Le interazioni

Statistica Applicata
Corso di Laurea in Informatica

cristiano.varin@unive.it

1 Discriminazione di Genere?

Il foglio elettronico `Gender Discrimination.csv`[1] contiene dati relativi ad uno studio per valutare la presenza di discriminazione di genere in un'azienda. Le informazioni disponibili sono:

`Gender` genere

`Experience` anni di esperienza

`Salary` salario annuale in US $

```
gender <- read.csv("Gender Discrimination.csv")
```

Prendiamo contatto con i dati

```
head(gender)
```

```
##   Gender Experience Salary
## 1 Female         15  78200
## 2 Female         12  66400
## 3 Female         15  61200
## 4 Female          3  61000
## 5 Female          4  60000
## 6 Female          4  68000
```

```
dim(gender)
```

```
## [1] 208   3
```

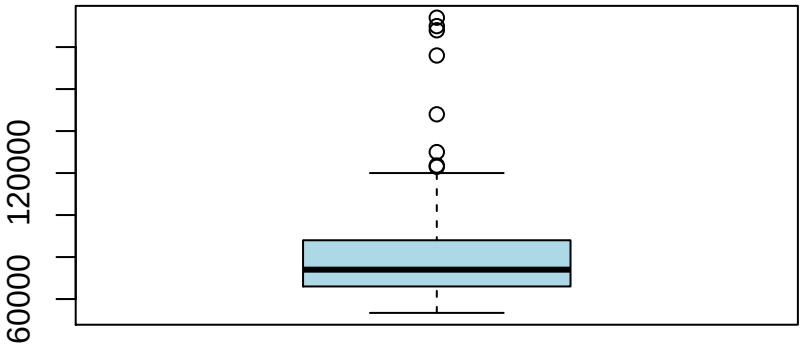
[1] Dati provenient da *Jank, W. (2011). Business Analytics for Managers. Springer.*

```r
summary(gender)
```

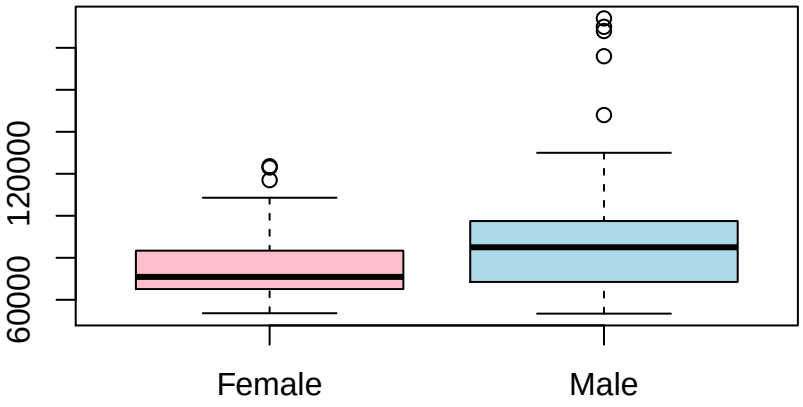
```
##     Gender      Experience       Salary
##  Female:140   Min.   : 2.0   Min.   : 53400
##  Male  : 68   1st Qu.: 7.0   1st Qu.: 66000
##               Median :10.0   Median : 74000
##               Mean   :12.1   Mean   : 79844
##               3rd Qu.:16.0   3rd Qu.: 88000
##               Max.   :39.0   Max.   :194000
```

Qualche figura per visualizzare i dati

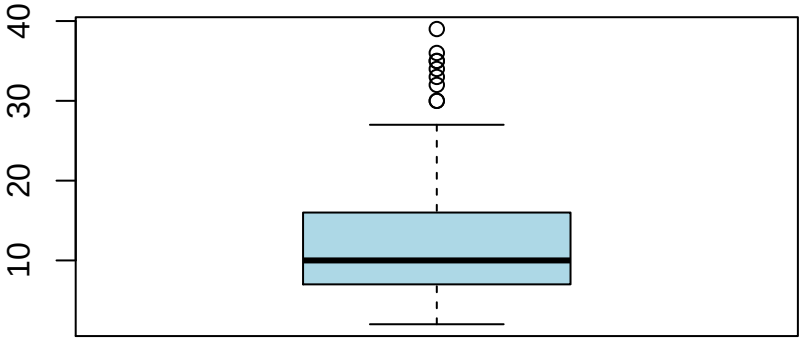
```r
boxplot(gender$Salary, col="lightblue")
```



```
boxplot(Salary~Gender, data=gender, col=c("pink", "lightblue"))
```




```
boxplot(gender$Experience, col="lightblue")
```



