LABORATORIO DI PROBABILITA' E STATISTICA

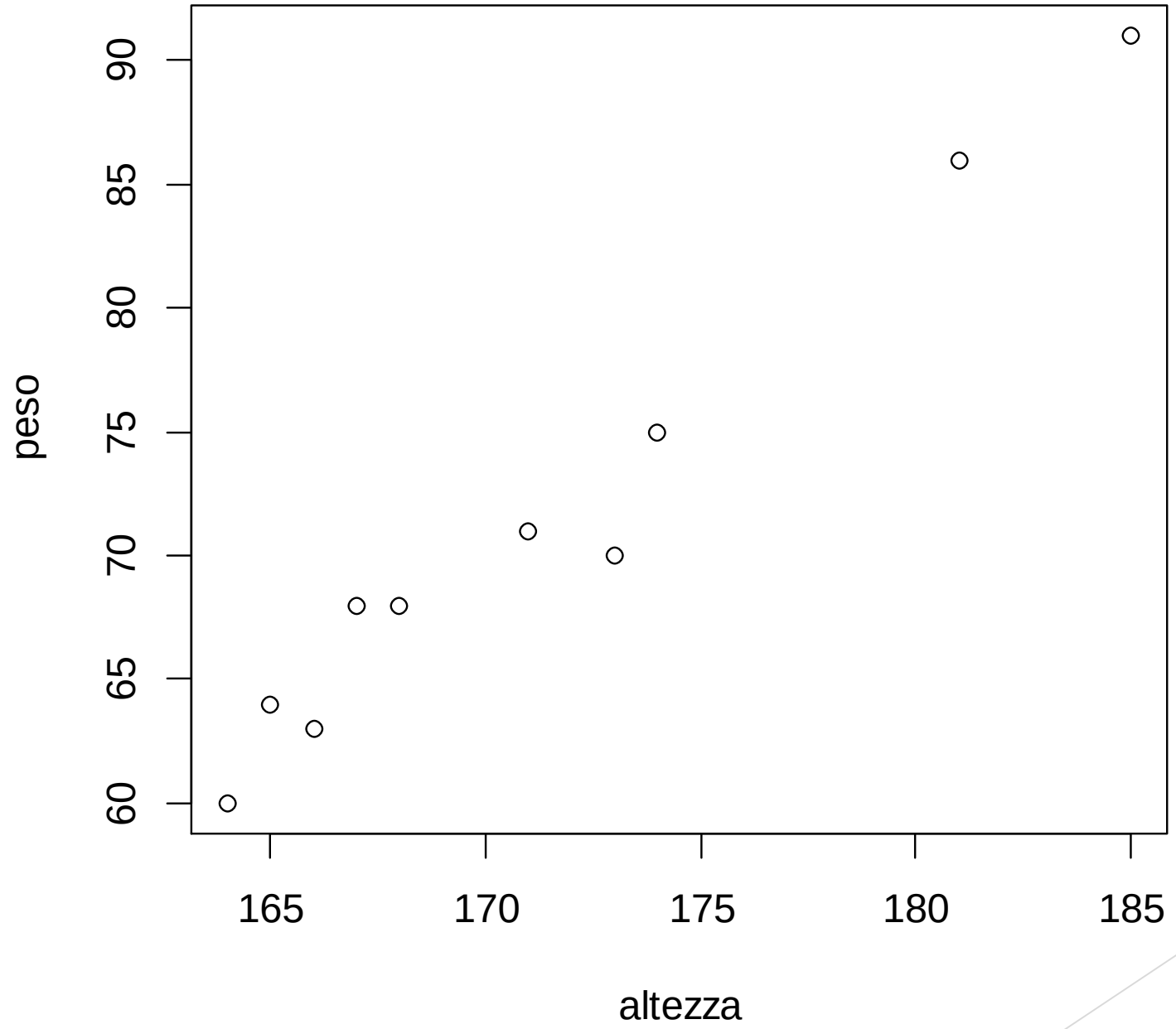
Docente: Bruno Gobbi

3 – LA REGRESSIONE LINEARE

ES. STUDIO RELAZIONE ALTEZZA - PESO

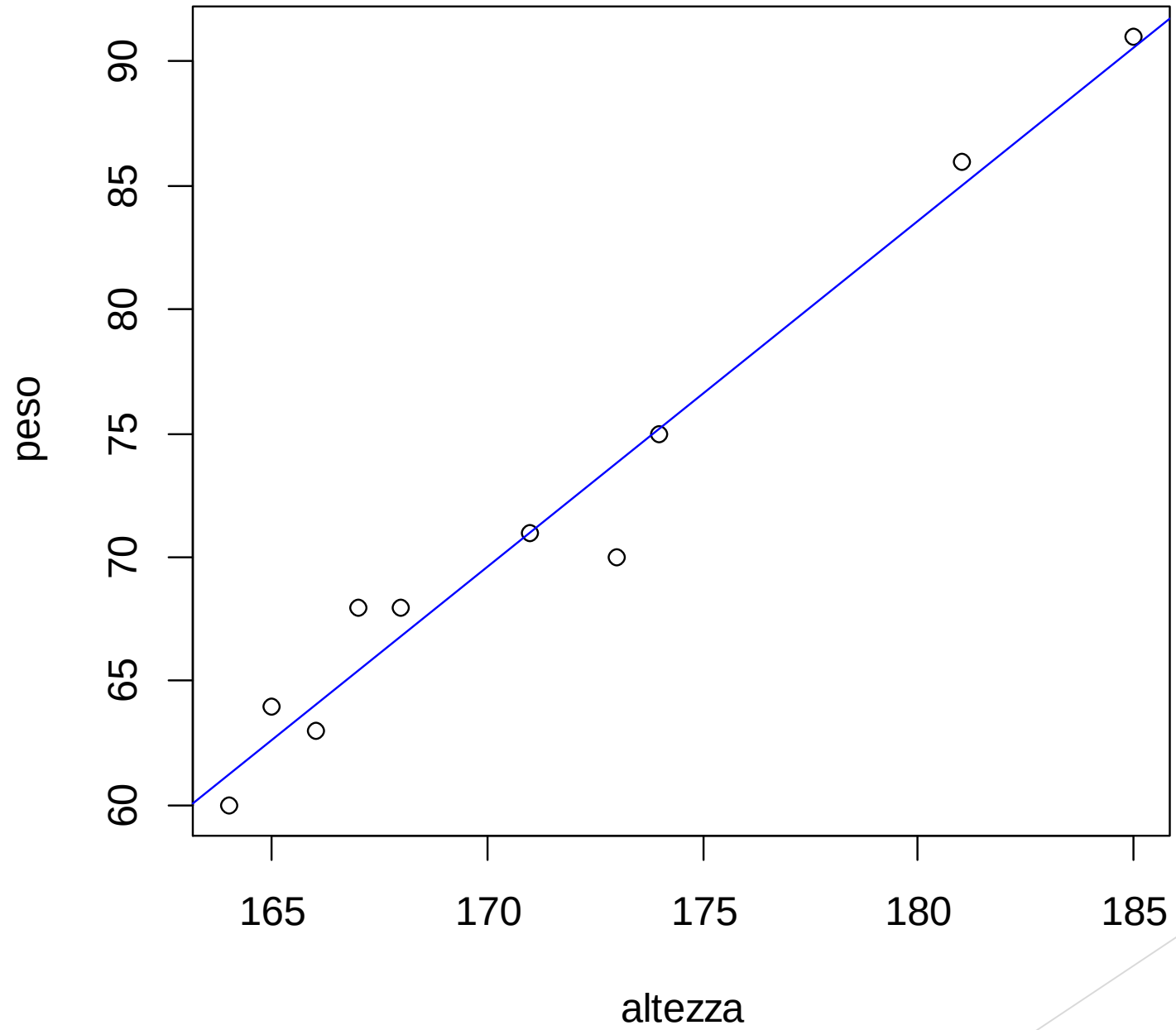
Soggetto	Altezza	Peso
A	174	75
B	166	63
C	173	70
D	171	71
E	168	68
F	167	68
G	165	64
H	164	60
I	181	86
L	185	91

