Il modello di regressione multivariato

Statistica Applicata
Corso di Laurea in Informatica

cristiano.varin@unive.it

Indice

1 Matrici

Definizione di un matrice

```
Z <- matrix(1:12, ncol = 3, nrow = 4)
Z

##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]    2    6   10
## [3,]    3    7   11
## [4,]    4    8   12

dim(Z)

## [1] 4 3

ncol(Z)

## [1] 3
```

```r
nrow(Z)

## [1] 4
```

Matrice trasposta

```r
t(Z)

##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    5    6    7    8
## [3,]    9   10   11   12
```

Prodotto matriciale

```r
Z %*% t(Z)

##      [,1] [,2] [,3] [,4]
## [1,]  107  122  137  152
## [2,]  122  140  158  176
## [3,]  137  158  179  200
## [4,]  152  176  200  224
```

Matrice inversa

```r
W <- matrix(c(4, 2, 7, 6), ncol = 2)
W

##      [,1] [,2]
## [1,]    4    7
## [2,]    2    6

solve(W)

##      [,1] [,2]
## [1,]  0.6 -0.7
## [2,] -0.2  0.4

W %*% solve(W)

##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1

solve(W) %*% W

##      [,1]      [,2]
## [1,]    1 8.882e-16
## [2,]    0 1.000e+00
```

Matrici diagonali

```
diag( c(2, 3, -1) )

##      [,1] [,2] [,3]
## [1,]    2    0    0
## [2,]    0    3    0
## [3,]    0    0   -1
```

Estrazione degli elementi sulla diagonale di un matrice

```
diag(W)

## [1] 4 6

diag(Z)

## [1]  1  6 11
```

2 Il modello di regressione multivariato

Lettura dati `HousePrices`[1]

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

Modello di regressione multivariato

```
mod <- lm( Price ~ SqFt + Bedrooms + Bathrooms + Offers,
data = house, x = TRUE, y = TRUE )
summary(mod)

##
## Call:
## lm(formula = Price ~ SqFt + Bedrooms + Bathrooms + Offers, data = house,
##     x = TRUE, y = TRUE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33608  -9889  -2968   9398  43243
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

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

```
## (Intercept) -17347.38   12724.90   -1.36     0.18
## SqFt            61.84       8.26    7.48  1.2e-11 ***
## Bedrooms      9319.75    2148.75    4.34  3.0e-05 ***
## Bathrooms    12646.35    3109.66    4.07  8.4e-05 ***
## Offers      -13601.01    1324.82  -10.27  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15000 on 123 degrees of freedom
## Multiple R-squared:  0.698,Adjusted R-squared:  0.688
## F-statistic: 71.1 on 4 and 123 DF,  p-value: <2e-16
```

Grazie agli argomenti `x=TRUE` e `y=TRUE` possiamo estrarre la matrice di regressione X e il vettore delle risposte Y

```
X <- mod$x
Y <- mod$y
dim(X)

## [1] 128   5

head(X)

##   (Intercept) SqFt Bedrooms Bathrooms Offers
## 1           1 1790        2         2      2
## 2           1 2030        4         2      3
## 3           1 1740        3         2      1
## 4           1 1980        3         2      3
## 5           1 2130        3         3      3
## 6           1 1780        3         2      2

head(Y)

##      1      2      3      4      5      6
## 114300 114200 114800  94700 119800 114600
```

Ricalcoliamo le stime ai minimi quadrati

```
beta.hat <- solve( t(X)%*% X ) %*% t(X) %*% Y
beta.hat

##                  [,1]
## (Intercept) -17347.38
## SqFt            61.84
## Bedrooms      9319.75
## Bathrooms    12646.35
## Offers      -13601.01
```

Calcoliamo la stima di σ^2

```
n <- nrow(X)
p <- ncol(X)
sigma2.hat <- sum( residuals(mod)^2 ) / (n-p)
sigma2.hat

## [1] 2.25e+08

sqrt(sigma2.hat)

