Laboratorio di ST1 - Lezione 5

Antonietta di Salvatore

Dipartimento di Matematica
Universitá degli Studi Roma Tre

Correlazione e regressione lineare

Un modello di regressione lineare assume la seguente struttura

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + u_i$$

dove:
$i = 1, \ldots, n$
y_i é la variabile dipendente;
x_{ji} per $j = 2, \ldots, k$, sono le variabili indipendente o regressori;
β_1 é l'intercetta della retta di regressione della popolazione;
β_j per $j = 2, \ldots, k$, sono i coefficienti angolari dell'iperpiano di regressione;
u_i é l'errore statistico.

Esempio: i dato che seguono si riferiscono al numero di anni di patente, allo stipendio mensile e al prezzo pagato per una macchina di 15 individui.

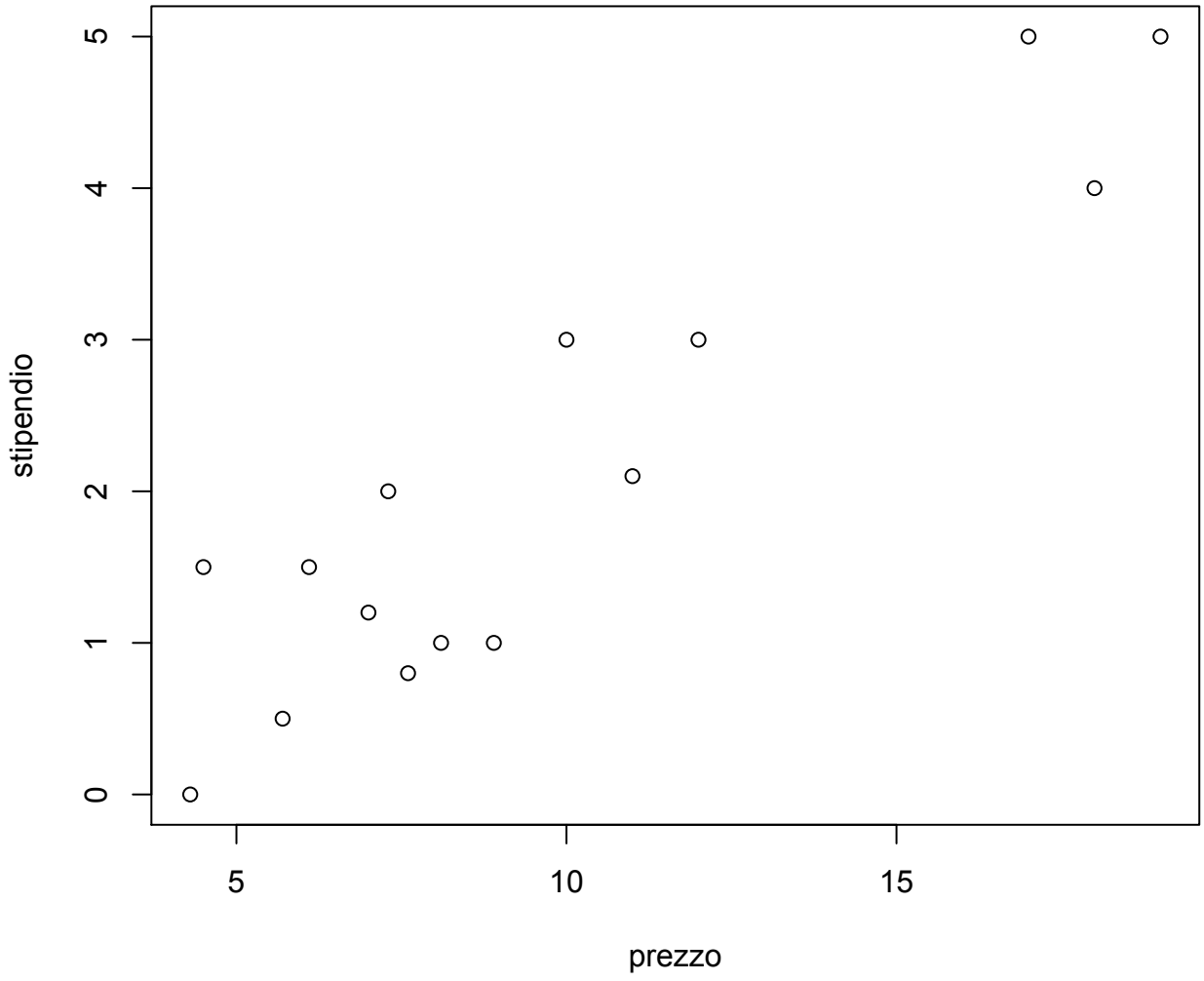
```
prezzo=c(18.0, 4.3, 4.5, 7.6, 7.0, 11.0, 17.0, 5.7, 8.1, 8.9,
7.3, 10.0, 19.0, 12.0, 6.1)
stipendio=c(4, 0, 1.5, 0.8, 1.2, 2.1, 5, 0.5, 1, 1, 2, 3, 5, 3,
1.5)
anni=c(6, 4, 0, 3, 3, 6, 2, 7, 6, 9, 2, 2, 4, 4, 0)
```