I VARI PUNTI SONO LE OSSERVAZIONI “Y” FRA PESO E ALTEZZA DI 10 PERSONE



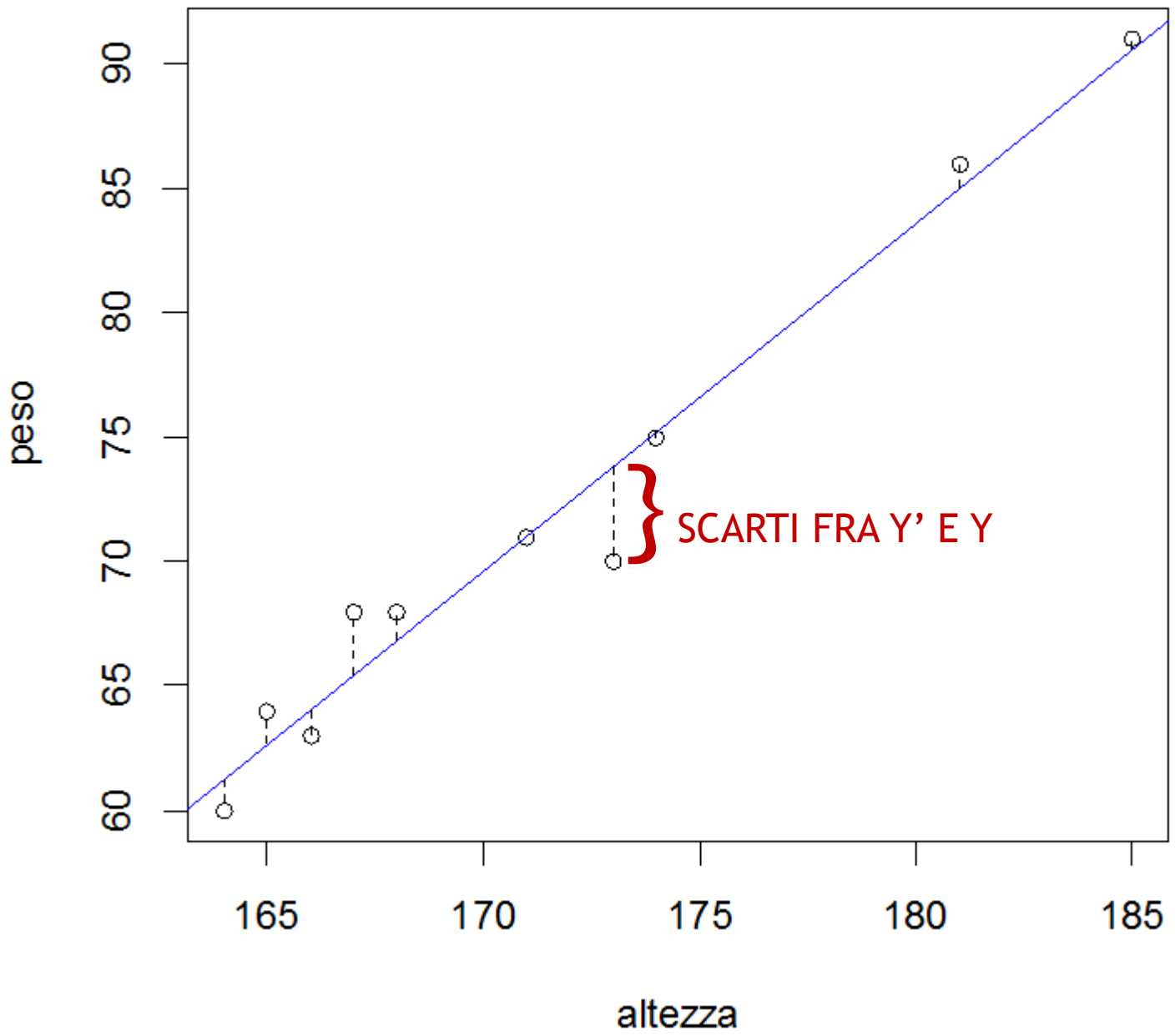
Soggetto	Altezza	Peso
A	174	75
B	166	63
C	173	70
D	171	71
E	168	68
F	167	68
G	165	64
H	164	60
I	181	86
L	185	91

CERCO DI INTERPOLARE CON UNA RETTA Y'= a + bX



Y' (RETTA DI REGRESSIONE)

MISURO GLI SCARTI FRA Y' (RETTA) E Y (OSSERVAZIONI)



Y' (RETTA DI REGRESSIONE)

METODO DEI MINIMI QUADRATI

- BISOGNA TROVARE QUELLA FUNZIONE Y’ CHE ‘’INTERPOLA’’ AL MEGLIO I DATI OSSERVATI Y

- IN QUESTO CASO Y’ SARA’ UNA RETTA DEL TIPO:

$$Y' = a + b\,X$$

METODO DEI MINIMI QUADRATI

- DOMANDA: QUALI PARAMETRI ‘’a’’ E ‘’b’’ PER LA RETTA Y’=a+bX?

- LA SCELTA RICADE SU QUELLI CHE DISEGNANO LA FUNZIONE CHE RENDE MINIMA LA SOMMA DEGLI SCARTI FRA Y’ (MODELLO TEORICO/RETTA) E LE VARIE Y (OSSERVAZIONI)

$$\min \Sigma (Y' - Y)$$

- GRAFICAMENTE GLI SCARTI SONO I SEGMENTI TRATTEGGIATI FRA LA RETTA E I SINGOLI PUNTI OSSERVATI
- POICHE’ CI POSSONO ESSERE DEGLI SCARTI CON VALORI POSITIVI O NEGATIVI FRA Y’ E Y, SI USA IL QUADRATO DEGLI SCARTI

$$\min \Sigma (Y' - Y)^2$$

METODO DEI MINIMI QUADRATI

- AL FINE DI MINIMIZZARE LA SOMMA DEI QUADRATI DEGLI SCARTI FRA Y' E Y, IL METODO MATEMATICO DA SEGUIRE E' QUELLO DI CALCOLARE LA DERIVATA PRIMA DI QUESTA FUNZIONE E POI DI PORLA UGUALE A 0

$$\mathbf{Der(\min \Sigma (Y' - Y)^2) = 0}$$

- SI OTTENGONO COSI' I PARAMETRI MIGLIORI PER LA FUNZIONE TEORICA, CHE NEL CASO DELLA RETTA SONO DATI DAL COEFFICIENTE ANGOLARE "b" E DALL'INTERSEZIONE CON L'ASSE DELLE ORDINATE "a".

ES. STUDIO RELAZIONE ALTEZZA - PESO

Soggetto	Altezza	Peso
A	174	75
B	166	63
C	173	70
D	171	71
E	168	68
F	167	68
G	165	64
H	164	60
I	181	86
L	185	91

```
> altezza = c(174, 166, 173, 171, 168, 167, 165, 164, 181, 185)

> peso = c(75, 63, 70, 71, 68, 68, 64, 60, 86, 91)
```

ES. STUDIO RELAZIONE ALTEZZA - PESO

 # CREIAMO IL GRAFICO DELLE VARIABILI

> plot(altezza, peso)

 # EFFETTUIAMO LA REGRESSIONE LINEARE FRA PESO E ALTEZZA

 # INVERTIRE L'ORDINE DELLE VARIABILI!!!

> retta = lm(peso ~ altezza) **# lm = LINEAR MODEL**

Per scrivere la tilde ~ in Ubuntu premere:
ALT GR + ì

 # PER DISEGNARE LA RETTA DI REGRESSIONE LINEARE

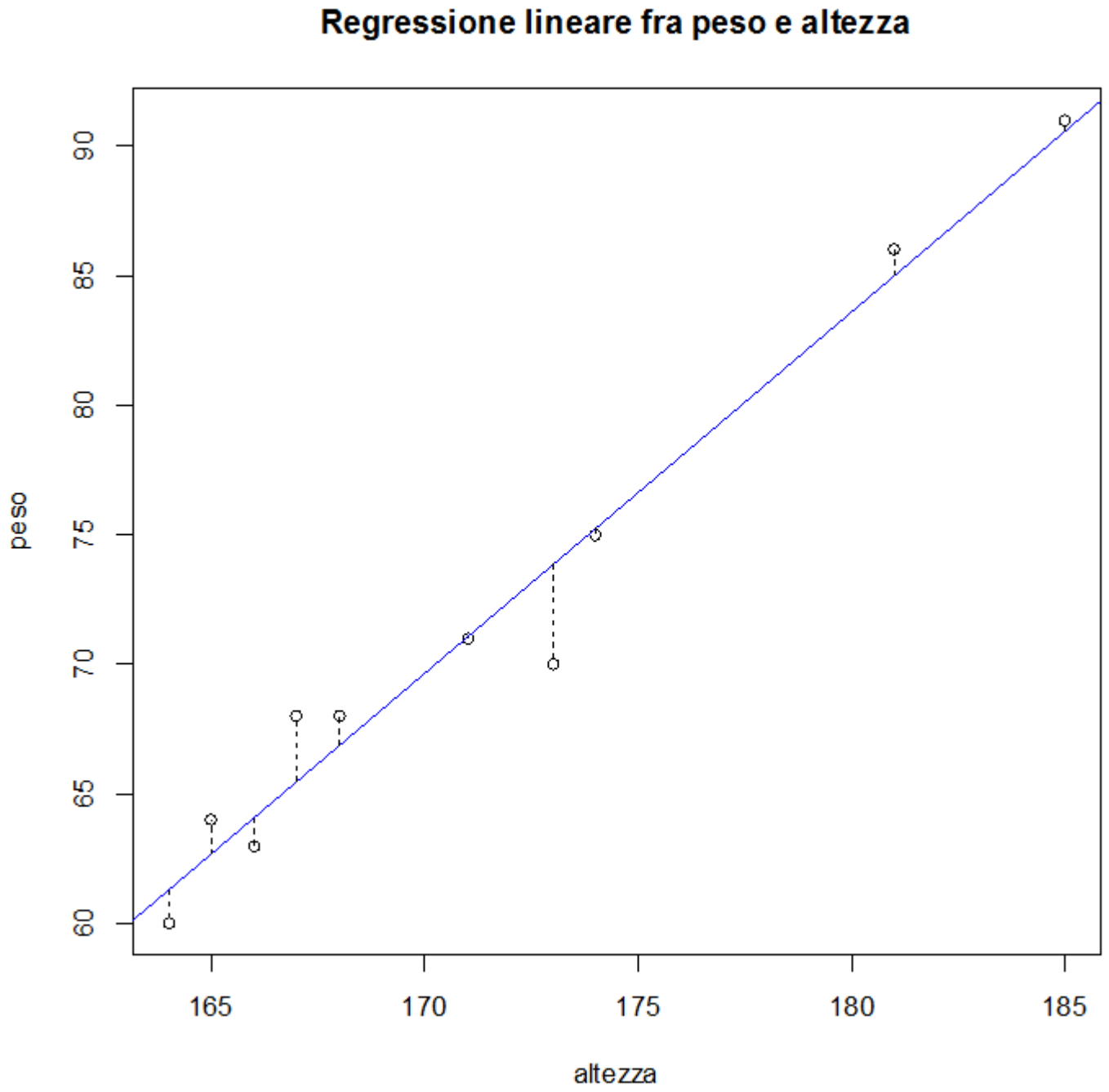
> abline (retta, col="blue") **# (PARAMETRI DELLA RETTA, COLORE)**

 # PER AGGIUNGERE DEI SEGMENTI CHE COLLEGANO LA RETTA AI SINGOLI PUNTI

> segments(altezza, fitted(retta), altezza, peso, lty=2)

 # (X DI PARTENZA, Y DI PARTENZA, X DI ARRIVO, Y DI ARRIVO, TIPO TRATTO)
(lty=2 →TRATTEGGIATO)

> title (main="Regressione lineare fra peso e altezza")


Regressione lineare fra peso e altezza
peso
altezza
60
65
70
75
80
85
90
165
170
175
180
185

OUTPUT DI > summary(retta)

```
Call:

lm(formula = peso ~ altezza)

Residuals:

    Min      1Q  Median      3Q     Max

-3.8336 -0.8535  0.1863  1.1094  2.5425


Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept) -167.67812   15.31583  -10.95 4.30e-06 ***

altezza        1.39602    0.08929   15.63 2.79e-07 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 1.878 on 8 degrees of freedom

Multiple R-squared:  0.9683,    Adjusted R-squared:  0.9643

F-statistic: 244.4 on 1 and 8 DF,  p-value: 2.793e-07
```

1 parte

2 parte

3 parte

PRIMA PARTE DELL'OUTPUT DI > summary(retta)

```
Call:

lm(formula = peso ~ altezza)

Residuals:

    Min      1Q  Median      3Q     Max

-3.8336 -0.8535  0.1863  1.1094  2.5425
```

1 parte

Nella **prima parte** dei risultati è riportata la descrizione dei residui del modello, cioè della distanza che c'è fra la retta di regressione in blu e i singoli punti.

