Test e intervalli di confidenza nel modello di regressione lineare

Statistica Applicata
Corso di Laurea in Informatica

cristiano.varin@unive.it

Indice

1 Retta di regressione

In questa lezione utilizzeremo dati e alcune funzioni del pacchetto `car`

```
library(car)
```

Iniziamo con il dataset `Davis` che riguarda uno studio sulla concordanza fra altezza e peso misurati e riportati da un gruppo di persone

```
data(Davis)
head(Davis)

##   sex weight height repwt repht
## 1   M     77    182    77   180
## 2   F     58    161    51   159
## 3   F     53    161    54   158
## 4   M     68    177    70   175
## 5   F     59    157    59   155
## 6   M     76    170    76   165
```

Maggiori informazioni sono disponibili nell'help in linea

```
help(Davis)
```

Retta di regressione fra il peso misurato e quello riportato

```
mod <- lm(weight~repwt, data=Davis)
summary(mod)

##
## Call:
## lm(formula = weight ~ repwt, data = Davis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7.05  -1.87  -0.73   0.60 108.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.3363     3.0369    1.76    0.081 .
## repwt         0.9278     0.0453   20.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.42 on 181 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.699,Adjusted R-squared:  0.697
## F-statistic:  420 on 1 and 181 DF,  p-value: <2e-16
```

Verifichiamo il calcolo dei test di significatività dei parametri di regressione

```
vcov(mod)

##             (Intercept)     repwt
## (Intercept)      9.2228 -0.134641
## repwt           -0.1346  0.002052

se <- sqrt( diag(vcov(mod)) )
se

## (Intercept)       repwt
##      3.0369      0.0453

tval <- coef(mod)/se
tval
```

```
## (Intercept)       repwt
##       1.757      20.484

N <- nobs(mod)
N

## [1] 183

p <- 2
pval <- 2*(  1-pt( abs(tval), df=N-p ) )
pval

## (Intercept)       repwt
##     0.08059     0.00000
```

Intervalli di confidenza

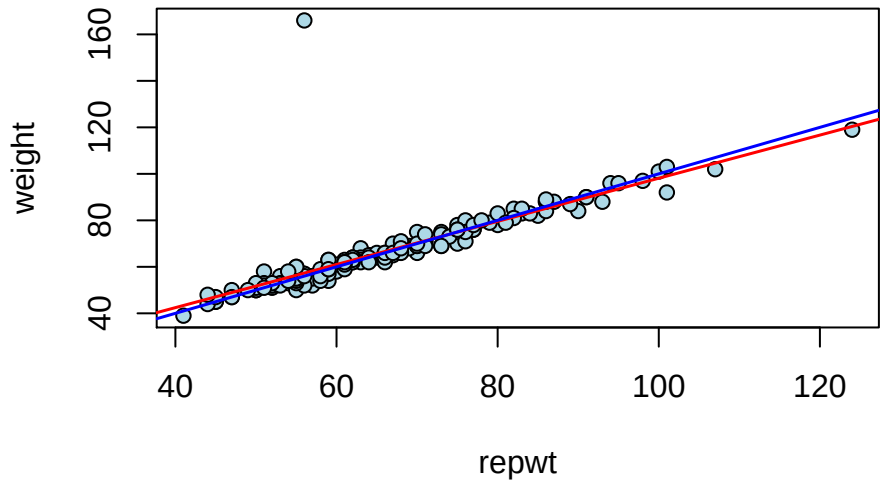
```
confint(mod)

##               2.5 % 97.5 %
## (Intercept) -0.6560 11.329
## repwt        0.8385  1.017

qt(0.025, df=N-p)

## [1] -1.973

qt(0.975, df=N-p)

## [1] 1.973

coef(mod) + qt(0.025, df=N-p) * se

## (Intercept)       repwt
##     -0.6560      0.8385

coef(mod) + qt(0.975, df=N-p) * se

## (Intercept)       repwt
##      11.329       1.017

confint(mod, level=0.9)

##                5 %   95 %
## (Intercept) 0.3153 10.357
## repwt       0.8530  1.003
```

```r
confint(mod, level=0.99)
```

```
##                0.5 % 99.5 %
## (Intercept) -2.5696 13.242
## repwt        0.8099  1.046
```

