R per l'analisi statistica multivariata

Unità L: regressione non lineare

Tommaso Rigon

Università Milano-Bicocca

Unità L

Argomenti affrontati

- Modelli linearizzabili
- Minimi quadrati non lineari
- Gli alberi di ciliegio nero

Riferimenti aggiuntivi

- **Unità K**, **Statistica I**: `https://tommasorigon.github.io/StatI/slides/sl_K.pdf`
- **Unità L**, **Statistica I**: `https://tommasorigon.github.io/StatI/slides/sl_L.pdf`
- Esercizi **R** associati: `https://tommasorigon.github.io/introR/exe/es_4.html`

Descrizione del problema

- Per $n = 31$ **alberi di ciliegio** nero sono disponibili le misure del diametro del tronco (misurato a circa 1m dal suolo) ed il volume ricavato dall'albero dopo l'abbattimento.
- Si vogliono utilizzare i dati per ottenere un'**equazione** che permetta di **prevedere** il volume, ottenibile solo dopo l'abbattimento dell'albero, avendo a disposizione il diametro, che è invece facilmente misurabile.
- In altri termini, stiamo cercando una qualche funzione $f(\cdot)$ tale che

$$(\text{volume}) \approx f(\text{diametro}).$$

- Una simile equazione ha differenti utilizzi.
- Ad esempio, può essere utilizzata per decidere quanti e quali alberi tagliare per ricavare un certo ammontare di legno, oppure per determinare il "prezzo" di un bosco.

I dati grezzi

Diametro

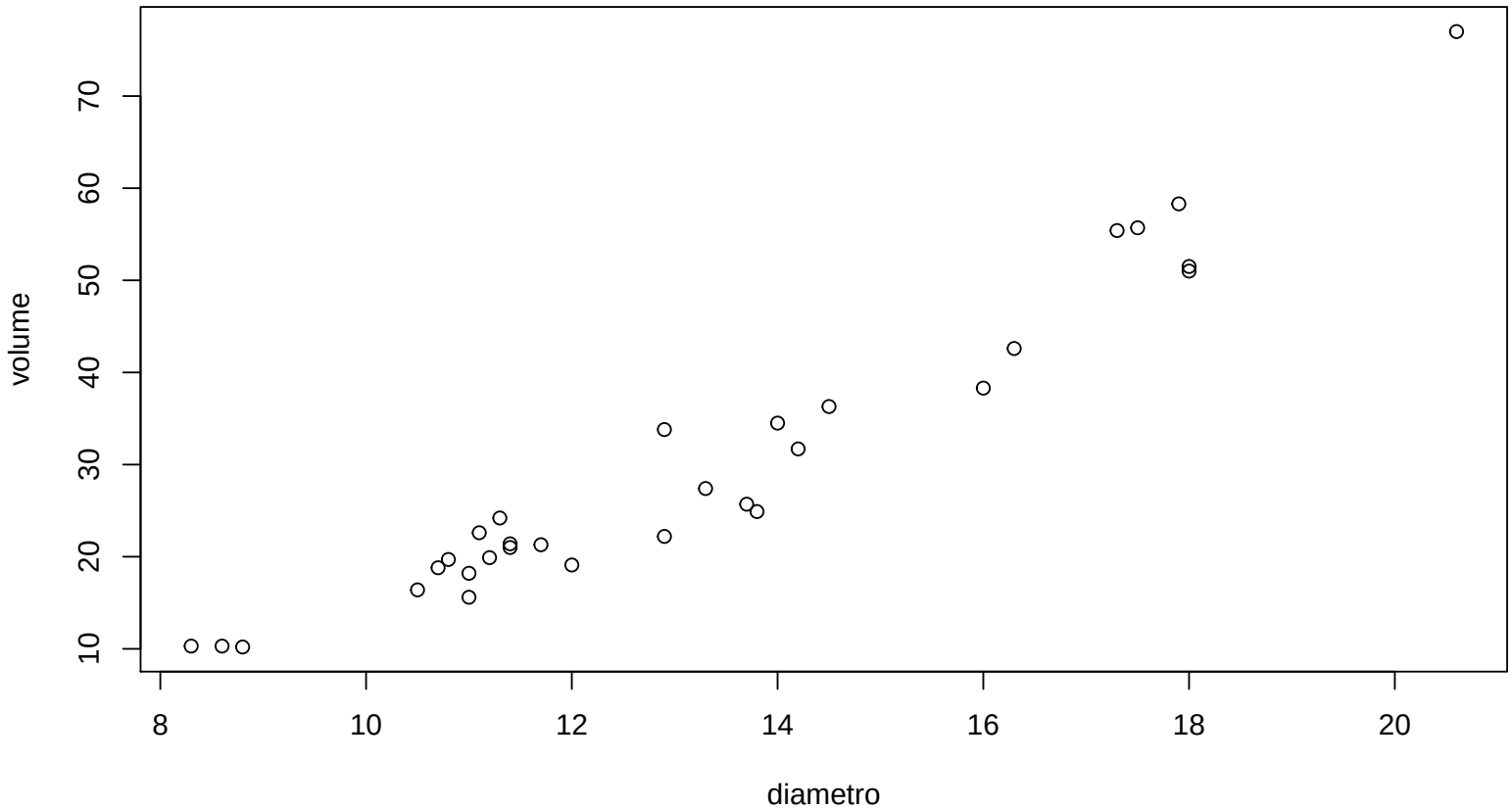
```
 [1]  8.3  8.6  8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 20.6 11.3
[13] 11.4 11.4 11.7 12.0 12.9 12.9 13.3 13.7 13.8 14.0 14.2 14.5
[25] 16.0 16.3 17.3 17.5 17.9 18.0 18.0
```