SECONDA PARTE DELL'OUTPUT DI > summary(retta)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -167.67812   15.31583  -10.95 4.30e-06 ***
altezza        1.39602    0.08929   15.63 2.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2 parte

Nella **seconda parte** sono riportati le stime dei coefficienti (Estimate). I coefficienti -167.67512 (riferito all'intercetta) e quello 1.39602 (riferito all'altezza) significano che il modello della retta di regressione (Y'=a+bX) sarà:

peso = -167.67812 + 1.39602 * altezza

I valori di Std. Error indicano il margine di errore che si commette con questa retta per ogni coefficiente. I due successivi t value e PR(>|t|) indicano la significatività di questi coefficienti così stimati, ossia quanto sono affidabili come stime.

TERZA PARTE DELL'OUTPUT DI > summary(retta)

```
Residual standard error: 1.878 on 8 degrees of freedom
Multiple R-squared:  0.9683,    Adjusted R-squared:  0.9643
F-statistic: 244.4 on 1 and 8 DF,  p-value: 2.793e-07
```

3 parte

Nella **terza parte** sono riportati l'errore standard dei residui (Residual standard error) e la bontà di adattamento del modello R^2 (Multiple R-Squared). Vedremo più avanti il significato di questo indicatore.

ANALISI DEI RESIDUI

Al fine di valutare la bontà della regressione lineare che abbiamo condotto, risulta utile fare l'**analisi dei residui**, ossia delle distanza fra ogni punto osservato e la retta di regressione.

Questi residui dovrebbero avere il più possibile una **media nulla** (perché sommando le distanze "positive" sopra la retta e "negative" sotto la retta dovremmo avere un valore vicino a 0) e dovrebbero essere **incorrelati** fra di loro.

ANALISI DEI RESIDUI

Y‘ = -167,67812 + 1,39602 * Altezza

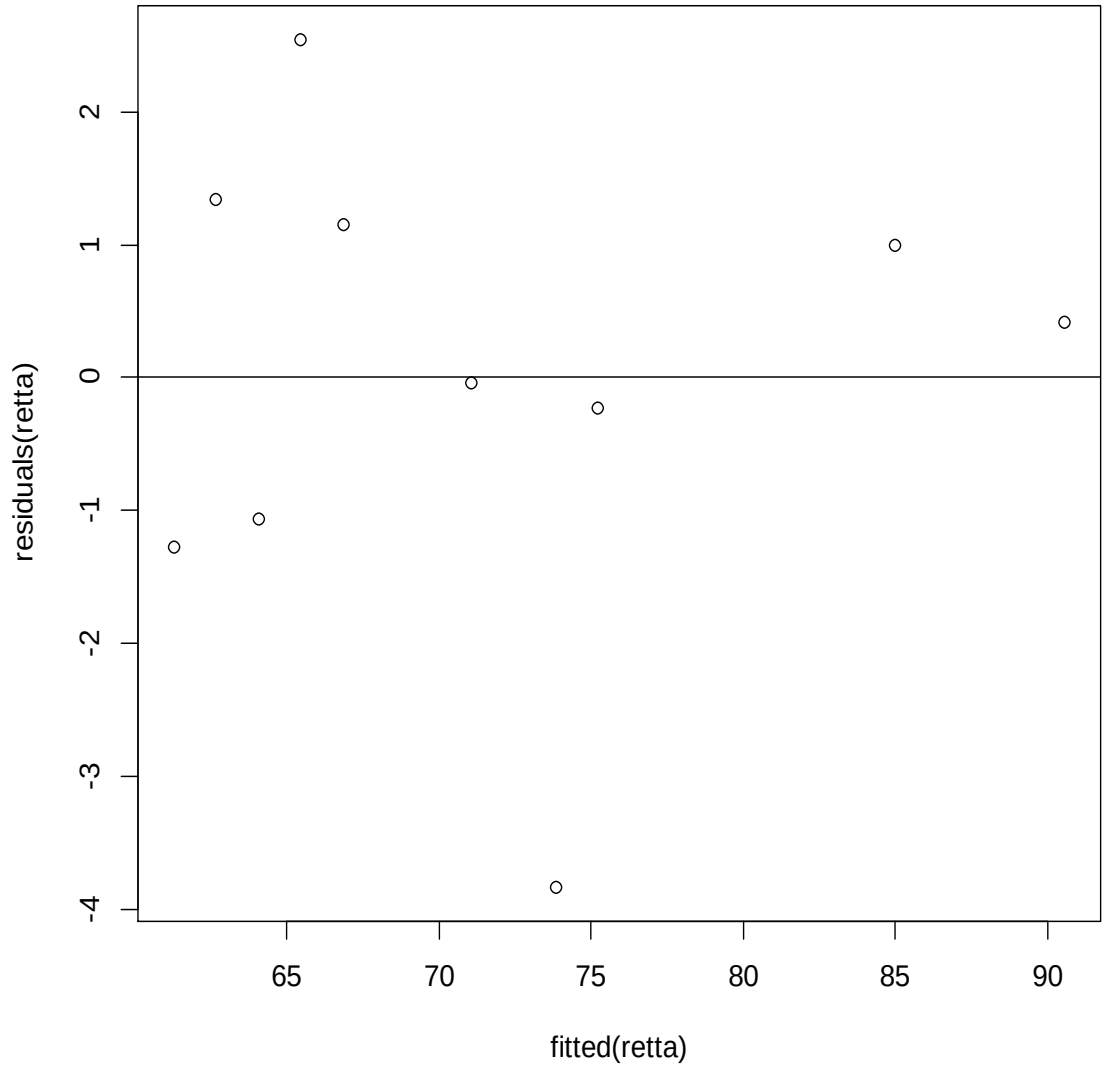
Altezza (X)	Peso (Y)	Y'	residui (Y-Y‘)
164	60	61,26916	-1,26916
165	64	62,66518	1,33482
166	63	64,0612	-1,0612
167	68	65,45722	2,54278
168	68	66,85324	1,14676
171	71	71,0413	-0,0413
173	70	73,83334	-3,83334
174	75	75,22936	-0,22936
181	86	85,0015	0,9985
185	91	90,58558	0,41442

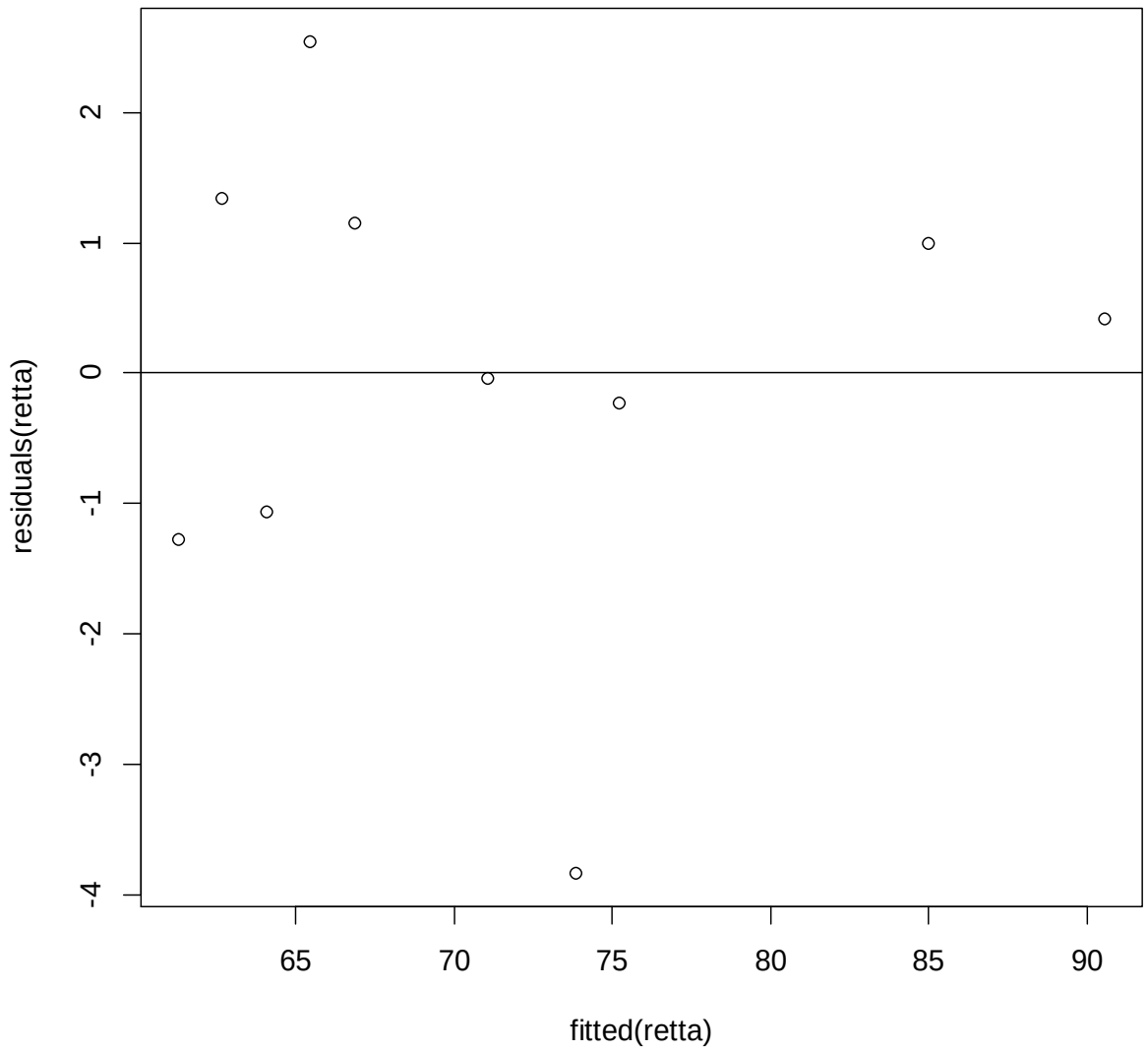
PER IL GRAFICO DEI RESIDUI

```
> plot(fitted(retta), residuals(retta))
```

PER DISEGNARE LA RETTA ORIZZONTALE DELLE ORDINATE ZERO

```
> abline(0, 0)  # (INTERCETTA E COEFFICIENTE ANGOLARE)
```





 # IL GRAFICO CONFERMA L'IPOTESI DI DISTRIBUZIONE CASUALE DEI RESIDUI, PERCHÉ I VALORI SONO EQUIDISTRIBUITI INTORNO ALLA RETTA 0 E SONO PRESENTI SIA SOPRA CHE SOTTO DI ESSA.