Boxplot per visualizzare la relazione fra esperienza e genere

```
boxplot(Experience~Gender, data=gender, col=c("pink","lightblue"))
```

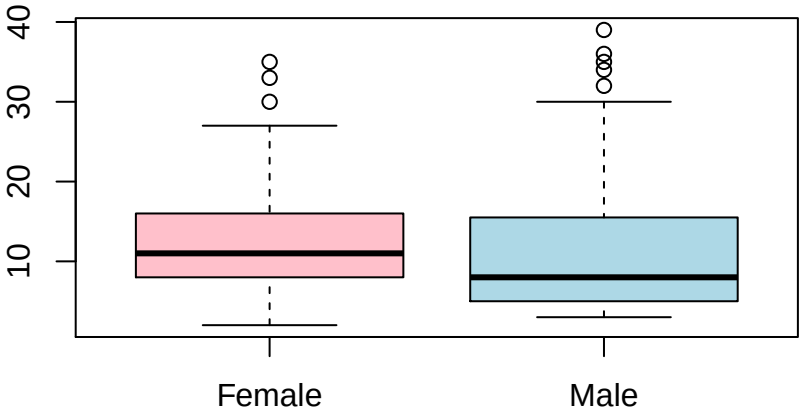
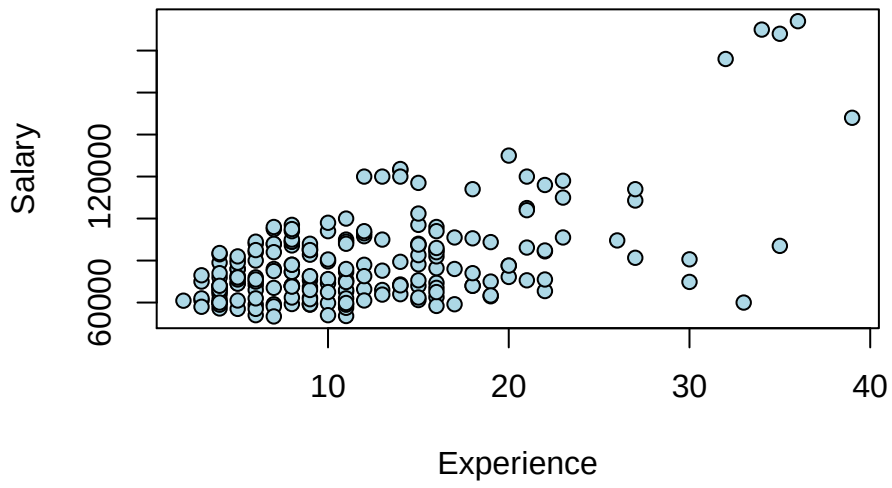


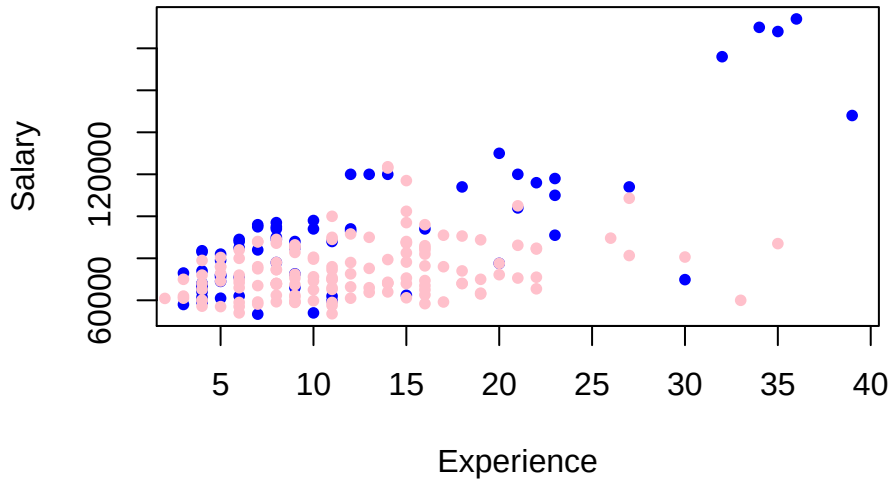
Grafico a dispersione per visualizzare la relazione fra salario ed esperienza

```
plot(Salary~Experience, data=gender, pch=21, bg="lightblue")
```