Volume

```
 [1] 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 77.0 24.2
[13] 21.0 21.4 21.3 19.1 22.2 33.8 27.4 25.7 24.9 34.5 31.7 36.3
[25] 38.3 42.6 55.4 55.7 58.3 51.5 51.0
```

```r
rm(list = ls())
ciliegi <- read.csv("https://tommasorigon.github.io/introR/data/ciliegi.csv", header = TRUE)
head(ciliegi)
#   diametro volume
# 1      8.3   10.3
# 2      8.6   10.3
# 3      8.8   10.2
# 4     10.5   16.4
# 5     10.7   18.8
# 6     10.8   19.7
```

Diagramma a dispersione




```
plot(ciliegi)
```

Alcune considerazioni geometriche (recap)

- Nelle **unità K** ed **L** del corso **Statistica I** abbiamo costruito dei modelli statistici del tipo (volume) $\approx f$(diametro) basati sulla geometria degli alberi.

- Dopo varie considerazioni di tipo geometrico, si era giunti ad una specificazione del tipo

$$(\text{volume}) = \eta\,(\text{diametro})^{\lambda},$$

  per due costanti **positive** $\eta, \lambda > 0$.

- Potremmo determinare i valori appropriati per η e λ utilizzando i **minimi quadrati**, ovvero considerando

$$(\hat{\eta}_{\text{ls}}, \hat{\lambda}_{\text{ls}}) = \arg\min_{\eta,\lambda} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \eta x_i^{\lambda}\right)^2.$$

- Purtroppo non esiste una **soluzione in forma chiusa** a questo problema, che infatti necessita dell'utilizzo di **tecniche numeriche**.

Minimi quadrati non-lineari

- La procedura di stima per $(\hat{\eta}_{\text{ls}}, \hat{\lambda}_{\text{ls}})$ prende il nome di **minimi quadrati non-lineari** e richiede una minimizzazione numerica, come quelle che abbiamo visto nell'**unità K**.

- Grazie ad **R** ed ai suoi strumenti computazionali, possiamo quindi svolgere un calcolo che nei corsi precedenti non era risolvibile. In particolare, possiamo usare `nlminb`.