Verifichiamo la presenza di dipendenza lineare tra le variabili prezzo e stipendio

```
plot(prezzo,stipendio)
```

```
cor(stipendio,prezzo)
```

```
cor.test(stipendio,prezzo)
```


stipendio
0
1
2
3
4
5
5
10
15
prezzo

Costruiamo un modello lineare per le variabili prezzo e stipendio. Scelgo il prezzo come variabile dipendente.

La stima dei parametri β_j puó avvenre mediante il Metodo dei Minimi Quadrati.
In **R** usiamo la funzione `lm(...)`

REGRESSIONE SEMPLICE

```
?lm
modello1=lm(prezzo ~ stipendio)
summary(modello1)
```

intervallo di confidenza per le stime dei parametri

```
confint(modello1)
```

valori predetti

```
fitted(modello1)
```

residui

```
residuals(modello1)
```

per vedere i dati

```
modello1$model
dim(modello1$model)
modello1$rank
```

Rappresentazione grafica linea, predetti e residui.

```
plot(stipendio,prezzo)
```

N.B. per far funzionare i comandi `abline()` e `segments` é importante lórdine delle variabili nel plot iniziale

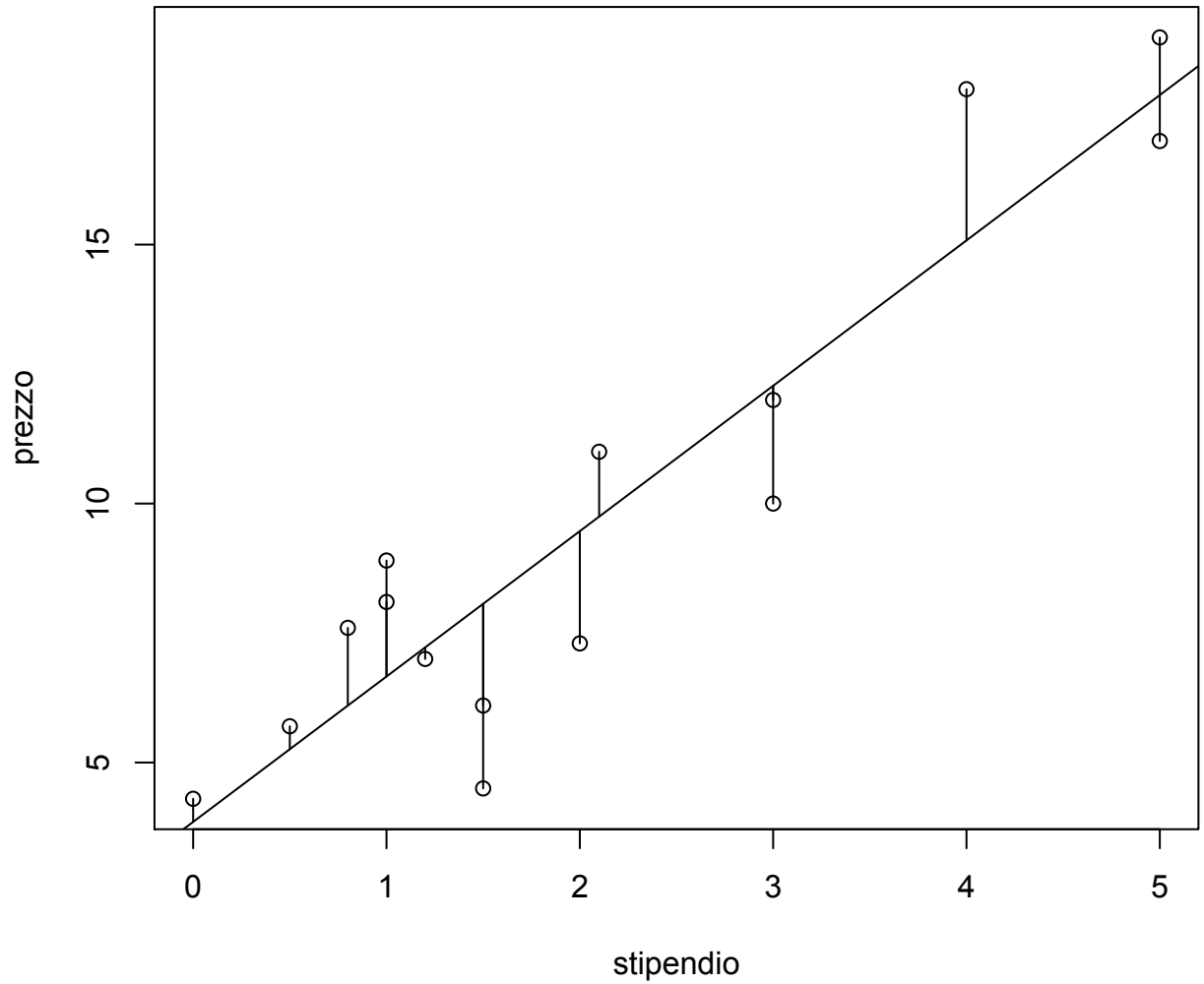
disegnamo la retta di regressione stimata, possiamo usare uno dei due seguenti comandi

```
abline(modello1)
```