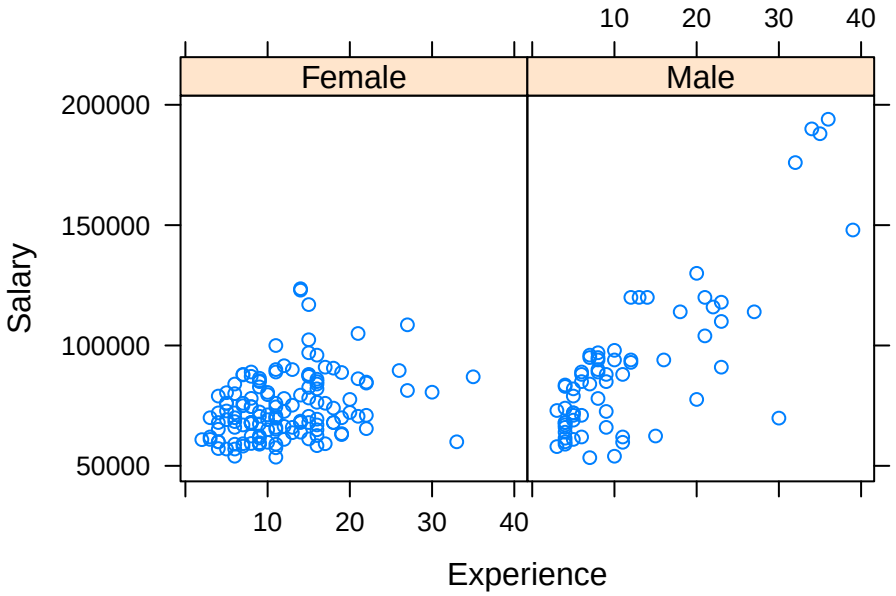
Il grafico a dispersione differenziato per genere

```
plot(Salary[Gender=="Male"]~Experience[Gender=="Male"], pch=20,
data=gender, col="blue", ylab="Salary", xlab="Experience")
points(Salary[Gender=="Female"]~Experience[Gender=="Female"], pch=20,
data=gender, col="pink")
```



Alternativamente, possiamo considerare un grafico a dispersione *condizionato*

```
library(lattice)
xyplot(Salary~Experience | Gender, data=gender)
```



Ora passiamo ai modelli di regressione. Come modello base consideriamo la retta di regressione del salario rispetto al genere

```
modelloA <- lm(Salary~Gender, data=gender)
summary(modelloA)

##
## Call:
## lm(formula = Salary ~ Gender, data = gender)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37611 -12868  -3720   8230 102989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    74420       1789    41.6  < 2e-16 ***
## GenderMale     16591       3129     5.3  2.9e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21200 on 206 degrees of freedom
## Multiple R-squared:  0.12,Adjusted R-squared:  0.116
## F-statistic: 28.1 on 1 and 206 DF,  p-value: 2.94e-07
```

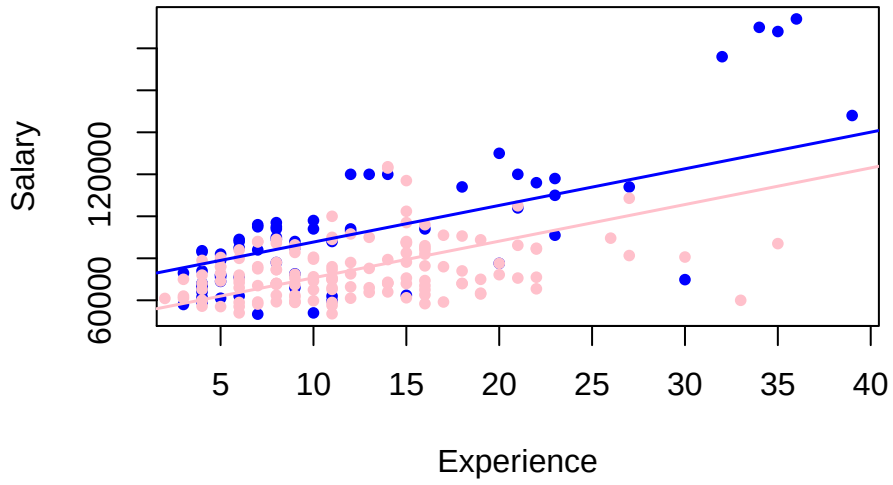
Proviamo ora ad aggiungere l'esperienza

```
modelloB <- lm(Salary~Experience+Gender, data=gender)
summary(modelloB)

##
## Call:
## lm(formula = Salary ~ Experience + Gender, data = gender)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52779  -9806   -121   8347  60913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    53260       2417   22.04  < 2e-16 ***
## Experience      1745        161   10.86  < 2e-16 ***
## GenderMale     17021       2500    6.81  1.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16900 on 205 degrees of freedom
## Multiple R-squared:  0.441,Adjusted R-squared:  0.436
## F-statistic:   81 on 2 and 205 DF,  p-value: <2e-16
```