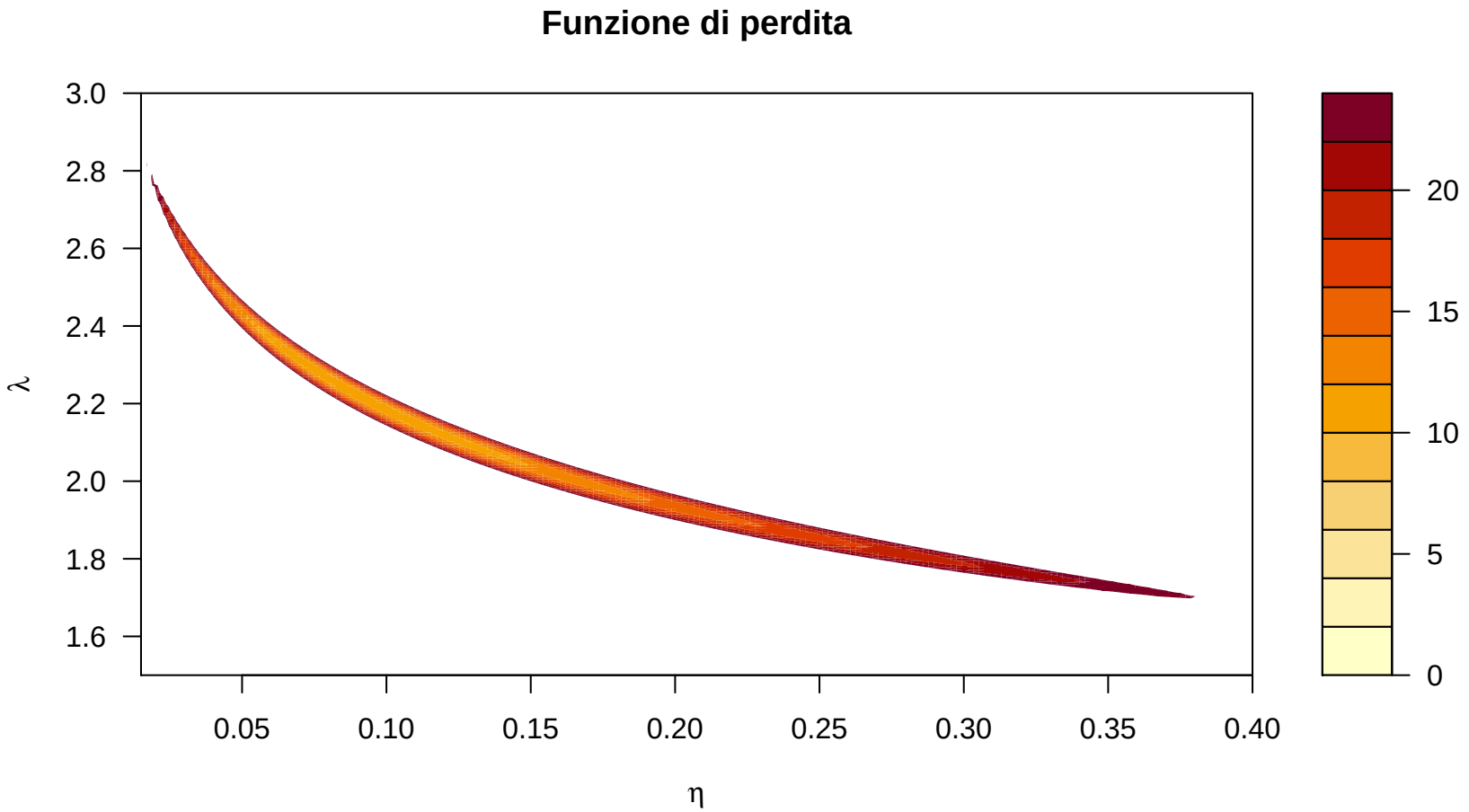
- In primo luogo, definiamo la **funzione obiettivo** o **funzione di perdita**:

```
# Funzione di perdita che vogliamo minimizzare
loss <- function(par, y, x) {
  mean((y - par[1] * x^par[2])^2)
}
```

- Ad esempio, tale funzione, valutata nel punto $(1, 1)$ vale circa 460.83, infatti:

```
loss(c(1, 1), ciliegi$volume, ciliegi$diametro)
# [1] 460.8329
```

La funzione di perdita



- **Esercizio**. Si riproduca il grafico di questa slide.

