Metodi statistici per la ricerca sociale

ANOVA e ANCOVA in 

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM–88 SOCIOLOGIA E RICERCA SOCIALE

Indice di massa corporea

- Campione casuale di $n = 100$ uomini adulti
- Variabili

bmi	=	Bodi mass index (indice di massa corporea)
af	=	Attività fisica (1 = per niente, 2 = 1 volta alla settimana, 3 = almeno 2 volte alla settimana)
dieta	=	Seguire una dieta equilibrata (0 = No; 1 = Si)
age	=	Età (in anni)

Indice di massa corporea: Dati

```
> rm(list=ls())
>
> load("Dati/BMI.RData")
> ls()
[1] "BMI"
> dim(BMI)
[1] 100   4
>
> head(BMI)
    bmi af dieta age
1 29.70  1     0  46
2 31.46  1     0  57
3 26.84  1     1  53
4 31.26  1     1  61
5 33.68  1     0  54
6 32.06  1     0  61
>
> attach(BMI)
```

Indice di massa corporea: ANOVA a un fattore

```
> ##ANALISI DELLA VARIANZA A UN FATTORE
> ##Variabile risposta = bmi
> ##Variabile esplicativa (fattore) = Attività fisica
>
> ##Dimensioni di gruppo
> table(af)
af
1  2  3
25 35 40
> by(bmi, af, length)
af: 1
[1] 25
------------------------------------------------------------
af: 2
[1] 35
------------------------------------------------------------
af: 3
[1] 40
```

Indice di massa corporea: ANOVA a un fattore

```
> ##Medie di gruppo e generale
> by(bmi, af,mean)
af: 1
[1] 30.872
------------------------------------------------------------
af: 2
[1] 26.43914
------------------------------------------------------------
af: 3
[1] 25.70425
>
>##Media generale
> mean(bmi)
[1] 27.2534
```

Indice di massa corporea: ANOVA a un fattore

```
> ##varianza e devianza totale
> var(bmi)
[1] 10.45611
> var(bmi)*{length(bmi)-1}
[1] 1035.155
>
> ##Varianze entro i  gruppi
> by(bmi, af, var)
af: 1
[1] 4.147658
------------------------------------------------------------
af: 2
[1] 9.262996
------------------------------------------------------------
af: 3
[1] 4.464415
```

Indice di massa corporea: ANOVA a un fattore

```
##Scomposizione della varianza
> aov(bmi~as.factor(af), data=BMI)
Call:
aov(formula = bmi ~ as.factor(af), data = BMI)

Terms:
as.factor(af) Residuals
Sum of Squares       446.5568  588.5979
Deg. of Freedom             2        97

Residual standard error: 2.463335
Estimated effects may be unbalanced
>
> summary(aov(bmi~as.factor(af), data=BMI))
              Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(af)  2  446.6  223.28    36.8 1.28e-12 ***
Residuals     97  588.6    6.07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Indice di massa corporea: ANOVA a un fattore

```
> ##Modello di regressione
> m1 <- lm(bmi~as.factor(af), data=BMI)
> summary(m1)

Call:
lm(formula = bmi ~ as.factor(af), data = BMI)

Residuals:
Min      1Q  Median      3Q     Max
-6.0491 -1.6976  0.0557  1.3270  9.0509

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     30.8720     0.4927  62.663  < 2e-16 ***
as.factor(af)2  -4.4329     0.6451  -6.872 6.11e-10 ***
as.factor(af)3  -5.1677     0.6280  -8.229 8.75e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.463 on 97 degrees of freedom
Multiple R-squared:  0.4314,Adjusted R-squared:  0.4197
F-statistic:  36.8 on 2 and 97 DF,  p-value: 1.284e-12
```

Indice di massa corporea: ANOVA a un fattore

```
> MX<-model.matrix(m1)
> MX[1:5,]
  (Intercept) as.factor(af)2 as.factor(af)3
1           1               0               0
2           1               0               0
3           1               0               0
4           1               0               0
5           1               0               0
```

Indice di massa corporea: ANOVA a un fattore

```
> yhat<- fitted(m1)
> table(af)
af
1  2  3
25 35 40
> table(yhat[af==1])