In realtà l'analisi svolta è imprecisa... guardiamo i dati

```r
plot( weight~repwt, data=Davis, pch=21, bg="lightblue" )
abline(mod, col="red", lwd=1.5)
abline(0, 1, col="blue", lwd=1.5)
```



Si nota un chiaro outlier

```r
outlier <- which.max(Davis$weight)
outlier
```

```
## [1] 12
```

```r
mod2 <- lm(weight~repwt, data=Davis, subset=-outlier)
summary(mod2)
```

```
##
## Call:
```

```
## lm(formula = weight ~ repwt, data = Davis, subset = -outlier)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.530 -1.101 -0.132  1.129  6.389
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7338     0.8148    3.36  0.00097 ***
## repwt         0.9584     0.0121   78.93  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.25 on 180 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.972,Adjusted R-squared:  0.972
## F-statistic: 6.23e+03 on 1 and 180 DF,  p-value: <2e-16
```

```
confint(mod2)
```

```
##              2.5 % 97.5 %
## (Intercept) 1.1260 4.3416
## repwt       0.9344 0.9823
```

2 Modelli di regressione multivariati

Consideriamo ora il dataset `Duncan` relativo ad uno studio sul prestigio di diverse occupazioni

```
data(Duncan)
head(Duncan)
```

```
##            type income education prestige
## accountant prof     62        86       82
## pilot      prof     72        76       83
## architect  prof     75        92       90
## author     prof     55        90       76
## chemist    prof     64        86       90
## minister   prof     21        84       87
```

I dati sono descritti nell'help in linea

```
help(Duncan)
```

Consideriamo il modello lineare in cui il prestigio è predetto dal salario e dall'educazione

```
mod <- lm( prestige~income+education, data=Duncan )
summary(mod)

##
## Call:
## lm(formula = prestige ~ income + education, data = Duncan)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29.54  -6.42   0.65   6.61  34.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.0647     4.2719   -1.42     0.16
## income        0.5987     0.1197    5.00  1.1e-05 ***
## education     0.5458     0.0983    5.56  1.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 42 degrees of freedom
## Multiple R-squared:  0.828,Adjusted R-squared:  0.82
## F-statistic:  101 on 2 and 42 DF,  p-value: <2e-16
```

Tabella dell'analisi della devianza del modello

```
anova(mod)

## Analysis of Variance Table
##
## Response: prestige
##           Df Sum Sq Mean Sq F value  Pr(>F)
## income     1  30665   30665   171.6 < 2e-16 ***
## education  1   5516    5516    30.9 1.7e-06 ***
## Residuals 42   7507     179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notate che la tabella tiene conto dell'ordine con cui entrano i predittori

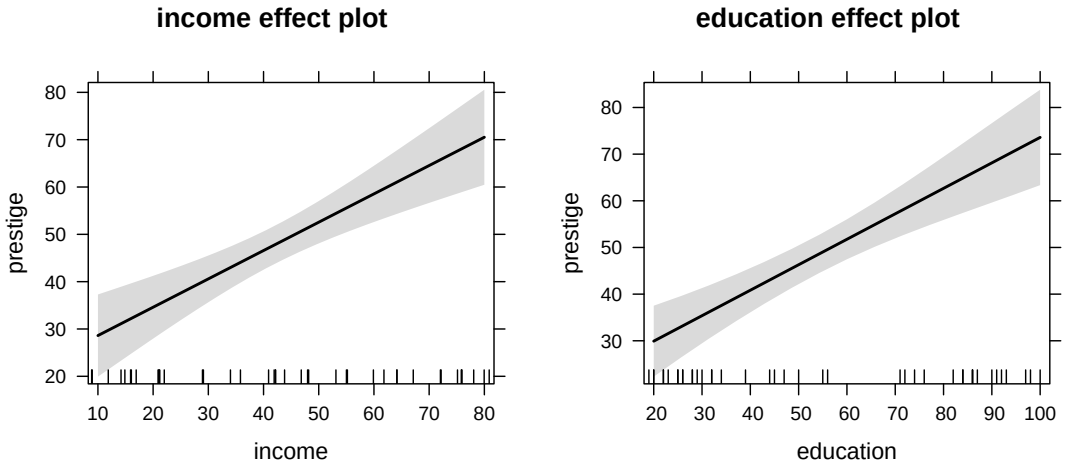
```r
mod2 <- lm( prestige~education+income, data=Duncan )
anova(mod2)