IL COEFFICIENTE DI CORRELAZIONE LINEARE

- PER CALCOLARE IL TIPO DI RELAZIONE FRA I FENOMENI SI USA IL COEFFICIENTE DI CORRELAZIONE LINEARE:

$$R = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \rho$$

- R = covarianza fra X e Y diviso s.q.m. di X per s.q.m. di Y
- SE R = -1 → PERFETTA RELAZIONE LINEARE INVERSA
- SE R = 0 → INDIPENDENZA LINEARE
- SE R = +1→ PERFETTA RELAZIONE LINEARE DIRETTA

IL COEFFICIENTE DI CORRELAZIONE LINEARE

$$R = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \rho$$

- IN Rstudio SI CALCOLA CON LA FUNZIONE DIRETTA:

  R = cor(peso, altezza)

- OPPURE SE PREFERIAMO SCRIVERE TUTTA LA FORMULA:

  R = var(peso, altezza)/(sd(peso)*sd(altezza))

IL COEFFICIENTE DI CORRELAZIONE LINEARE

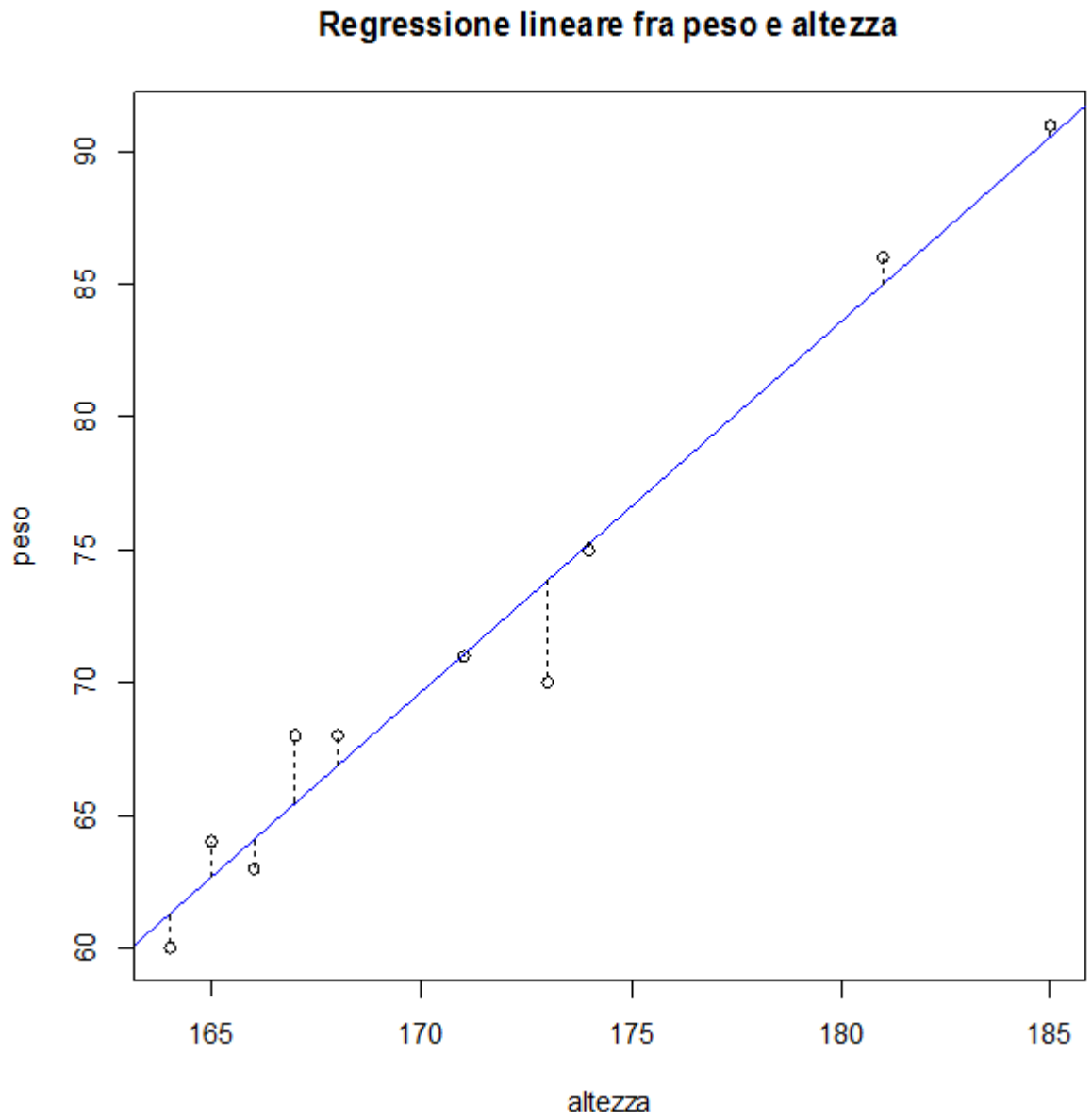
- NEL NOSTRO ESEMPIO:

```
> R = cor(peso, altezza)
> R
[1] 0.9840273
```

- DATO CHE IL VALORE DEL COEFFICIENTE DI CORRELAZIONE LINEARE E' MOLTO VICINO A 1, ALLORA DICIAMO CHE C'E' UNA FORTE RELAZIONE LINEARE DIRETTA FRA I DUE FENOMENI.

- IN ALTRE PAROLE: AL CRESCERE DEL PESO, L'ALTEZZA CRESCE QUASI SEMPRE E LO FA IN MANIERA MOLTO SIMILE AL PESO.

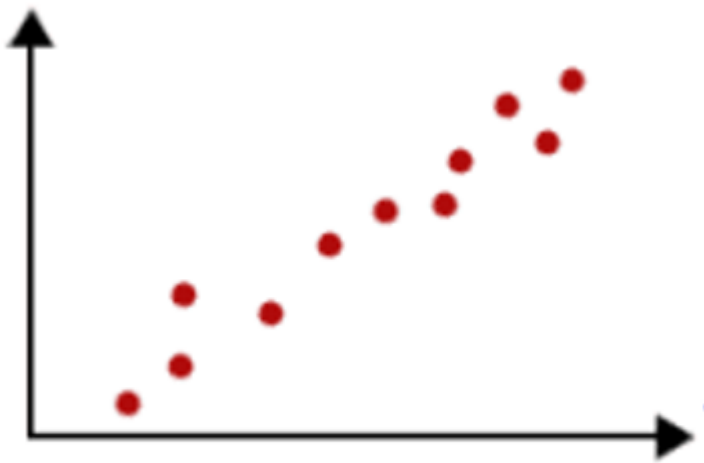
IL COEFFICIENTE DI CORRELAZIONE LINEARE



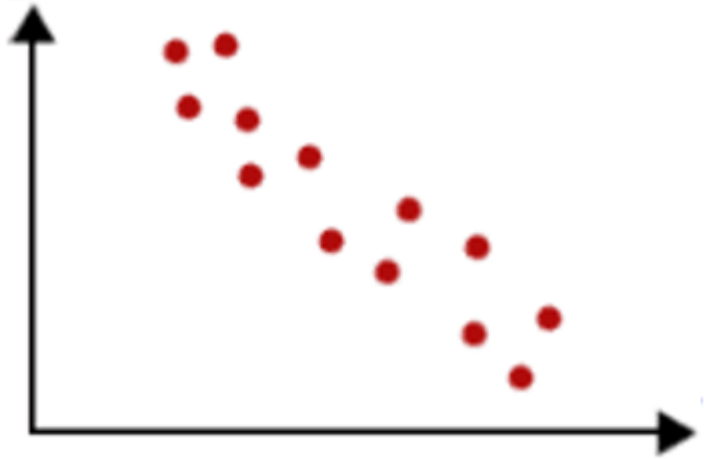
R = 0.9840273

FORTE RELAZIONE LINEARE DIRETTA

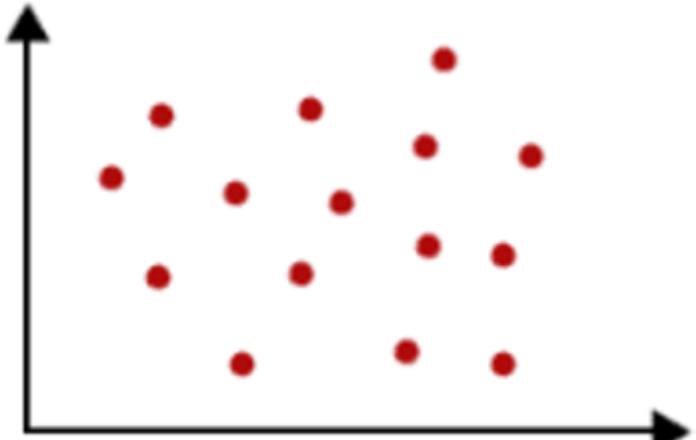
ES. DI COEFFICIENTI DI CORRELAZIONE LINEARE



