La multicollinearità (e gli outliers...)

Statistica Applicata
Corso di Laurea in Informatica

cristiano.varin@unive.it

Il foglio elettronico `Sales-and-Assets.csv`[1] contiene informazioni sui profitti, le vendite e il patrimonio per un campione delle compagnie presenti elencati in *Fortune 500*. I dati sono espressi in milioni di dollari.
Obiettivo: valutare come vendite e patrimonio siano legate ai profitti.
Leggiamo i dati

```
sales <- read.csv("Sales-and-Assets.csv")
```

Statistiche descrittive

```
summary(sales)
```

```
##      Profit          Sales           Assets
##  Min.   :-925   Min.   : 2161   Min.   :  713
##  1st Qu.: -15   1st Qu.: 2744   1st Qu.: 1117
##  Median : 112   Median : 4788   Median : 1886
##  Mean   : 147   Mean   : 9318   Mean   : 8206
##  3rd Qu.: 228   3rd Qu.: 8353   3rd Qu.: 3053
##  Max.   :1508   Max.   :53913   Max.   :86972
```

Matrice di correlazione

```
round( cor(sales), 2)
```

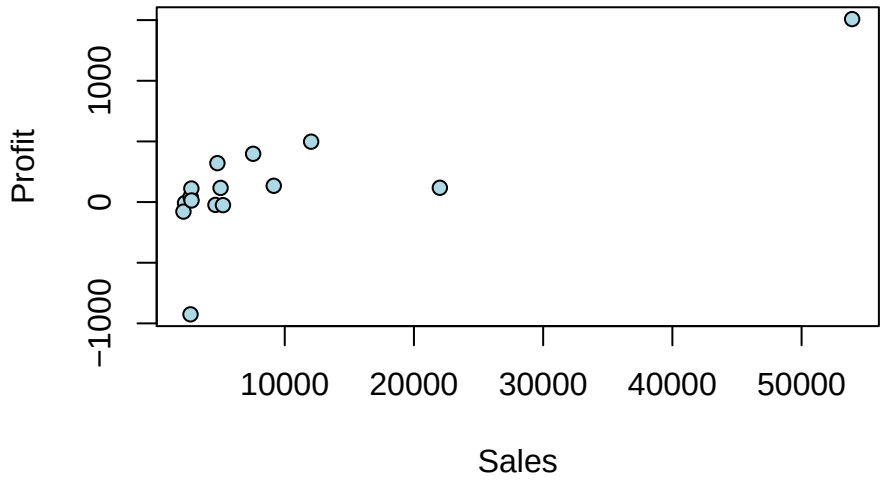
```
##        Profit Sales Assets
## Profit   1.00  0.79   0.78
## Sales    0.79  1.00   0.95
## Assets   0.78  0.95   1.00
```

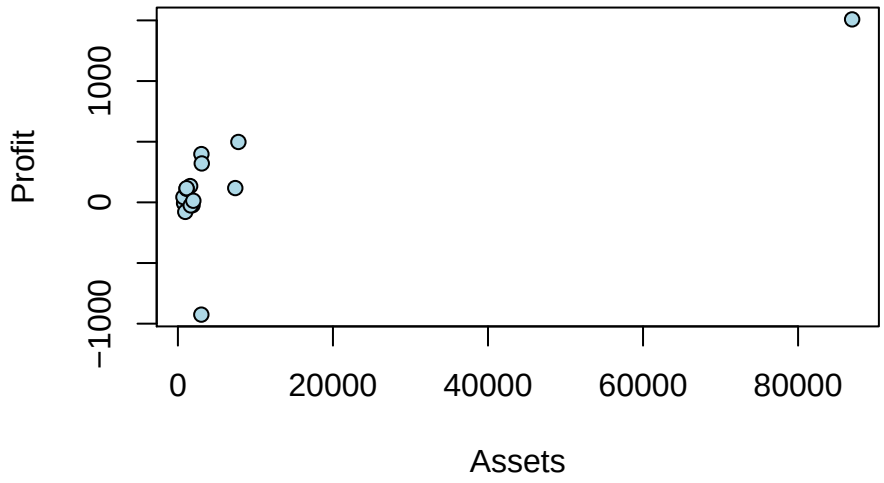
La correlazione fra `Sales` e `Assets` è piuttosto alta...

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

Ispezione grafica della relazione fra profitto e vendite e fra profitto e patrimonio

```
par(mfrow=c(2,1))
with( sales, plot(Profit~Sales, pch=21, bg="lightblue") )
with( sales, plot(Profit~Assets, data=sales, pch=21, bg="lightblue") )
```





Si nota la presenza di uno, forse due outlier...