## Analysis of Variance Table
##
## Response: prestige
##           Df Sum Sq Mean Sq F value  Pr(>F)
## education  1  31707   31707     177 < 2e-16 ***
## income     1   4474    4474      25 1.1e-05 ***
## Residuals 42   7507     179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Intervalli di confidenza

```r
confint(mod)

##                2.5 % 97.5 %
## (Intercept) -14.6858 2.5565
## income        0.3572 0.8402
## education     0.3476 0.7441
```

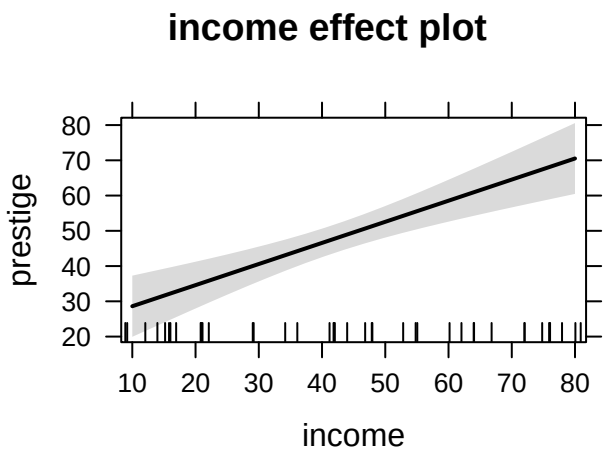
Gli effetti marginali dei due predittori possono essere visualizzati con un grafico 'effetto' in cui viene mostrato l'intervallo di confidenza di un predittore mentre gli altri predittori del modello sono posti pari al loro valore medio

```r
library(effects)
plot( allEffects(mod) )
```



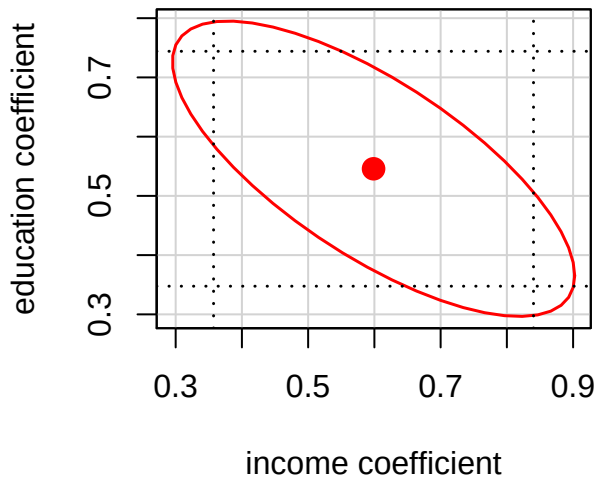
Oppure se siamo interessati solo ad un predittore

```
plot( effect("income", mod) )
```



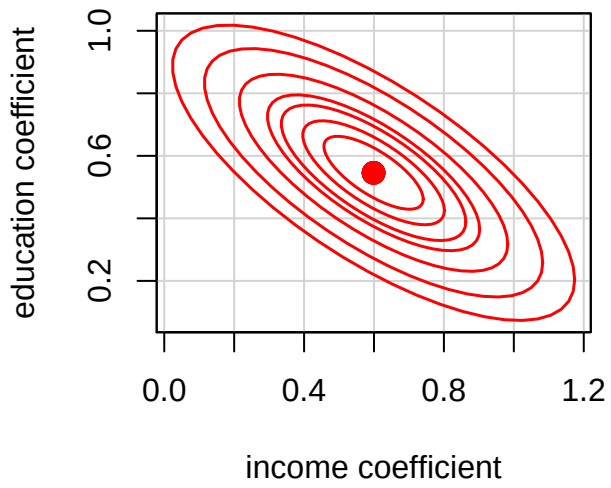
Possiamo anche visualizzare l'intervallo di confidenza simultaneo di `income` e `education`

```
## la funzione confidenceEllipse e' contenuta in car
ci <- confint(mod)
confidenceEllipse(mod, lwd=1.5)
for(i in 1:2) abline(v=ci[2, i], lty="dotted", lwd=1.5)
for(i in 1:2) abline(h=ci[3, i], lty="dotted", lwd=1.5)
```



