La retta di regressione

Statistica Applicata
Corso di Laurea in Informatica

cristiano.varin@unive.it

Indice

1 La stima ai minimi quadrati 1

2 L'effetto degli outlier 5

3 Stimatori robusti 7

1 La stima ai minimi quadrati

Lettura dati `HousePrices`[1]

```
## The following objects are masked from house (position 3):
##
##     Bathrooms, Bedrooms, Brick, HomeID, Neighborhood, Offers,
##     Price, SqFt
```

```
house <- read.csv(file = "HousePrices.csv")
attach(house)
```

Stima ai minimi quadrati della retta di regressione con risposta `Price` e predittore `SqFt`

```
mod <- lm( Price ~ SqFt )
mod

##
## Call:
```

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

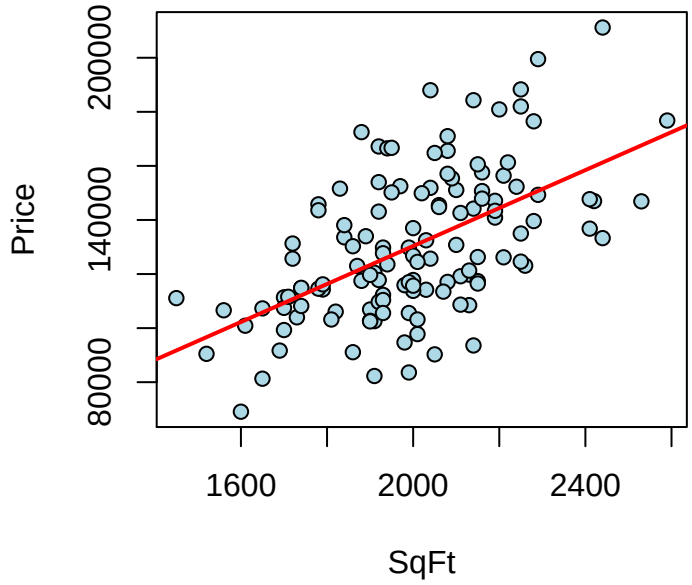
```
## lm(formula = Price ~ SqFt)
##
## Coefficients:
## (Intercept)         SqFt
##    -10091.1         70.2

summary(mod)

##
## Call:
## lm(formula = Price ~ SqFt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46593 -16644  -1610  15124  54829
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10091.13   18966.10   -0.53      0.6
## SqFt            70.23       9.43    7.45  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22500 on 126 degrees of freedom
## Multiple R-squared:  0.306,Adjusted R-squared:   0.3
## F-statistic: 55.5 on 1 and 126 DF,  p-value: 1.3e-11
```

Rappresentazione grafica della retta stimata ai minimi quadrati

```
plot( Price ~ SqFt, pch = 21, bg = "lightblue" )
abline( mod, col = "red", lwd = 2 )
```



Verifichiamo le stime

```
mod

##
## Call:
## lm(formula = Price ~ SqFt)
##
## Coefficients:
## (Intercept)         SqFt
##    -10091.1         70.2

beta <- cov(Price, SqFt) / var(SqFt)
beta

## [1] 70.23

alpha <- mean(Price) - beta * mean(SqFt)
alpha

## [1] -10091
```

Controlliamo i residui

```
res <- residuals(mod)
myres <- Price - predict(mod)
max( abs(res-myres) )

## [1] 6.803e-10

mean(res)

## [1] -1.17e-14

var(res)

## [1] 501172057

var(Price) - cov(Price, SqFt)^2 / var(SqFt)

## [1] 501172057

var(Price) * ( 1 - cor(Price, SqFt)^2 )

## [1] 501172057
```

Controlliamo l'R^2

```
summary(mod)

##
## Call:
## lm(formula = Price ~ SqFt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46593 -16644  -1610  15124  54829
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10091.13   18966.10   -0.53      0.6
## SqFt            70.23       9.43    7.45  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22500 on 126 degrees of freedom
## Multiple R-squared:  0.306,Adjusted R-squared:   0.3
## F-statistic: 55.5 on 1 and 126 DF,  p-value: 1.3e-11
```

```
sum.mod <- summary(mod)
names(sum.mod)

##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"

sum.mod$r.squared

## [1] 0.3058

R2 <- 1 - var(res) / var(Price)
R2

## [1] 0.3058

cor(Price, SqFt)^2

## [1] 0.3058
```

