```
campione <- sample( 1:N, 150 )
head( campione )

## [1] 305 832 593 805 293 141
```

Ristiamo il modello sui dati campionari

```
mod.camp <- lm( Price ~ SqFt, data = data.sim, subset = campione )
mod.camp

##
## Call:
## lm(formula = Price ~ SqFt, data = data.sim, subset = campione)
##
## Coefficients:
## (Intercept)         SqFt
##    -37570.9         84.7
```

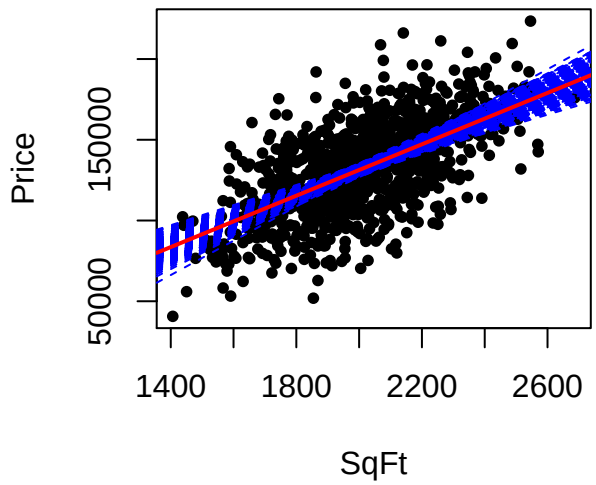
Ripetiamo l'operazione 1000 volte usando campioni di dimensione 200

```
stima.camp <- function( subset ){
coef( lm( Price ~ SqFt, data = data.sim, subset = subset ) )
}
coeff <- replicate( 1000, stima.camp( sample( 1:N, 200 ) ) )
summary( coeff[1,] )

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -82500  -36900  -27600  -27700  -18600   15000

summary( coeff[2,] )

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    57.9    75.1    79.4    79.5    84.2   106.0

summary( coeff[1,] - coef(mod.sim)[1] )

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -54700   -9050     250      95    9190   42800

summary( coeff[2,] - coef(mod.sim)[2] )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -21.600  -4.490  -0.139  -0.034   4.660  26.700

summary( ( coeff[1,] - coef(mod.sim)[1] ) / coef(mod.sim)[1] )

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.5400 -0.3300 -0.0090 -0.0034  0.3250  1.9700

summary( ( coeff[2,] - coef(mod.sim)[2] ) / coef(mod.sim)[2] )

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.2720 -0.0564 -0.0018 -0.0004  0.0586  0.3350
```

Rappresentazione grafica delle 1000 rette di regressione (in blu) rispetto alla 'vera' retta di regressione calcolata su tutti i dati (in rosso)

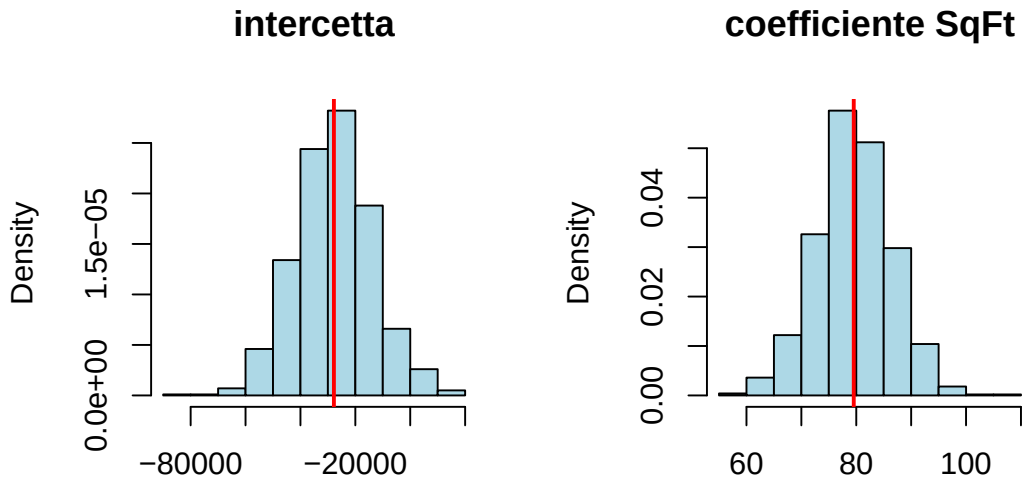
```
plot( Price ~ SqFt, data = data.sim, pch = 20 )
for( i in 1:1000 )
abline( a = coeff[1, i], b = coeff[2, i], col = "blue", lty = "dashed" )
abline( mod.sim, col = "red", lwd = 2 )
```




Istogrammi della distribuzione delle stime

```
par( mfrow = c(1, 2) )
hist( coeff[1, ], freq = FALSE, main = "intercetta", xlab = "",
```

```
col = "lightblue" )
abline( v = coef( mod.sim )[1], col = "red", lwd = 2 )
hist( coeff[2, ], freq = FALSE, main = "coefficiente SqFt",
xlab = "", col = "lightblue" )
abline( v = coef( mod.sim )[2], col = "red", lwd = 2 )
```




### 4.2.1 Precisione delle stime

Ripetiamo l'esercizio di simulazione variando la dimensione dei campioni: 100, 200, 400

```
coeff.array <- array( dim = c(3, 2, 1000 ) )
coeff.array[1,,] <- replicate( 1000, stima.camp( sample( 1:N, 100 ) ) )
coeff.array[2,,] <- replicate( 1000, stima.camp( sample( 1:N, 200 ) ) )
coeff.array[3,,] <- replicate( 1000, stima.camp( sample( 1:N, 400 ) ) )
```

Gli istogrammi delle stime mostrano chiare differenze in termini di precisione

```
par( mfrow=c(3, 2) )
for( i in 1:3 ){
hist( coeff.array[i, 1, ], freq = FALSE, main = "intercetta", xlab = "",
xlim = c( range( coeff.array[,1,] ) ), col = "lightblue" )
abline( v = coef( mod.sim )[1], col = "red", lwd = 2 )
hist( coeff.array[i, 2, ], freq = FALSE, main = "coefficiente SqFt",
xlab = "", xlim = c( range( coeff.array[,2,] ) ), col = "lightblue" )
```

```
abline( v = coef( mod.sim )[2], col = "red", lwd = 2 )
}
```