30.872
25
> table(yhat[af==2])

26.4391428571429
35
> table(yhat[af==3])

25.70425
40
```

Indice di massa corporea: ANOVA a un fattore

```
> by(bmi, af,mean)
af: 1
[1] 30.872
------------------------------------------------------------
af: 2
[1] 26.43914
------------------------------------------------------------
af: 3
[1] 25.70425
>
> new<- data.frame(af=c(1,2,3))
> predict(m1, new)
       1        2        3
30.87200 26.43914 25.70425
```

Indice di massa corporea: ANOVA a un fattore

```
> ##Auto-costruzione delle variabili binarie
> A1<- as.numeric(af==1)
> A2<- as.numeric(af==2)
> A3<- as.numeric(af==3)
>
> lm(bmi~A2+A3)

Call:
lm(formula = bmi ~ A2 + A3)

Coefficients:
(Intercept)          A2          A3
30.872       -4.433       -5.168

> m1

Call:
lm(formula = bmi ~ as.factor(af), data = BMI)

Coefficients:
(Intercept)  as.factor(af)2  as.factor(af)3
30.872          -4.433          -5.168
```

Indice di massa corporea: ANOVA a un fattore

```
> ##Cambiare il valore di riferimento
> lm(bmi~A1+A2)

Call:
lm(formula = bmi ~ A1 + A2)

Coefficients:
(Intercept)        A1            A2
25.7043       5.1678       0.7349

> af.factor <- as.factor(af)
> af.factor <- relevel(af.factor, ref="3")
> lm(bmi~af.factor)

Call:
lm(formula = bmi ~ af.factor)

Coefficients:
(Intercept)   af.factor1   af.factor2
25.7043       5.1678       0.7349
```

Indice di massa corporea: ANOVA a un fattore

```
> ##Modello a un fattore versus modello nullo
> m0<-lm(bmi~1, data=BMI)
> m0

Call:
lm(formula = bmi ~ 1, data = BMI)

Coefficients:
(Intercept)
27.25
>
> mean(bmi)
[1] 27.2534
>
> anova(m0, m1)
Analysis of Variance Table

Model 1: bmi ~ 1
Model 2: bmi ~ as.factor(af)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     99 1035.2
2     97  588.6  2    446.56 36.796 1.284e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Indice di massa corporea: ANOVA a due fattori

```
> ##ANOVA a due fattori senza interazione
> table(af, dieta)
dieta
af 0  1
1 12 13
2 18 17
3 29 11
> by(bmi, list(af=af, dieta=dieta),mean)
af: 1
dieta: 0
[1] 31.72
------------------------------------------------------------
af: 2
dieta: 0
[1] 27.93222
------------------------------------------------------------
af: 3
dieta: 0
[1] 26.11966
------------------------------------------------------------
af: 1
dieta: 1
[1] 30.08923
------------------------------------------------------------
af: 2
dieta: 1
[1] 24.85824
------------------------------------------------------------
af: 3
dieta: 1
[1] 24.60909
```

Indice di massa corporea: ANOVA a due fattori

```
> table(af)
af
1  2  3
25 35 40
> by(bmi, af,mean)
af: 1
[1] 30.872
------------------------------------------------------------
af: 2
[1] 26.43914
------------------------------------------------------------
af: 3
[1] 25.70425
>
> table(dieta)
dieta
0  1
59 41
> by(bmi, dieta, mean)
dieta: 0
[1] 27.81169
------------------------------------------------------------
dieta: 1
[1] 26.45
```

Indice di massa corporea: ANOVA a due fattori

```
> m2 <- lm(bmi~as.factor(af)+as.factor(dieta), data=BMI)
> summary(m2)

Call:
lm(formula = bmi ~ as.factor(af) + as.factor(dieta), data = BMI)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9493 -1.4167 -0.0117  1.6356  8.0121

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        31.9841     0.5107  62.626  < 2e-16 ***
as.factor(af)2     -4.5062     0.5879  -7.664 1.45e-11 ***
as.factor(af)3     -5.6917     0.5836  -9.753 5.06e-16 ***
as.factor(dieta)1  -2.1386     0.4684  -4.566 1.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.244 on 96 degrees of freedom
Multiple R-squared:  0.5328,Adjusted R-squared:  0.5182
F-statistic:  36.5 on 3 and 96 DF,  p-value: 7.915e-16
```