## [1] 14999
```

e ora la varianza delle stime ai minimi quadrati

```
var.beta.hat <- solve( t(X)%*% X ) * sigma2.hat
var.beta.hat

##              (Intercept)      SqFt Bedrooms Bathrooms
## (Intercept)   161922986 -84798.99  1357864    618333
## SqFt             -84799     68.29    -6090    -10107
## Bedrooms        1357864  -6090.46  4617146  -1435485
## Bathrooms        618333 -10107.14 -1435485   9669998
## Offers          1510673  -3380.63   147134    116113
##               Offers
## (Intercept) 1510673
## SqFt          -3381
## Bedrooms     147134
## Bathrooms    116113
## Offers      1755144
```

da cui si ottengono gli standard errors

```
sqrt( diag(var.beta.hat) )

## (Intercept)        SqFt    Bedrooms   Bathrooms      Offers
##   12724.896       8.264    2148.754    3109.662    1324.819
```

Infine la statistica R^2

```
1 - sum( residuals(mod)^2 ) / sum( ( Y-mean(Y) )^2 )

## [1] 0.6982
```

e la sua versione aggiustata

```
var.res <- sum( residuals(mod)^2 ) / (n-p)
1 - var.res / var(Y)

## [1] 0.6884
```

3 Costruzione del modello

Matrice di correlazione

```
attach(house)

## The following objects are masked from house (position 4):
##
##     Bathrooms, Bedrooms, Brick, HomeID, Neighborhood,
##     Offers, Price, SqFt
## The following objects are masked from house (position 5):
##
##     Bathrooms, Bedrooms, Brick, HomeID, Neighborhood,
##     Offers, Price, SqFt

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
```

Modello di partenza

```
mod0 <- lm(Price ~ SqFt)
summary(mod0)

##
## 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
```

Residui

```
res0 <- residuals(mod0)
```

Correlazione fra residui e predittori

```
cor0 <- cor( cbind(res0, SqFt, Bedrooms, Bathrooms, Offers ) )
round(cor0, 2)

##             res0 SqFt Bedrooms Bathrooms Offers
## res0        1.00 0.00     0.31      0.28  -0.60
## SqFt        0.00 1.00     0.48      0.52   0.34
## Bedrooms    0.31 0.48     1.00      0.41   0.11
## Bathrooms   0.28 0.52     0.41      1.00   0.14
## Offers    -0.60 0.34     0.11      0.14   1.00
```

Aggiungiamo `Offers`

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

##
## Call:
## lm(formula = Price ~ SqFt + Offers)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -36185 -12885  -2874  10456  47057
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21841.69   14728.84   -1.48     0.14
## SqFt            94.36       7.75   12.18  < 2e-16 ***
## Offers      -14170.77    1532.61   -9.25  7.9e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17400 on 125 degrees of freedom
## Multiple R-squared:  0.588,Adjusted R-squared:  0.581
## F-statistic: 89.1 on 2 and 125 DF,  p-value: <2e-16
```

Abbiamo fatto bene?

```
summary( update(mod0, .~. + Bathrooms) )$r.squared

## [1] 0.3813

summary( update(mod0, .~. + Bedrooms) )$r.squared

## [1] 0.393
```

Un altro passo

```
res1 <- residuals(mod1)
cor1 <- cor( cbind(res1, SqFt, Bedrooms, Bathrooms, Offers ) )
round(cor1, 2)

##           res1 SqFt Bedrooms Bathrooms Offers
## res1      1.00 0.00     0.36      0.34   0.00
## SqFt      0.00 1.00     0.48      0.52   0.34
## Bedrooms  0.36 0.48     1.00      0.41   0.11
## Bathrooms 0.34 0.52     0.41      1.00   0.14
## Offers    0.00 0.34     0.11      0.14   1.00
```

Aggiungiamo `Bedrooms`

```
mod2 <- update(mod1, . ~ . + Bedrooms)
summary(mod2)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Offers + Bedrooms)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32804 -10973  -1091   7804  46160
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18156.03   13497.02   -1.35     0.18
## SqFt            75.06       8.06    9.31  5.8e-16 ***
## Offers      -13752.86    1404.82   -9.79  < 2e-16 ***
## Bedrooms     11197.07    2226.19    5.03  1.7e-06 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15900 on 124 degrees of freedom
## Multiple R-squared:  0.658,Adjusted R-squared:  0.649
## F-statistic: 79.4 on 3 and 124 DF,  p-value: <2e-16
```

e infine `Bathrooms`

```
mod3 <- update(mod2, . ~ . + Bathrooms)
summary(mod3)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Offers + Bedrooms + Bathrooms)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33608  -9889  -2968   9398  43243
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17347.38   12724.90   -1.36     0.18
## SqFt            61.84       8.26    7.48  1.2e-11 ***
## Offers      -13601.01    1324.82  -10.27  < 2e-16 ***
## Bedrooms      9319.75    2148.75    4.34  3.0e-05 ***
## Bathrooms    12646.35    3109.66    4.07  8.4e-05 ***
## ---
## Signif. codes:
```

```
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15000 on 123 degrees of freedom
## Multiple R-squared:  0.698,Adjusted R-squared:  0.688
## F-statistic: 71.1 on 4 and 123 DF,  p-value: <2e-16
```