Variando il livello di confidenza

```
confidenceEllipse(mod, lwd=1.5, levels=c(0.5, 0.75, 0.9, 0.95, 0.99,
0.999, 0.9999))
```



Infine, proviamo ad inserire anche un predittore categoriale (`type`)

```
mod2 <- update(mod, .~.+type)
summary(mod2)

##
## Call:
## lm(formula = prestige ~ income + education + type, data = Duncan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.889  -5.744  -1.755   5.440  28.971
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1850     3.7138   -0.05   0.9605
## income        0.5975     0.0894    6.69  5.1e-08 ***
## education     0.3453     0.1136    3.04   0.0042 **
## typeprof     16.6575     6.9930    2.38   0.0221 *
## typewc      -14.6611     6.1088   -2.40   0.0211 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.74 on 40 degrees of freedom
## Multiple R-squared:  0.913,Adjusted R-squared:  0.904
## F-statistic:  105 on 4 and 40 DF,  p-value: <2e-16
```

```
anova(mod2)
```

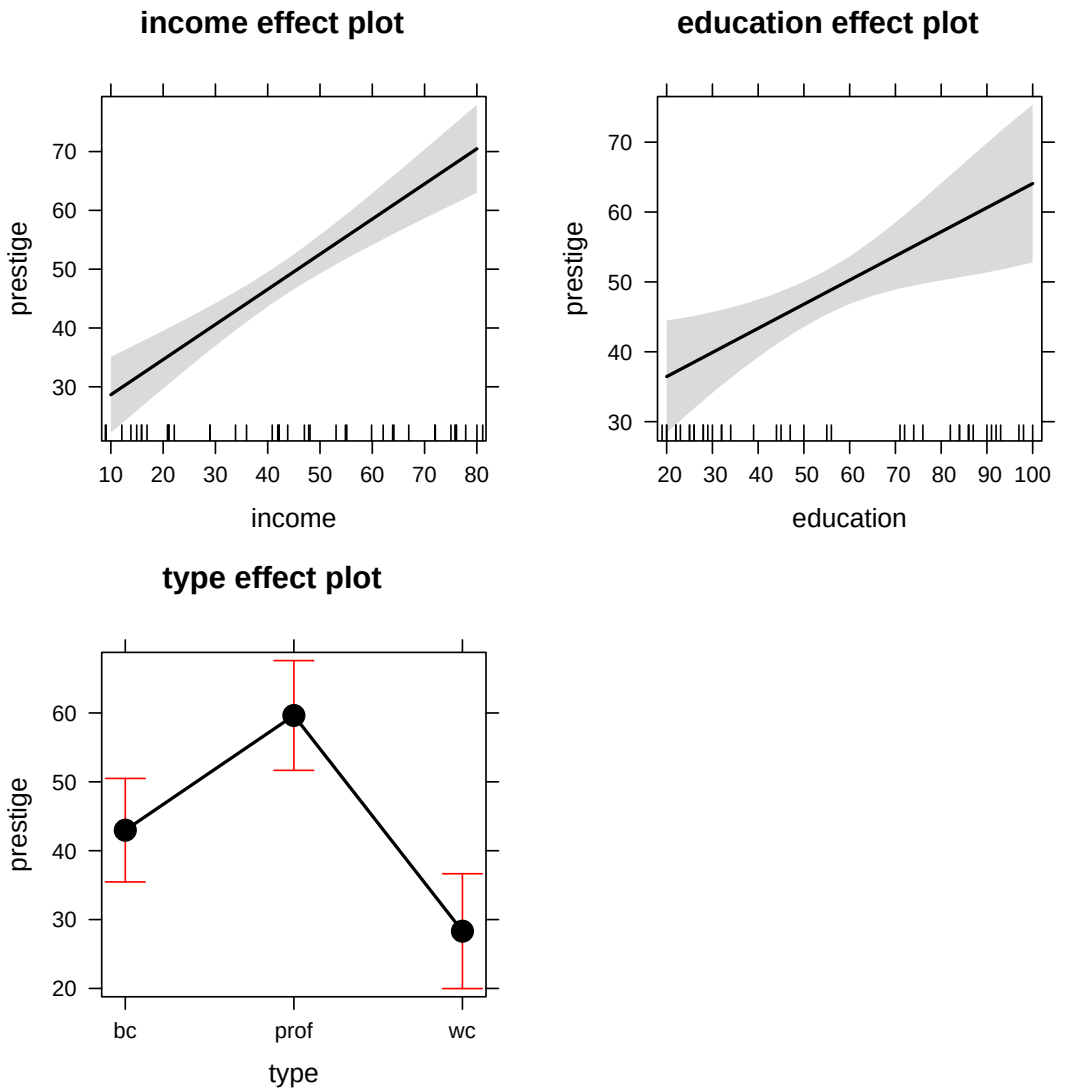
```
## Analysis of Variance Table
##
## Response: prestige
##           Df Sum Sq Mean Sq F value  Pr(>F)
## income     1  30665   30665   323.0 < 2e-16 ***
## education  1   5516    5516    58.1 2.6e-09 ***
## type       2   3709    1854    19.5 1.2e-06 ***
## Residuals 40   3798      95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(mod2)
```

```
##                2.5 %  97.5 %
## (Intercept)  -7.6908  7.3208
## income        0.4170  0.7781
## education     0.1157  0.5749
## typeprof      2.5241 30.7909
## typewc      -27.0074 -2.3148
```