Indice di massa corporea: ANOVA a due fattori

```
> model.matrix(m2)[1:5,]
  (Intercept) as.factor(af)2 as.factor(af)3 as.factor(dieta)1
1           1              0              0                 0
2           1              0              0                 0
3           1              0              0                 1
4           1              0              0                 1
5           1              0              0                 0
```

Indice di massa corporea: ANOVA a due fattori

```
> ##Modello nullo versus Modello con 2 fattori senza interazione
> anova(m0,m2)
Analysis of Variance Table

Model 1: bmi ~ 1
Model 2: bmi ~ as.factor(af) + as.factor(dieta)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     99 1035.2
2     96  483.6  3    551.56 36.497 7.915e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ##Modello con il fattore af versus Modello con 2 fattori senza interazione
> anova(m1,m2)
Analysis of Variance Table

Model 1: bmi ~ as.factor(af)
Model 2: bmi ~ as.factor(af) + as.factor(dieta)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     97 588.6
2     96 483.6  1       105 20.844 1.477e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Indice di massa corporea: ANOVA a due fattori

```
> m1b <-lm(bmi~as.factor(dieta), data=BMI)
> m1b

Call:
lm(formula = bmi ~ as.factor(dieta), data = BMI)

Coefficients:
(Intercept)  as.factor(dieta)1
27.812              -1.362
>
> anova(m1b,m2)
Analysis of Variance Table

Model 1: bmi ~ as.factor(dieta)
Model 2: bmi ~ as.factor(af) + as.factor(dieta)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     98 990.3
2     96 483.6  2    506.71 50.294 1.144e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Indice di massa corporea: ANOVA a due fattori

```
> ####ANOVA con due fattori con interazione
> m3 <- lm(bmi~as.factor(af)+as.factor(dieta) +
+ as.factor(af):as.factor(dieta), data=BMI)
> summary(m3)

Call:
lm(formula = bmi ~ as.factor(af) + as.factor(dieta) +
   as.factor(af):as.factor(dieta), data = BMI)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4682 -1.5341  0.0798  1.5710  7.5578

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      31.7200     0.6463  49.078  < 2e-16 ***
as.factor(af)2                   -3.7878     0.8344  -4.540 1.67e-05 ***
as.factor(af)3                   -5.6003     0.7685  -7.287 9.68e-11 ***
as.factor(dieta)1                -1.6308     0.8963  -1.819    0.072 .
as.factor(af)2:as.factor(dieta)1 -1.4432     1.1733  -1.230    0.222
as.factor(af)3:as.factor(dieta)1  0.1202     1.1966   0.100    0.920
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.239 on 94 degrees of freedom
Multiple R-squared:  0.5448,Adjusted R-squared:  0.5206
F-statistic:  22.5 on 5 and 94 DF,  p-value: 8.959e-15
```

Indice di massa corporea: ANOVA a due fattori

```
> new<- data.frame(af=c(1,1,2,2,3,3), dieta=c(0,1,0,1,0,1))
> predict(m3, new)
1        2        3        4        5        6
31.72000 30.08923 27.93222 24.85824 26.11966 24.60909
>
> by(bmi, list(af=af, dieta=dieta),mean)
af: 1
dieta: 0
[1] 31.72
------------------------------------------------------------
af: 2
dieta: 0
[1] 27.93222
------------------------------------------------------------
af: 3
dieta: 0
[1] 26.11966
------------------------------------------------------------
af: 1
dieta: 1
[1] 30.08923
------------------------------------------------------------
af: 2
dieta: 1
[1] 24.85824
------------------------------------------------------------
af: 3
dieta: 1
[1] 24.60909
```

Indice di massa corporea: ANOVA a due fattori

```
> anova(m2,m3)
Analysis of Variance Table

Model 1: bmi ~ as.factor(af) + as.factor(dieta)
Model 2: bmi ~ as.factor(af) + as.factor(dieta) + as.factor(af):as.factor(dieta)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     96 483.60
2     94 471.19  2    12.404 1.2373 0.2948
```