RELAZIONE LINEARE DIRETTA
$R \rightarrow 1$



RELAZIONE LINEARE INDIRETTA
$R \rightarrow -1$



NESSUNA RELAZIONE LINEARE
$R \rightarrow 0$

NESSUNA RELAZIONE LINEARE
$R \rightarrow 0$

IL COEFFICIENTE DI DETERMINAZIONE: R^2

- AL FINE DI MISURARE QUANTO IL MODELLO TEORICO SIA BUONO (BONTA' DI ACCOSTAMENTO), SI USA IL COEFFICIENTE DI DETERMINAZIONE R^2:

$$R^2 = \frac{\sum(Y' - M(Y))^2}{\sum(Y - M(Y))^2} = \rho^2$$

ovvero

$$R^2 = \frac{dev(Regressione)}{dev\,(Totale)} = \rho^2$$

- QUESTO INDICATORE SI PUO' ANCHE CALCOLARE SEMPLICEMENTE FACENDO IL QUADRATO DEL COEFFICIENTE DI CORRELAZIONE LINEARE ''R'' (CFR. ANCHE 3° PARTE OUTPUT)

IL COEFFICIENTE DI DETERMINAZIONE : R^2

$$0 \leq R^2 \leq 1$$

- SE $R^2 = 0$ IL MODELLO TEORICO Y' NON RIESCE A SPIEGARE NULLA DELLA VARIABILITA' DELLE OSSERVAZIONI Y
- SE $R^2 = 1$ IL MODELLO TEORICO Y' SPIEGA IN MANIERA PERFETTA LA VARIABILITA' DELLE OSSERVAZIONI Y
  → IL MODELLO TEORICO ASSUME GLI STESSI VALORI DI Y
  → Y'=Y

```
> R2=R^2
[1] 0.9683098
```

POICHE' R^2 E' VICINO A 1, IL MODELLO TROVATO E' MOLTO BUONO E RIESCE A SPIEGARE QUASI COMPLETAMENTE LA VARIABILITA' DELLE Y

ES. STUDIO RELAZIONE ORE DI STUDIO – VOTO ESAME

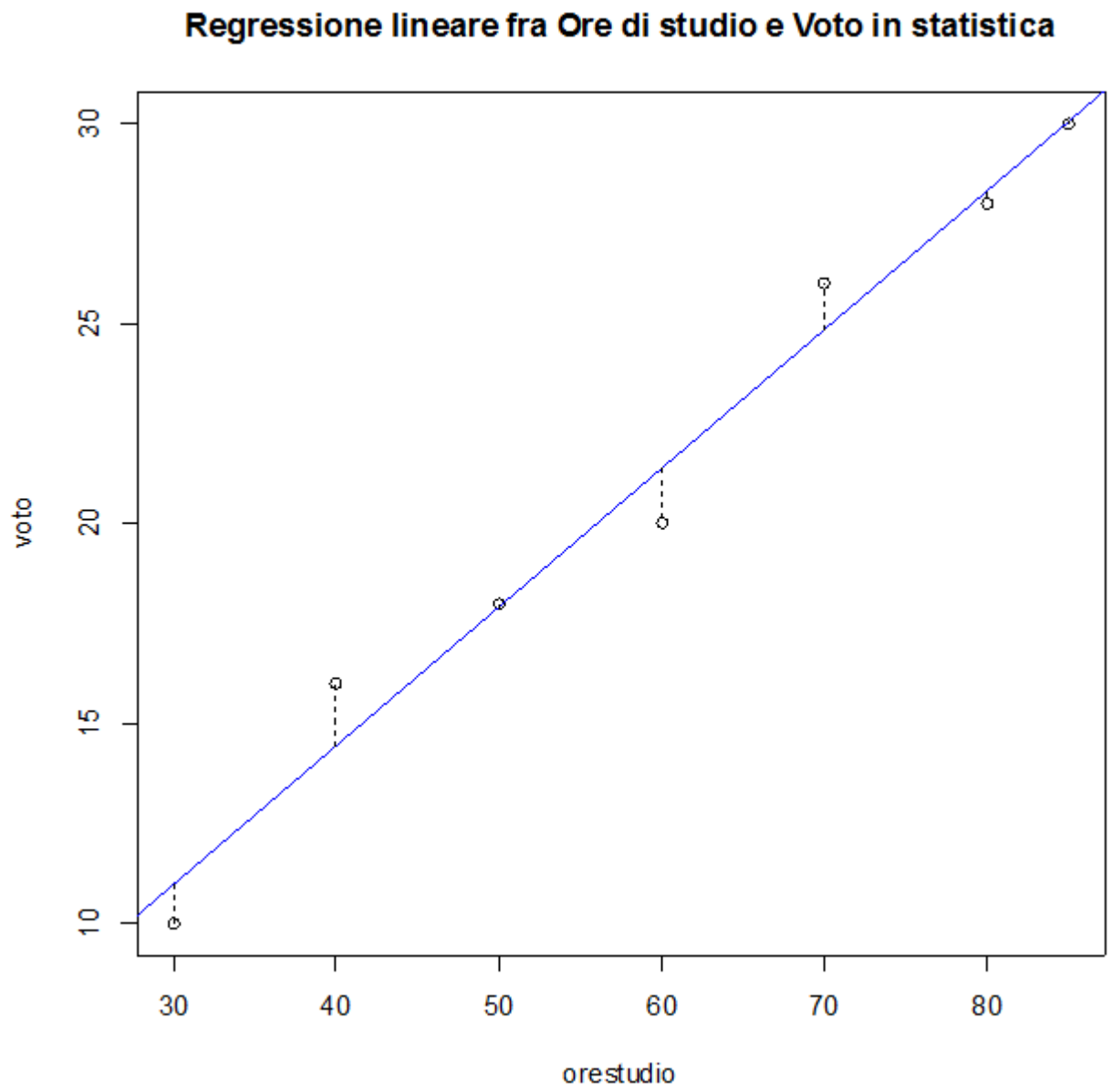
Ore studio	Voto
30	10
50	18
40	16
85	30
60	20
80	28
70	26

Studiare la relazione fra le Ore di studio e il Voto ottenuto all'esame di 7 studenti utilizzando la regressione lineare, disegnando il grafico, calcolando i parametri della retta interpolante, i residui con grafico, il coefficiente di correlazione lineare e giudicandone la bontà di accostamento.

ES. STUDIO RELAZIONE ORE DI STUDIO – VOTO ESAME

```
> orestudio=c(30, 50, 40, 85, 60, 80, 70)
> voto=c(10, 18, 16, 30, 20, 28, 26)
> rettastat=lm(voto~orestudio)
> plot(orestudio, voto)
> abline(rettastat, col="blue")
> segments(orestudio, fitted(rettastat), orestudio, voto, lty=2)
> title(main="Regressione lineare fra Ore di studio e Voto in statistica")
```

ES. STUDIO RELAZIONE ORE DI STUDIO – VOTO ESAME



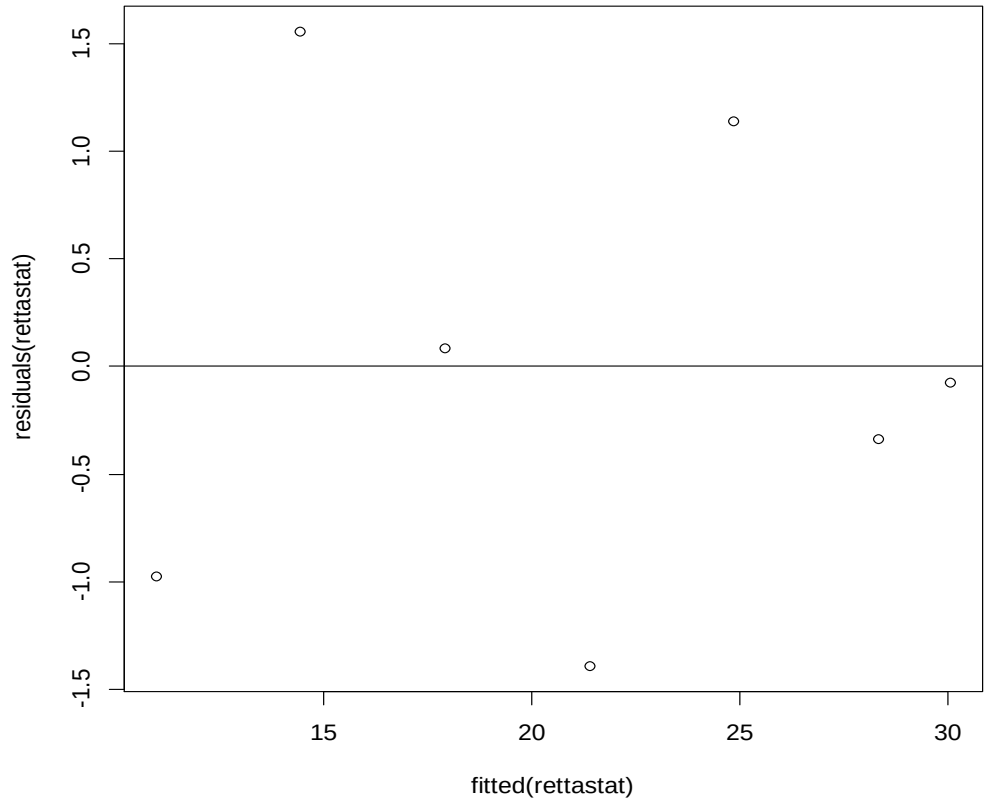
ES. STUDIO RELAZIONE ORE DI STUDIO – VOTO ESAME

```
> summary (rettastat)

Call:
lm(formula = voto ~ orestudio)

Residuals:
      1       2       3       4       5       6       7
-0.97167  0.08215  1.55524 -0.07365 -1.39093 -0.33711  1.13598

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.55241    1.43653   0.385    0.716
orestudio    0.34731    0.02308  15.050 2.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.159 on 5 degrees of freedom
Multiple R-squared:  0.9784,    Adjusted R-squared:  0.9741
F-statistic: 226.5 on 1 and 5 DF,  p-value: 2.346e-05
```

ES. STUDIO RELAZIONE ORE DI STUDIO – VOTO ESAME

 # I PARAMETRI TROVATI SONO a=0,55241 E b=0,37431
QUINDI IL MODELLO TEORICO SARA':
Y'=0.55241+0.37431*orestudio

 # EFFETTUIAMO L'ANALISI DEI RESIDUI
> plot(fitted(rettastat), residuals(rettastat))
>abline(0, 0)



L'analisi dei residui conferma che questi si distribuiscono in maniera uniforme e apparentemente casuale attorno all'asse zero, quindi si può confermare l'ipotesi di distribuzione casuale degli stessi, con media nulla e incorrelazione.