2 L'effetto degli outlier

Costruiamo un `data.frame` con le sole variabili `Price` e `SqFt`

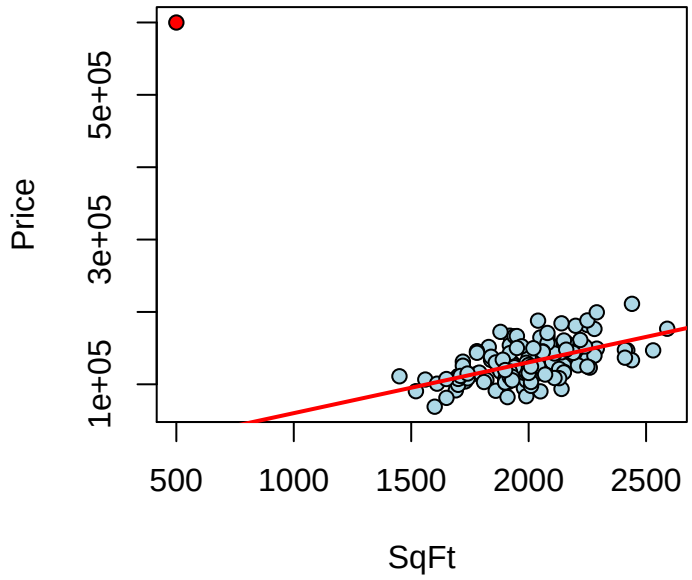
```
dati <- data.frame( Price, SqFt )
dim(dati)

## [1] 128   2
```

Consideriamo una nuova osservazione "anomala" rispetto ai dati osservati

```
newSqFt <- 500
newPrice <- 600000
```

```
plot( Price ~ SqFt, pch = 21, bg = "lightblue",
xlim = c( newSqFt, max(SqFt) ),
ylim = c( min(Price), newPrice ) )
abline( mod, col = "red", lwd = 2 )
points( newSqFt, newPrice, pch = 21, bg = "red" )
```



Aggiungiamo la nuova osservazione "anomala" ai dati osservati

```
dati <- rbind( dati, c(newPrice, newSqFt) )
dim(dati)

## [1] 129   2

tail(dati)

##      Price SqFt
## 124 119700 1900
## 125 147900 2160
## 126 113500 2070
## 127 149900 2020
## 128 124600 2250
## 129 600000  500
```

Calcoliamo il modello di regressione con i nuovi dati

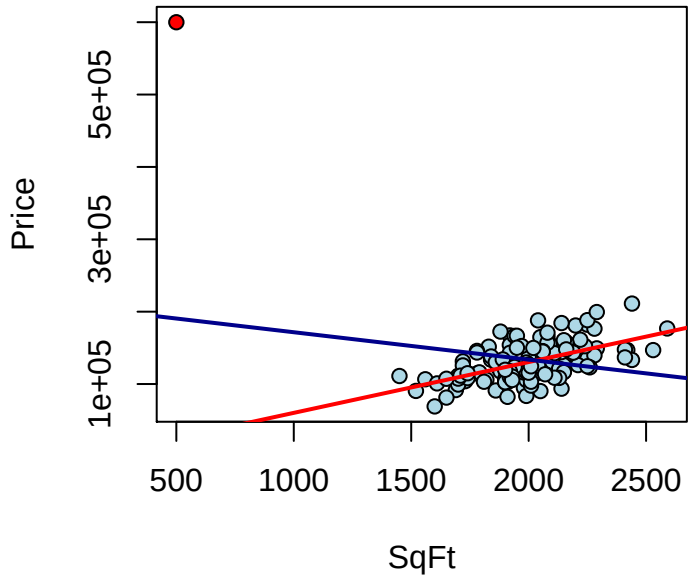
```
mod2 <- lm( Price ~ SqFt, data = dati )
mod2

##
```

```
## Call:
## lm(formula = Price ~ SqFt, data = dati)
##
## Coefficients:
## (Intercept)         SqFt
##    209444.4        -37.9
```

Aggiungiamo al grafico precedente la nuova retta ai minimi quadrati

```
abline( mod2, col = "darkblue", lwd = 2 )
```



3 Stimatori robusti

Abbiamo visto come la retta ai minimi quadrati non sia robusta alla presenza di 'outlier'. La funzione `rlm` ('robust linear model') implementa uno stimatore robusto di tipo M della retta di regressione

```
library(MASS) ## rlm appartiene a questa libreria
mod3 <- rlm( Price ~ SqFt, data = dati )
mod3
```

```
## Call:
## rlm(formula = Price ~ SqFt, data = dati)
## Converged in 6 iterations
##
## Coefficients:
## (Intercept)        SqFt
##    14435.82       57.67
##
## Degrees of freedom: 129 total; 127 residual
## Scale estimate: 23300
```

Aggiungiamo al grafico precedente la nuova retta ai minimi quadrati

```
abline( mod3, col = "black", lwd = 2 )
```