Modello di regressione lineare con sia `Sales` che `Assets`

```
modello1 <- lm( Profit~Sales+Assets, data=sales )
summary(modello1)
```

```
##
## Call:
## lm(formula = Profit ~ Sales + Assets, data = sales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -909.1  -39.2   30.8  117.8  316.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -88.15665  127.56975   -0.69     0.50
## Sales         0.02007    0.02018    0.99     0.34
## Assets        0.00587    0.01234    0.48     0.64
##
## Residual standard error: 319 on 12 degrees of freedom
## Multiple R-squared:  0.639,Adjusted R-squared:  0.578
## F-statistic: 10.6 on 2 and 12 DF,  p-value: 0.00223
```

Retta di regressione con solo `Sales`

```
modello2 <- lm( Profit~Sales, data=sales )
summary(modello2)
```

```
##
## Call:
## lm(formula = Profit ~ Sales, data = sales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -879.3  -25.1   57.0  125.5  306.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.25e+02   9.85e+01   -1.27   0.2273
## Sales        2.92e-02   6.18e-03    4.72   0.0004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 310 on 13 degrees of freedom
```

```
## Multiple R-squared:  0.632,Adjusted R-squared:  0.603
## F-statistic: 22.3 on 1 and 13 DF,  p-value: 0.000399
```

La statistica 'R^2 aggiustato' è migliorata togliendo `Sales`

Retta di regressione con solo `Assets`

```
modello3 <- lm( Profit~Assets, data=sales )
summary(modello3)

##
## Call:
## lm(formula = Profit ~ Assets, data = sales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -981.4  -41.0  -14.8   98.9  358.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2913    88.3905    0.04   0.9709
## Assets        0.0175     0.0039    4.50   0.0006 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319 on 13 degrees of freedom
## Multiple R-squared:  0.609,Adjusted R-squared:  0.579
## F-statistic: 20.2 on 1 and 13 DF,  p-value: 6e-04
```

Qui abbiamo un R^2 aggiustato di poco più alto del modello di partenza.

Quanto pesano gli outlier sui nostri risultati? Innanzitutto identifichiamoli

```
outlier1 <- which( sales$Assets > 80000)
outlier1

## [1] 4

outlier1 <- which( sales$Sales > 50000)
outlier1

## [1] 4

outlier2 <- which( sales$Profit < -500)
outlier2

## [1] 9

outliers <- c(outlier1, outlier2)
```

Senza gli outlier la matrice di correlazione diventa

```
round( cor(sales[-outliers,]), 2)

##        Profit Sales Assets
## Profit   1.00  0.39   0.63
## Sales    0.39  1.00   0.86
## Assets   0.63  0.86   1.00
```

Ora possiamo ristimare i vari modelli senza gli outlier

```
modello1bis <- lm( Profit~Sales+Assets, data=sales, subset=-outliers )
summary(modello1bis)

##
## Call:
## lm(formula = Profit ~ Sales + Assets, data = sales, subset = -outliers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -145.8 -124.4    5.2   86.3  253.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.1737    60.7071    0.45     0.66
## Sales        -0.0176     0.0140   -1.26     0.24
## Assets        0.0826     0.0327    2.52     0.03 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 139 on 10 degrees of freedom
## Multiple R-squared:  0.484,Adjusted R-squared:  0.381
## F-statistic: 4.69 on 2 and 10 DF,  p-value: 0.0366
```

```
modello2bis <- lm( Profit~Sales, data=sales, subset=-outliers )
summary(modello2bis)

##
## Call:
## lm(formula = Profit ~ Sales, data = sales, subset = -outliers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -202.9 -125.9  -35.1   32.8  302.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.35965   73.59488    0.60     0.56
## Sales        0.01257    0.00885    1.42     0.18
##
## Residual standard error: 170 on 11 degrees of freedom
## Multiple R-squared:  0.155,Adjusted R-squared:  0.0784
## F-statistic: 2.02 on 1 and 11 DF,  p-value: 0.183
```

```
modello3bis <- lm( Profit~Assets, data=sales, subset=-outliers )
summary(modello3bis)

##
## Call:
## lm(formula = Profit ~ Assets, data = sales, subset = -outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -236.40 -107.66    5.41   59.90  250.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.2212    59.4387    0.07     0.94
## Assets        0.0473     0.0174    2.72     0.02 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143 on 11 degrees of freedom
## Multiple R-squared:  0.402,Adjusted R-squared:  0.348
## F-statistic:  7.4 on 1 and 11 DF,  p-value: 0.0199
```

Cosa notate?