```
lines(fitted(modello1),stipendio)
```

evidenziamo i residui

```
segments(stipendio, fitted(modello1), stipendio, prezzo)
```


prezzo
5
10
15
0
1
2
3
4
5
stipendio

Verifichiamo se esiste dipendenza lineare tra le variabili prezzo e anni

```
plot(anni,prezzo)
```

```
cor(anni,prezzo)
```

```
cor.test(anni,prezzo)
```

REGRESSIONE MULTIPLA

Inseriamo la variabile anni tra i regressori

```
modello2=lm(prezzo ~ stipendio + anni)
summary(modello2)
```

Quale modello é preferibile (spiega meglio la variabile dipendente) tra il modello1 e il modello2?

Consideriamo una ultima variabile da introdure nel modello

x=c(1.00, 2.10, 1.69, 1.69, 1.69, 2.61, 1.69, 2.39, 1.69, 1.69, 1.00, 2.39, 2.61, 1.69, 1.69)

```
plot(x,prezzo)
```

```
cor(anni,prezzo)
```

```
cor.test(x,prezzo)
```

```
modello3=lm(prezzo ~ stipendio + anni + x)
```

```
summary(modello3)
```

Dal confronto del modello2 con il modello3 cosa possiamo dire?
Osservare come variano gli indici `R-squared` e `Adjusted R-squared`.

I residui del modello.

```
sum(residuals(modello1))
```

```
sum(residuals(modello2))
```

```
sum(residuals(modello3))
```