Stima ai minimi quadrati

- La stima ai minimi quadrati si ottiene quindi usando `nlminb`. In questo caso, siamo effettivamente interessati a **minimizzare** una funzione.

```
fit_ls <- nlminb(start = c(1, 1), function(param) loss(param, ciliegi$volume, ciliegi$diametro),
                 lower = c(1e-6, 1e-6))
fit_ls
# $par
# [1] 0.08661007 2.23638534
#
# $objective
# [1] 10.12108
#
# $convergence
# [1] 0
#
# $iterations
# [1] 27
#
# $evaluations
# function gradient
#       41       61
#
# $message
# [1] "relative convergence (4)"

# Salvo i risultati
param_hat_ls <- fit_ls$par
```

Commenti ai risultati

- I comandi precedenti quindi implicano che la **stima ai minimi quadrati** (**non lineari**) è pari a

$$\hat{\eta}_{\text{ls}} = 0.0866, \qquad \hat{\lambda}_{\text{ls}} = 2.2364.$$

- Inoltre, la **varianza residuale** è pari a

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\eta}_{\text{ls}} x_i^{\hat{\lambda}_{\text{ls}}}\right)^2 = 10.121.$$

- Come ricorderete, questo problema di stima può essere affrontato alternativamente tramite la procedura di **linearizzazione** del modello.

- Supponendo che la relazione sia del tipo $(\text{volume}) = \eta\,(\text{diametro})^{\lambda}$, allora applicando la funzione log ambo i lati, si ottiene

$$\log(\text{volume}) = \log\eta + \lambda \log(\text{diametro}).$$

- Quindi, la **relazione non lineare** che abbiamo supposto tra `diametro` e `volume` corrisponde ad una **relazione lineare** tra i logaritmi delle due variabili.

Il modello linearizzato

- La relazione in scala logaritmica descrive un modello **linearizzato**. Si tratta di un modello di regressione lineare semplice in cui

$$z_i = \log y_i, \qquad w_i = \log x_i, \qquad i = 1, \ldots, n.$$

- Introducendo esplicitamente il termine di errore, avremo quindi che

$$z_i = \alpha + \beta w_i + \epsilon_i$$

in cui $\alpha = \log \eta$ e $\beta = \lambda$.

- Possiamo determinare i **parametri trasformati** ottimali $\hat{\alpha}$ e $\hat{\beta}$ ed **parametri originali** $\hat{\eta}_{\text{ols}}$ ed $\hat{\lambda}_{\text{ols}}$ utilizzando il criterio dei minimi quadrati sulla scala trasformata, ovvero

$$\min_{\alpha,\beta} \frac{1}{n} \sum_{i=1}^{n} (z_i - \alpha - \beta w_i)^2 = \min_{\eta,\lambda} \frac{1}{n} \sum_{i=1}^{n} (\log y_i - \log \eta - \lambda \log x_i)^2 .$$

- Varrà quindi la relazione $\hat{\eta}_{\text{ols}} = \exp\{\hat{\alpha}\}$ e che $\hat{\lambda}_{\text{ols}} = \hat{\beta}$.

Stima ai minimi quadrati (modello linearizzato)

- La stima ai minimi quadrati in scala trasformata ammette una **soluzione esplicita**.