ES. STUDIO RELAZIONE ORE DI STUDIO – VOTO ESAME

 # CALCOLIAMO IL COEFFICIENTE DI CORRELAZIONE LINEARE:

```
> R=cor(orestudio, voto)
> R
[1] 0.9891421
```

 # POICHE' R E' MOLTO VICINO A 1 POSSIAMO AFFERMARE CHE C'E' UNA FORTE RELAZIONE LINEARE DIRETTA FRA LE DUE VARIABILI

 # CALCOLIAMO IL COEFFICIENTE DI DETERMINAZIONE FACENDO IL QUADRATO DI R PER GIUDICARE LA BONTA' DI ACCOSTAMENTO:

```
> R2=R^2
> R2
[1] 0.978402
```

 # DATO CHE R2 E' QUASI UGUALE A 1, DICIAMO CHE IL MODELLO TEORICO USATO SI ADATTA MOLTO BENE AI VALORI OSSERVATI
A TITOLO DI VERIFICA, LO STESSO VALORE E' PRESENTE ANCHE NELLA TERZA PARTE DELL'OUTPUT DELLA summary

ES. LONGLEY – POPOLAZIONE E OCCUPATI

Utilizzando la serie storica di ‘’longley”, presente nei dataset precaricati di RStudio (usare il comando ‘’data()” per ottenerne una lista), analizzare la relazione fra le variabili:

- **Population** (per richiamare i dati: pop=longley$Population)
- **Employed** (per richiamare i dati: occu=longley$Employed)

Attraverso una regressione lineare determinare:

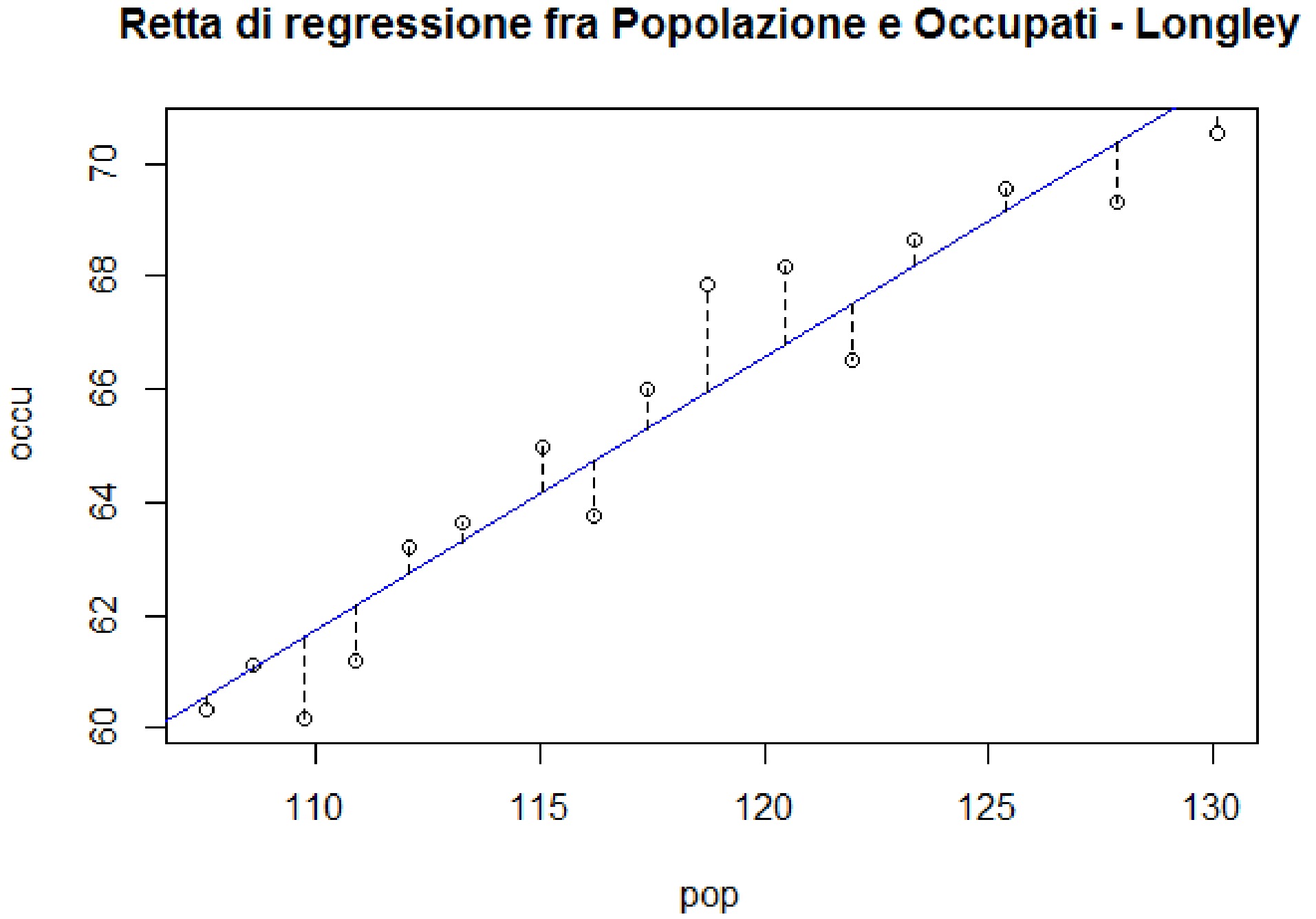
- Grafico del modello teorico Y’=a+bX
- Coefficiente angolare e intersezione con l’asse delle ordinate della retta di regressione
- Analisi dei residui con relativo grafico
- Verificare il tipo di relazione con R
- Giudicare la bontà di accostamento

ES. LONGLEY – POPOLAZIONE E OCCUPATI

```
> longley
     GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
1947         83.0 234.289      235.6        159.0    107.608 1947   60.323
1948         88.5 259.426      232.5        145.6    108.632 1948   61.122
1949         88.2 258.054      368.2        161.6    109.773 1949   60.171
1950         89.5 284.599      335.1        165.0    110.929 1950   61.187
1951         96.2 328.975      209.9        309.9    112.075 1951   63.221
1952         98.1 346.999      193.2        359.4    113.270 1952   63.639
1953         99.0 365.385      187.0        354.7    115.094 1953   64.989
1954        100.0 363.112      357.8        335.0    116.219 1954   63.761
1955        101.2 397.469      290.4        304.8    117.388 1955   66.019
1956        104.6 419.180      282.2        285.7    118.734 1956   67.857
1957        108.4 442.769      293.6        279.8    120.445 1957   68.169
1958        110.8 444.546      468.1        263.7    121.950 1958   66.513
1959        112.6 482.704      381.3        255.2    123.366 1959   68.655
1960        114.2 502.601      393.1        251.4    125.368 1960   69.564
1961        115.7 518.173      480.6        257.2    127.852 1961   69.331
1962        116.9 554.894      400.7        282.7    130.081 1962   70.551
```

ES. LONGLEY – POPOLAZIONE E OCCUPATI

```
> pop=longley$Population
> occu=longley$Employed
> plot(pop, occu)
> rettalong=lm(occu ~ pop)
> abline (rettalong, col="blue")
> segments(pop, fitted(rettalong), pop, occu,
lty=2)
> title(main="Retta di regressione fra Popolazione e
Occupati - Longley")
```


Retta di regressione fra Popolazione e Occupati - Longley
occu
60
62
64
66
68
70
110
115
120
125
130
pop

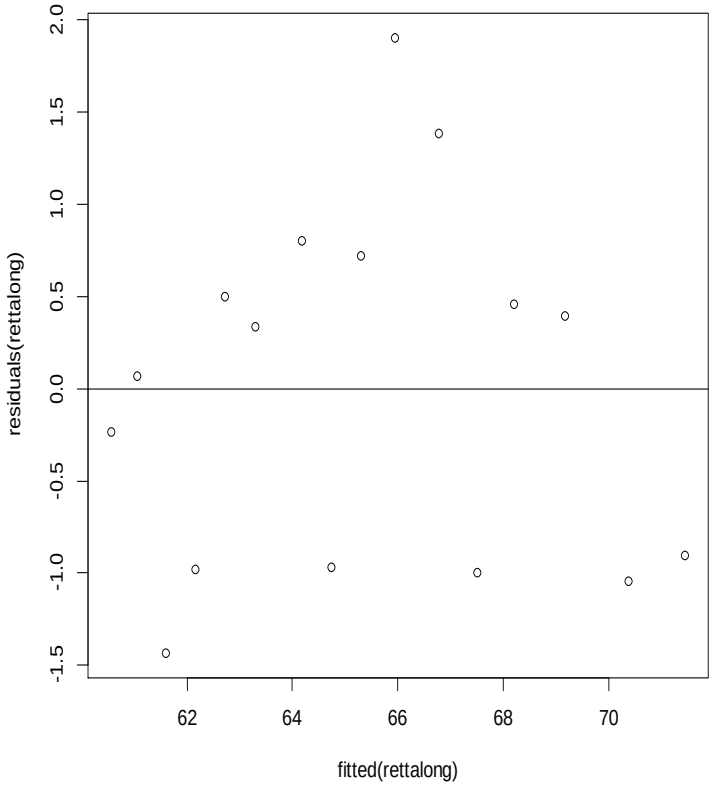
ES. LONGLEY – POPOLAZIONE E OCCUPATI

```
> summary(rettalong)

Call:
lm(formula = occu ~ pop)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4362 -0.9740  0.2021  0.5531  1.9048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.3807     4.4224   1.895   0.0789 .
pop           0.4849     0.0376  12.896 3.69e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.013 on 14 degrees of freedom
Multiple R-squared:  0.9224,    Adjusted R-squared:  0.9168
F-statistic: 166.3 on 1 and 14 DF,  p-value: 3.693e-09
```

ES. LONGLEY – POPOLAZIONE E OCCUPATI

 # I PARAMETRI TROVATI SONO a=8.3807 E b=0.4849
QUINDI IL MODELLO TEORICO SARA':
Y'=8.3807+0.4849*pop

 # EFFETTUIAMO L'ANALISI DEI RESIDUI
> plot(fitted(rettalong), residuals(rettalong))
> abline(0, 0)



L'analisi dei residui conferma che questi si distribuiscono in maniera uniforme e apparentemente casuale attorno all'asse zero, quindi si può confermare l'ipotesi di distribuzione casuale degli stessi, con media nulla e incorrelazione.