Indice di massa corporea: ANCOVA

```
> m1.add <-  lm(bmi ~as.factor(af)+ age, data=BMI)
> summary(m1.add)

Call:
lm(formula = bmi ~ as.factor(af) + age, data = BMI)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9193 -1.2163 -0.0611  1.4783  8.9449

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    19.18969    2.52814   7.590 2.07e-11 ***
as.factor(af)2 -2.11899    0.76475  -2.771  0.00671 **
as.factor(af)3 -1.93514    0.89344  -2.166  0.03279 *
age             0.20596    0.04387   4.695 8.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.233 on 96 degrees of freedom
Multiple R-squared:  0.5376,Adjusted R-squared:  0.5231
F-statistic:  37.2 on 3 and 96 DF,  p-value: 4.875e-16
```

Indice di massa corporea: ANCOVA

```
> m1.int<-lm(bmi ~ as.factor(af)*age, data=BMI)
> summary(m1.int)

Call:
lm(formula = bmi ~ as.factor(af) * age, data = BMI)

Residuals:
Min      1Q  Median      3Q     Max
-5.6002 -1.1824  0.1209  1.2416  8.8662

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        19.22590    4.23184   4.543 1.64e-05 ***
as.factor(af)2     -9.12046    5.53855  -1.647  0.10295
as.factor(af)3      3.12541    5.13832   0.608  0.54448
age                 0.20533    0.07421   2.767  0.00682 **
as.factor(af)2:age  0.15377    0.10776   1.427  0.15692
as.factor(af)3:age -0.12360    0.10239  -1.207  0.23042
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.178 on 94 degrees of freedom
Multiple R-squared:  0.5694,Adjusted R-squared:  0.5465
F-statistic: 24.86 on 5 and 94 DF,  p-value: 7.046e-16
```

Indice di massa corporea: ANCOVA

```
> mean(age)
[1] 46.51
> BMI$age.c<- age-mean(age)
> m1.int.c<-lm(bmi ~ as.factor(af)*age.c, data=BMI)
> summary(m1.int.c)

Call:
lm(formula = bmi ~ as.factor(af) * age.c, data = BMI)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6002 -1.1824  0.1209  1.2416  8.8662

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          28.77562    0.87397  32.925  < 2e-16 ***
as.factor(af)2       -1.96866    0.95169  -2.069  0.04133 *
as.factor(af)3       -2.62309    1.01592  -2.582  0.01137 *
age.c                 0.20533    0.07421   2.767  0.00682 **
as.factor(af)2:age.c  0.15377    0.10776   1.427  0.15692
as.factor(af)3:age.c -0.12360    0.10239  -1.207  0.23042
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.178 on 94 degrees of freedom
Multiple R-squared:  0.5694,Adjusted R-squared:  0.5465
F-statistic: 24.86 on 5 and 94 DF,  p-value: 7.046e-16
```

Indice di massa corporea: ANCOVA

```
> anova(m1.add, m1.int)
Analysis of Variance Table

Model 1: bmi ~ as.factor(af) + age
Model 2: bmi ~ as.factor(af) * age
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     96 478.69
2     94 445.77  2    32.923 3.4712 0.03512 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Indice di massa corporea: ANCOVA

```
> #Un modello di regressione per il body mass index...
> mm <- lm(bmi ~  as.factor(af) + as.factor(dieta) + age.c, data=BMI)
> summary(mm)

Call:
lm(formula = bmi ~ as.factor(af) + as.factor(dieta) + age.c,
data = BMI)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0033 -1.3896  0.0036  1.1453  7.8960

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       29.87318    0.59890  49.880  < 2e-16 ***
as.factor(af)2    -2.17325    0.67761  -3.207  0.00183 **
as.factor(af)3    -2.43622    0.79733  -3.055  0.00292 **
as.factor(dieta)1 -2.15769    0.41291  -5.226 1.03e-06 ***
age                0.20772    0.03887   5.344 6.22e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.978 on 95 degrees of freedom
Multiple R-squared:  0.6408,Adjusted R-squared:  0.6257
F-statistic: 42.37 on 4 and 95 DF,  p-value: < 2.2e-16
```