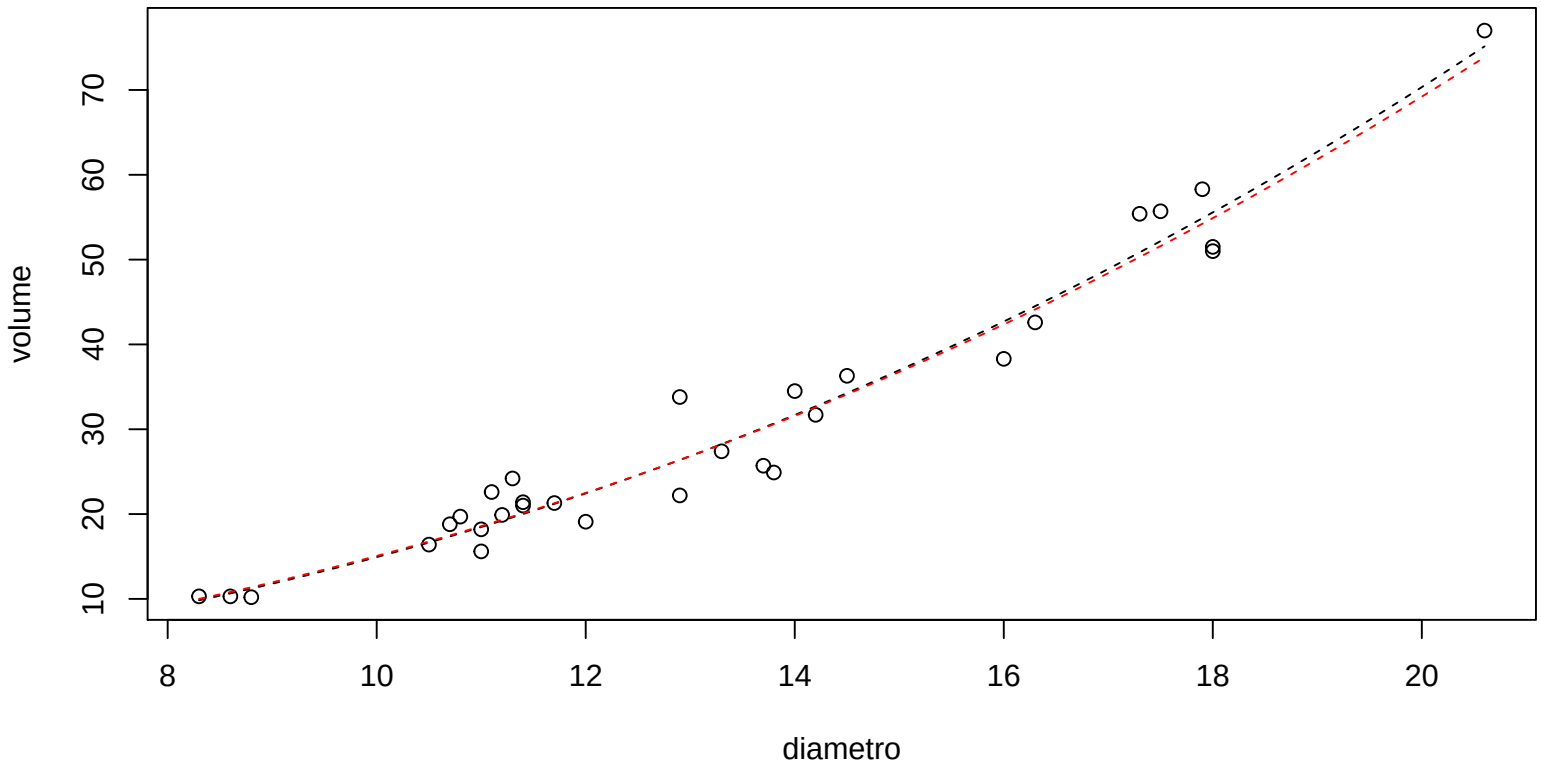
```
z <- log(ciliegi$volume)
w <- log(ciliegi$diametro)

beta_hat_ols <- cov(w, z) / var(w)
alpha_hat_ols <- mean(z) - mean(w) * beta_hat_ols

# Stima ai minimi quadrati, scala trasformata
param_hat_ols <- c(exp(alpha_hat_ols), beta_hat_ols)
param_hat_ols
# [1] 0.09505259 2.19996993

# Varianza residuale
loss(param_hat_ols, ciliegi$volume, ciliegi$diametro)
# [1] 10.2531
```

- I comandi precedenti quindi implicano che la **stima ai minimi quadrati** (**modello linearizzato**) è pari a

$$\hat{\eta}_{\text{ols}} = 0.095, \qquad \hat{\lambda}_{\text{ols}} = 2.200.$$

- Inoltre, la varianza residuale $n^{-1} \sum_{i=1}^{n} (y_i - \hat{\eta}_{\text{ols}} x_i^{\hat{\lambda}_{\text{ols}}})^2 = 10.253$ è superiore a quella ottenuta in precedenza (**come mai**?), anche se di poco.

Confronto tra modelli




```
plot(ciliegi)
curve(param_hat_ls[1] * x^param_hat_ls[2], add = TRUE, lty = "dashed") # Non-lineari
curve(param_hat_ols[1] * x^param_hat_ols[2], add = TRUE, lty = "dashed", col = "red") #Modello linearizzato
```

Commenti conclusivi

- I due approcci sono **sostanzialmente equivalenti** in questo specifico esempio, nel senso che producono risultati quasi indistinguibili.
- Si noti che il **modello** è lo stesso, abbiamo solo cambiato **metodo di stima**!
- Tuttavia, in generale non è detto che sia possibile linearizzare il modello originale. In questi casi, non esiste un'alternativa semplice.
- Infine, a seconda della funzione di perdita utilizzata, due diversi metodi di stima potrebbero differire di molto nonostante il modello sia lo stesso.
- Ad esempio, alcuni stimatori sono più **robusti** di altri rispetto alla presenza di outlier.