Visualizziamo il modello stimato

```
plot(Salary[Gender=="Male"]~Experience[Gender=="Male"], pch=20,
data=gender, col="blue", ylab="Salary", xlab="Experience")
points(Salary[Gender=="Female"]~Experience[Gender=="Female"], pch=20,
data=gender, col="pink")
coefB <- coef(modelloB)
abline(coefB[1], coefB[2], col="pink", lwd=1.5)
abline(coefB[1]+coefB[3], coefB[2], col="blue", lwd=1.5)
```



Ora consideriamo, invece, un modello in cui i predittori esperienza e genere *interagiscono*

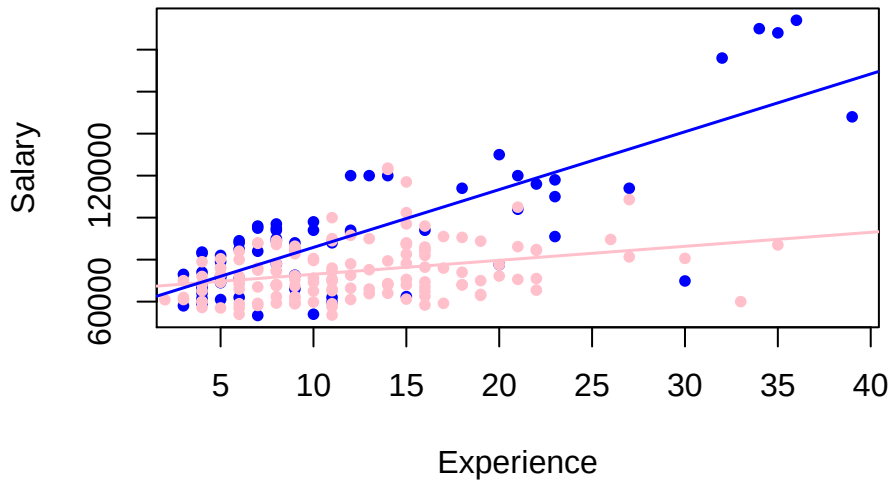
```
modelloC <- lm(Salary~Experience*Gender, data=gender)
summary(modelloC)

##
## Call:
## lm(formula = Salary ~ Experience * Gender, data = gender)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -71048  -9278  -1701   9166  47932
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              66334       2812   23.59   <2e-16 ***
## Experience                 667        206    3.23   0.0015 **
## GenderMale               -8034       4111   -1.95   0.0520 .
## Experience:GenderMale     2086        287    7.26    8e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15100 on 204 degrees of freedom
## Multiple R-squared:  0.556,Adjusted R-squared:  0.55
```

```
## F-statistic: 85.2 on 3 and 204 DF,  p-value: <2e-16
```

Infine visualizziamo il modello con il termine di interazione

```
plot(Salary[Gender=="Male"]~Experience[Gender=="Male"], pch=20,
data=gender, col="blue", ylab="Salary", xlab="Experience")
points(Salary[Gender=="Female"]~Experience[Gender=="Female"], pch=20,
data=gender, col="pink")
coefC <- coef(modelloC)
abline(coefC[1], coefC[2], col="pink", lwd=1.5)
abline(coefC[1]+coefC[3], coefC[2]+coefC[4], col="blue", lwd=1.5)
```



Commenti?