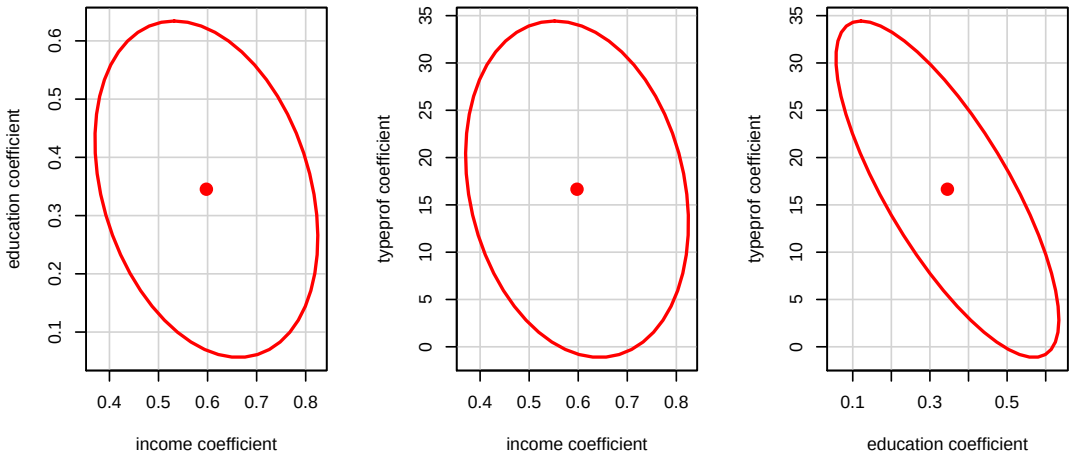
Grafici 'effetto'

```
plot( allEffects(mod2) )
```



Intervalli di confidenza simultanei per ogni coppia di coefficienti

```
par(mfrow=c(1,3))
confidenceEllipse(mod2, which.coef=c(2,3))
confidenceEllipse(mod2, which.coef=c(2,4))
confidenceEllipse(mod2, which.coef=c(3,4))
```


education coefficient
0.1
0.2
0.3
0.4
0.5
0.6
0.4
0.5
0.6
0.7
0.8
income coefficient
typeprof coefficient
0
5
10
15
20
25
30
35
0.4
0.5
0.6
0.7
0.8
income coefficient
typeprof coefficient
0
5
10
15
20
25
30
35
0.1
0.3
0.5
education coefficient