4 Previsioni

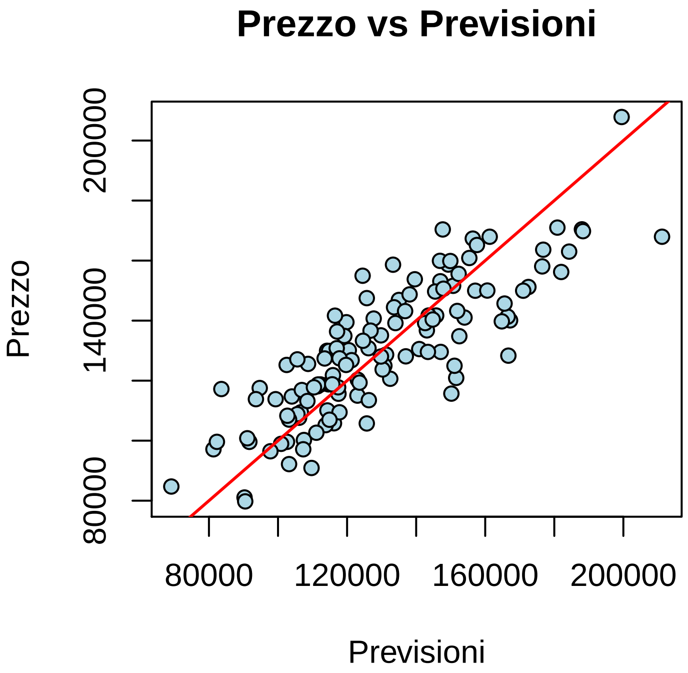
La funzione `predict` permette di estrarre i valori predetti dal modello

```
pred <- predict(mod3)
head(pred)
```

```
##      1      2      3      4      5      6
## 110076 129956 129905 117545 139467 118778
```

La qualità delle previsioni può essere visualizzata con un grafico a dispersione

```
plot( mod$y, pred, ylab = "Prezzo", xlab = "Previsioni", pch = 21,
bg = "lightblue", main = "Prezzo vs Previsioni" )
abline(a = 0, b = 1, col = "red", lwd = 1.5)
```



Previsione del prezzo di un immobile di 2000 piedi quadri con 2 stanze da letto, 2 bagni e che ha ricevuto un'offerta

```
predict( mod3, newdata = data.frame( SqFt = 2000, Bedrooms = 2,
Bathrooms = 2, Offers = 1) )

##      1
## 136664
```

Conviene costruire un'ulteriore stanza da letto?

```
predict( mod3, newdata = data.frame( SqFt = 2000, Bedrooms = 3,
Bathrooms = 2, Offers = 1) )

##      1
## 145983
```

5 Predittori categoriali

Aggiungiamo la variabile `Neighborhood`

```
mod4 <- update(mod3, . ~ . + Neighborhood, x = TRUE)
summary(mod4)

##
## Call:
## lm(formula = Price ~ SqFt + Offers + Bedrooms + Bathrooms + Neighborhood,
##     x = TRUE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31028  -9082   -688   9531  39126
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5335.8    12143.7    0.44  0.66117
## SqFt                    53.4        7.3    7.32  3.0e-11 ***
## Offers               -9026.4     1376.9   -6.56  1.4e-09 ***
## Bedrooms              3348.1     2030.5    1.65  0.10176
## Bathrooms            10443.3     2669.8    3.91  0.00015 ***
## NeighborhoodNorth    -2307.9     2999.3   -0.77  0.44312
## NeighborhoodWest     21597.5     3222.5    6.70  6.9e-10 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12800 on 121 degrees of freedom
## Multiple R-squared:  0.785,Adjusted R-squared:  0.775
## F-statistic: 73.7 on 6 and 121 DF,  p-value: <2e-16
```