ES. LONGLEY – POPOLAZIONE E OCCUPATI

 # CALCOLIAMO IL COEFFICIENTE DI CORRELAZIONE LINEARE:

```
> R=cor(pop, occu)
> R
[1] 0.9603906
```

 # POICHE' R E' VICINISSIMO A 1, POSSIAMO AFFERMARE CHE C'E' UNA FORTE RELAZIONE LINEARE DIRETTA FRA LE DUE VARIABILI

 # CALCOLIAMO IL COEFFICIENTE DI DETERMINAZIONE FACENDO IL QUADRATO DI R PER GIUDICARE LA BONTA' DI ACCOSTAMENTO:

```
> R2=R^2
> R2
[1] 0.9223501
```

 # DATO CHE R2 E' MOLTO VICINO A 1, DICIAMO CHE IL MODELLO TEORICO USATO SI ADATTA MOLTO BENE AI VALORI OSSERVATI
A TITOLO DI VERIFICA, LO STESSO VALORE E' PRESENTE ANCHE NELLA TERZA PARTE DELL'OUTPUT DELLA summary

ES. WOMEN – ALTEZZA E PESO AMERICANE

Utilizzando la serie storica di ‘women’’, presente nei dataset precaricati di RStudio (usare il comando ‘’data()’’ per ottenerne una lista), analizzare la relazione fra le variabili:

- **height**
- **weight**

Attraverso una regressione lineare determinare:

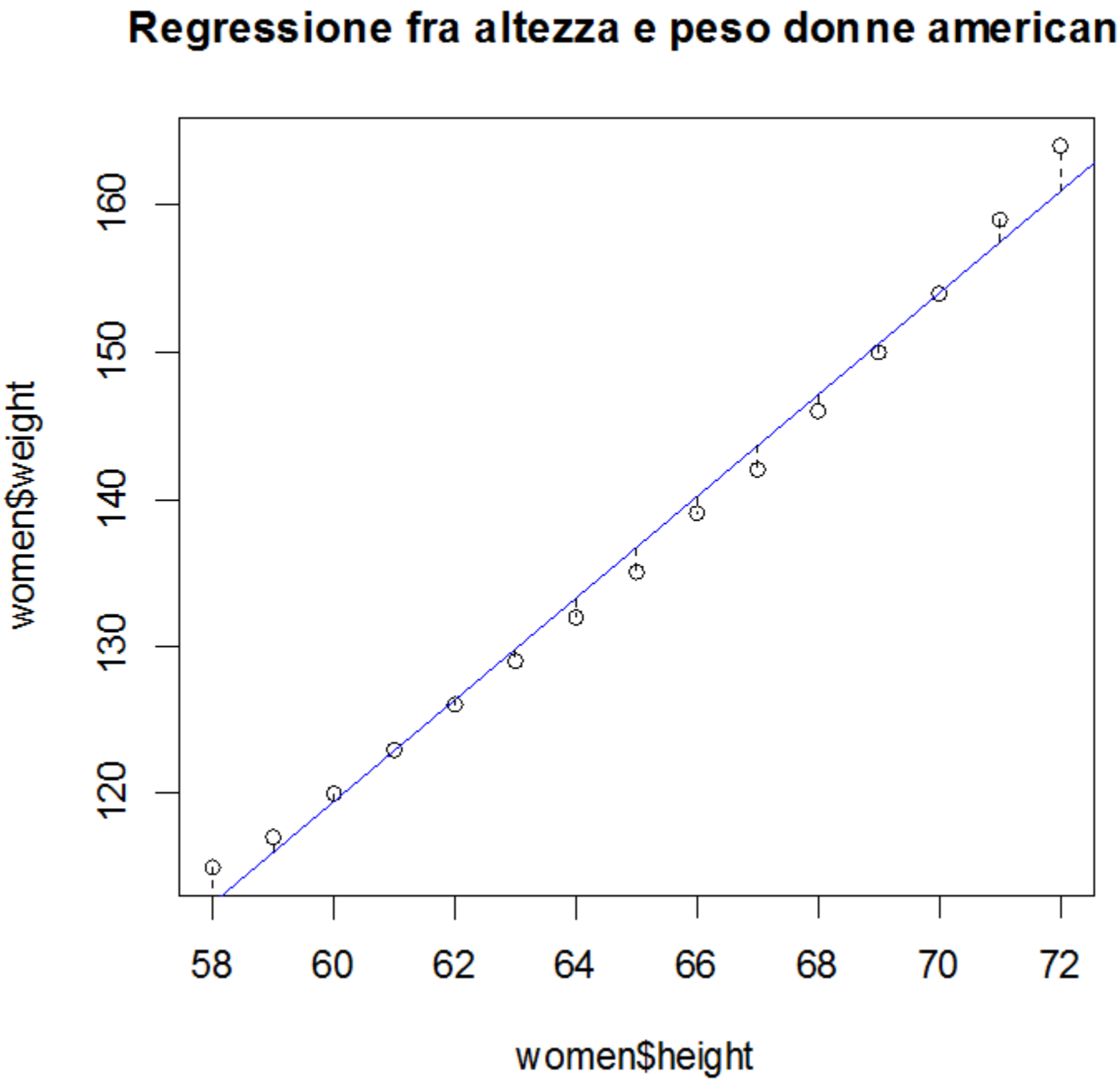
- Grafico del modello teorico Y’=a+bX
- Coefficiente angolare e intersezione con l’asse delle ordinate della retta di regressione
- Analisi dei residui con relativo grafico
- Verificare il tipo di relazione con R
- Giudicare la bontà di accostamento

ES. WOMEN – ALTEZZA E PESO AMERICANE

```
> women
   height weight
1      58    115
2      59    117
3      60    120
4      61    123
5      62    126
6      63    129
7      64    132
8      65    135
9      66    139
10     67    142
11     68    146
12     69    150
13     70    154
14     71    159
15     72    164
```

ES. WOMEN – ALTEZZA E PESO AMERICANE

```
> plot(women$height, women$weight)

> retta=lm(women$weight ~ women$height)

> abline(retta, col="blue")

> segments(women$height, fitted(retta), women$height, women$weight, lty=2)

> title(main="Regressione fra altezza e peso donne americane")
```


Regressione fra altezza e peso donne americane
women$weight
120
130
140
150
160
58
60
62
64
66
68
70
72
women$height

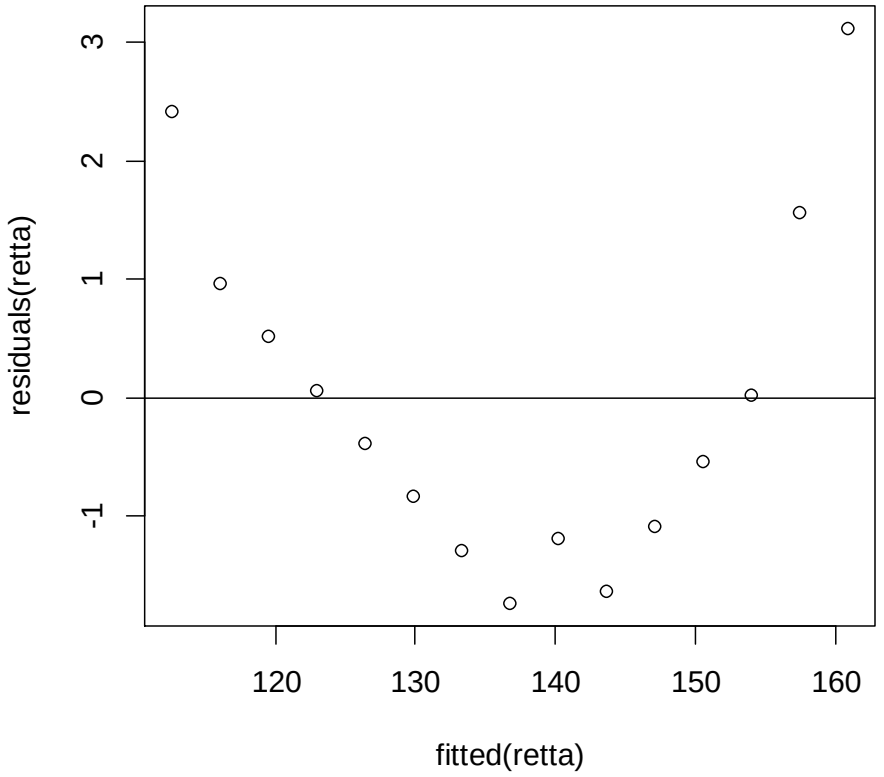
ES. WOMEN – ALTEZZA E PESO AMERICANE

```
> summary(retta)

Call:
lm(formula = women$weight ~ women$height)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7333 -1.1333 -0.3833  0.7417  3.1167

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -87.51667    5.93694  -14.74 1.71e-09 ***
women$height   3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991,     Adjusted R-squared:  0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

ES. WOMEN – ALTEZZA E PESO AMERICANE

 # I PARAMETRI TROVATI SONO a=8.3807 E b=0.4849
QUINDI IL MODELLO TEORICO SARA':
Y'= -87.51667 + 3.45000 *height

 # EFFETTUIAMO L'ANALISI DEI RESIDUI
> plot(fitted(rettalong), residuals(rettalong))
> abline(0, 0)



Questa volta l'analisi dei residui evidenzia un andamento a parabola. Questo significa che **il modello della retta di regressione non è il migliore** per questo caso e che sarebbe opportuno riprovare a fare la regressione con una parabola del tipo:

$Y' = a + bX + cX^2$

ES. WOMEN – ALTEZZA E PESO AMERICANE

 # CALCOLIAMO COMUNQUE LA CORRELAZIONE LINEARE:

```
> R=cor(women$height, women$weight)
> R
[1] 0.9954948
```

 # POICHE' R E' VICINISSIMO A 1, POSSIAMO AFFERMARE CHE C'E' UNA FORTE RELAZIONE LINEARE DIRETTA FRA LE DUE VARIABILI

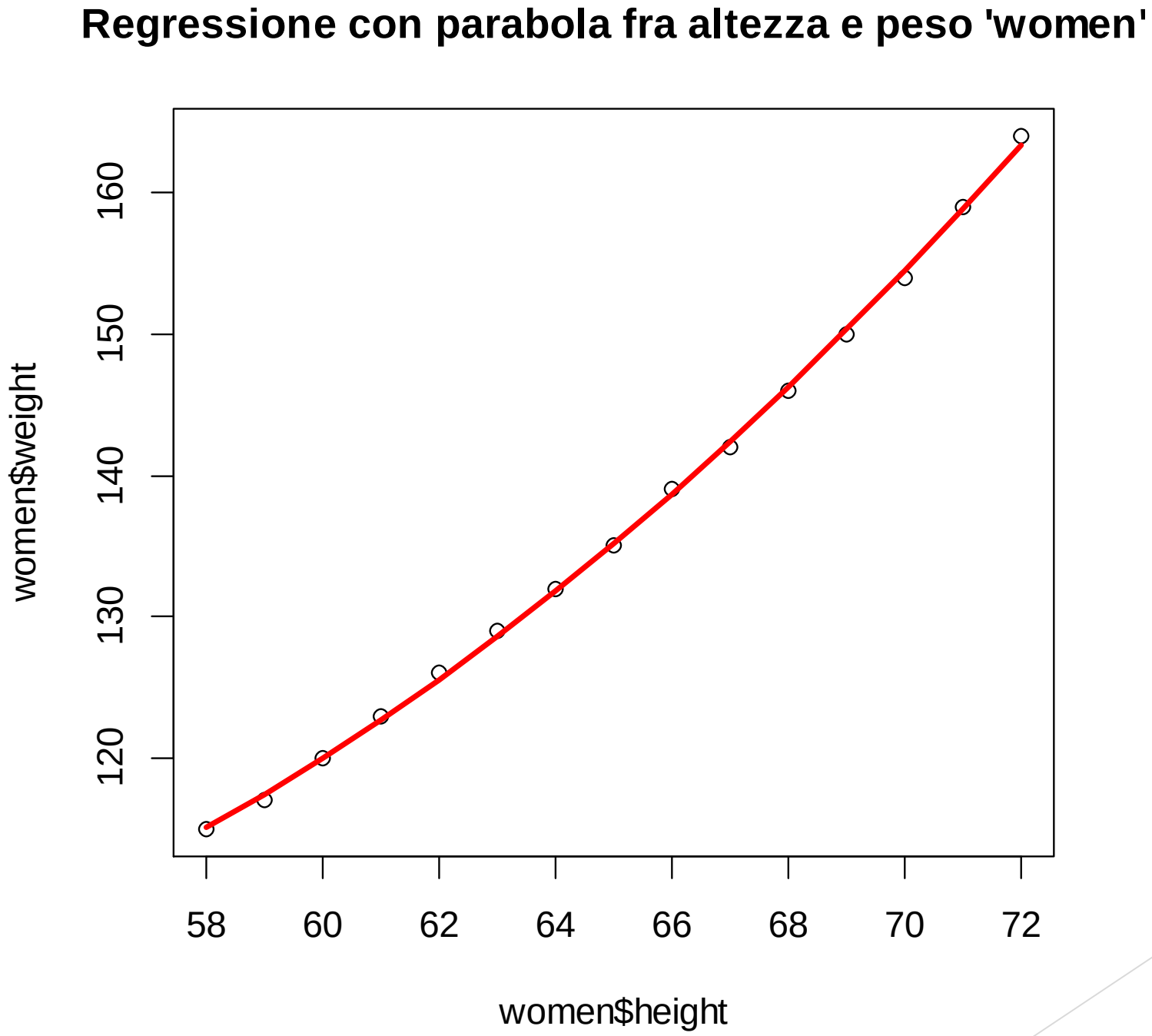
 # CALCOLIAMO IL COEFFICIENTE DI DETERMINAZIONE FACENDO IL QUADRATO DI R PER GIUDICARE LA BONTA' DI ACCOSTAMENTO:

```
> R2=R^2
> R2
[1] 0.9910098
```

 # DATO CHE R2 E' MOLTO VICINO A 1, DICIAMO CHE IL MODELLO TEORICO USATO SI ADATTA MOLTO BENE AI VALORI OSSERVATI
A TITOLO DI VERIFICA, ANCHE SE COME ABBIAMO VISTO DALL'ANALISI DEI RESIDUI SI PUO' FARE DI MEGLIO

ES. WOMEN – REGRESSIONE CON MODELLO PARABOLA
EXTRA – NON VERRA' CHIESTO ALL'ESAME!

```
> plot(women$height, women$weight)

# CREO UN MODELLO PARABOLICO DEL TIPO Y'=a+bX+cX^2
> height2=women$height^2
> parabola=lm(women$weight ~ women$height + height2)

# DISEGNO LA PARABOLA
> lines(women$height, fitted(parabola), col = "red", lwd = 3)
```


Regressione con parabola fra altezza e peso 'women'
women$weight
120
130
140
150
160
58
60
62
64
66
68
70
72
women$height

ES. WOMEN – REGRESSIONE CON MODELLO PARABOLA

```
> summary(parabola)

Call:
lm(formula = women$weight ~ women$height + height2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50941 -0.29611 -0.00941  0.28615  0.59706

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  261.87818   25.19677  10.393 2.36e-07 ***
women$height  -7.34832    0.77769  -9.449 6.58e-07 ***
height2        0.08306    0.00598  13.891 9.32e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3841 on 12 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9994
F-statistic: 1.139e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

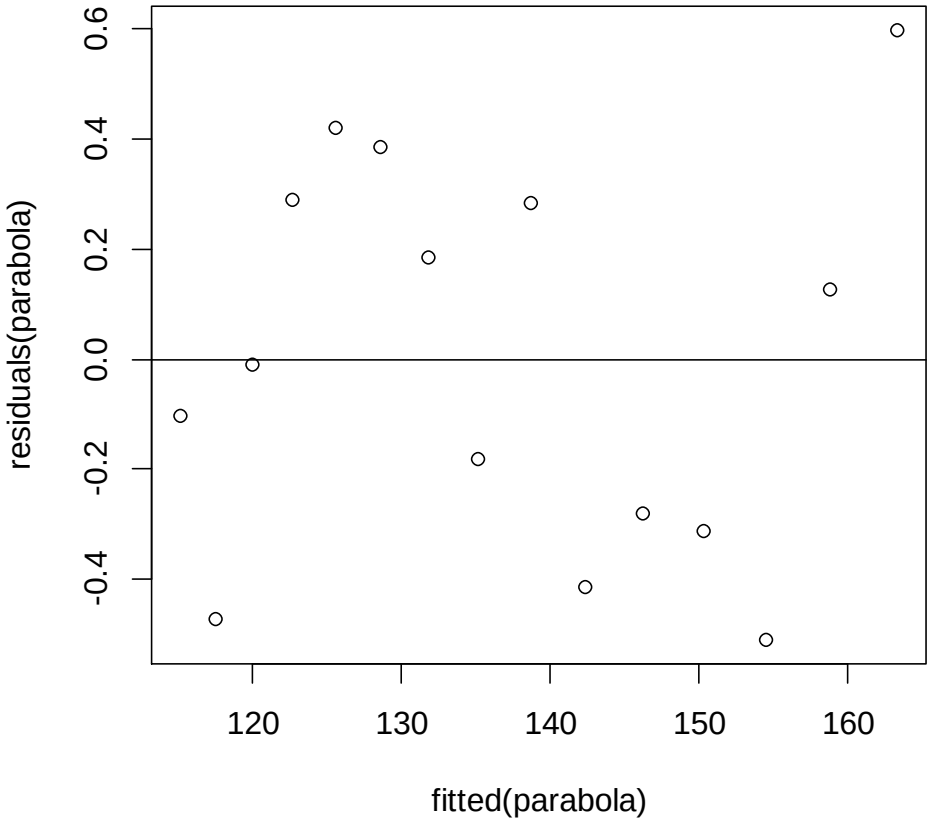
ES. WOMEN – REGRESSIONE CON MODELLO PARABOLA

 # IL MODELLO TEORICO SARA':

Y'= 261.87818 -7.34832 * height + 0.08306 * height^2

 # EFFETTUIAMO L'ANALISI DEI RESIDUI

```
> plot(fitted(parabola), residuals(parabola))
> abline(0, 0)
```



Ora l'analisi dei residui non ha più l'andamento parabolico di prima, ma rispetta la distribuzione casuale intorno alla retta 0 che ci si aspetta. S Si può quindi confermare l'ipotesi di distribuzione casuale degli stessi, con media nulla e incorrelazione.

ES. WOMEN – REGRESSIONE CON MODELLO PARABOLA

 # DATO CHE PER QUESTO CASO NON ABBIAMO UNA RETTA COME MODELLO TEORICO, IL CALCOLO DEL COEFFICIENTE DI CORRELAZIONE LINEARE NON HA MOLTO SENSO.

 # TUTTAVIA POSSIAMO VERIFICARE COME L'R2 SIA PIU' ALTO RISPETTO AL CASO PRECEDENTE DELLA RETTA (0.991), ARRIVANDO A 0.9995, COME EVIDENZIATO DALL'OUTPUT DELLA SUMMARY. QUESTO CONFERMA CHE IL MODELLO DELLA PARABOLA SI ADATTA MEGLIO AI DATI OSSERVATI RISPETTO AL MODELLO LINEARE, ARRIVANDO A ESSERE VICINISSIMO A 1 E QUINDI QUASI PERFETTAMENTE SOVRAPPOSTO ALLE OSSERVAZIONI Y.