Come è stata codificata la variabile `Neighborhood`?

```
head(mod4$x)
```

```
##   (Intercept) SqFt Offers Bedrooms Bathrooms
## 1           1 1790      2        2         2
## 2           1 2030      3        4         2
## 3           1 1740      1        3         2
## 4           1 1980      3        3         2
## 5           1 2130      3        3         3
## 6           1 1780      2        3         2
##   NeighborhoodNorth NeighborhoodWest
## 1                 0                0
## 2                 0                0
## 3                 0                0
## 4                 0                0
## 5                 0                0
## 6                 1                0
```

```
summary(mod4$x)
```

```
##   (Intercept)      SqFt          Offers        Bedrooms
##  Min.   :1    Min.   :1450   Min.   :1.00   Min.   :2.00
##  1st Qu.:1    1st Qu.:1880   1st Qu.:2.00   1st Qu.:3.00
##  Median :1    Median :2000   Median :3.00   Median :3.00
##  Mean   :1    Mean   :2001   Mean   :2.58   Mean   :3.02
##  3rd Qu.:1    3rd Qu.:2140   3rd Qu.:3.00   3rd Qu.:3.00
##  Max.   :1    Max.   :2590   Max.   :6.00   Max.   :5.00
##    Bathrooms    NeighborhoodNorth NeighborhoodWest
##  Min.   :2.00   Min.   :0.000     Min.   :0.000
##  1st Qu.:2.00   1st Qu.:0.000     1st Qu.:0.000
##  Median :2.00   Median :0.000     Median :0.000
##  Mean   :2.44   Mean   :0.344     Mean   :0.305
##  3rd Qu.:3.00   3rd Qu.:1.000     3rd Qu.:1.000
##  Max.   :4.00   Max.   :1.000     Max.   :1.000
```

Infine, la variabile `Brick`

```
mod5 <- update(mod4, . ˜ . + Brick, x = TRUE)
summary(mod5)
```

```
##
## Call:
## lm(formula = Price ˜ SqFt + Offers + Bedrooms + Bathrooms + Neighborhood +
##     Brick, x = TRUE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27337  -6549    -42   5803  27359
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          598.92    9552.20    0.06   0.9501
## SqFt                  52.99       5.73    9.24  1.1e-15 ***
## Offers             -8267.49    1084.78   -7.62  6.5e-12 ***
## Bedrooms            4246.79    1597.91    2.66   0.0089 **
## Bathrooms           7883.28    2117.04    3.72   0.0003 ***
## NeighborhoodNorth   1560.58    2396.77    0.65   0.5162
## NeighborhoodWest   22241.62    2531.76    8.79  1.3e-14 ***
## BrickYes           17297.35    1981.62    8.73  1.8e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10000 on 120 degrees of freedom
## Multiple R-squared:  0.869,Adjusted R-squared:  0.861
## F-statistic:  113 on 7 and 120 DF,  p-value: <2e-16
```

```
summary(mod5$x)
```

```
##   (Intercept)      SqFt          Offers         Bedrooms
##  Min.   :1    Min.   :1450   Min.   :1.00   Min.   :2.00
##  1st Qu.:1    1st Qu.:1880   1st Qu.:2.00   1st Qu.:3.00
##  Median :1    Median :2000   Median :3.00   Median :3.00
##  Mean   :1    Mean   :2001   Mean   :2.58   Mean   :3.02
##  3rd Qu.:1    3rd Qu.:2140   3rd Qu.:3.00   3rd Qu.:3.00
##  Max.   :1    Max.   :2590   Max.   :6.00   Max.   :5.00
##    Bathrooms    NeighborhoodNorth NeighborhoodWest
##  Min.   :2.00   Min.   :0.000     Min.   :0.000
##  1st Qu.:2.00   1st Qu.:0.000     1st Qu.:0.000
##  Median :2.00   Median :0.000     Median :0.000
##  Mean   :2.44   Mean   :0.344     Mean   :0.305
##  3rd Qu.:3.00   3rd Qu.:1.000     3rd Qu.:1.000
##  Max.   :4.00   Max.   :1.000     Max.   :1.000
##     BrickYes
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.328
##  3rd Qu.:1.000
##  Max.   :1.000
```