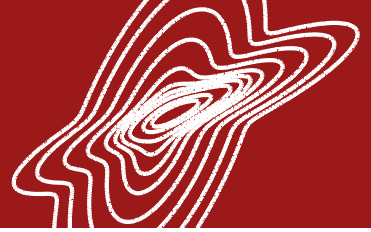


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

METODI STATISTICI PER LA BIOINGEGNERIA

Laboratorio 9

A.A. 2024-2025

Enrico Longato

Dal lab 8: Ciclo for escludendo un indice

```
[1 2 3 4 5 6 7]
```

```
[  2 3 4 5 6 7]  [1]
[1   3 4 5 6 7]  [2]
[1 2   4 5 6 7]  [3]
[1 2 3   5 6 7]  [4]
[1 2 3 4   6 7]  [5]
[1 2 3 4 5   7]  [6]
[1 2 3 4 5 6  ]  [7]
```

Traduco in sequenze

```
[1:7]
```

```
[1:0   2:7]  [1]
[1:1   3:7]  [2]
[1:2   4:7]  [3]
[1:3   5:7]  [4]
[1:4   6:7]  [5]
[1:5   7:7]  [6]
[1:6   8:7]  [7]
```

Queste sono "ovvie": sono le **sequenze che si fermano "un numero prima" e riprendono "un numero dopo"** di quello che vogliamo escludere

Queste due, invece, sfruttano una proprietà di MATLAB per cui, se l'inizio di una sequenza con passo 1 (implicito) è maggiore del valore di fine, allora la sequenza è vuota, quindi `1:0 == 8:7 == []`

In generale (quello che serve per un ciclo `for j=1:N_variabili`)

```
[1 … j … end]                          [1:end]
[1 … (j-1) (j+1) … end] [j]    [1:(j-1)  (j+1):end] [j]
```

Dal lab 8: Interpretazione test sui $\hat{\beta}$

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    ________    ________    __________

    (Intercept)       -46.002      19.282     -2.3857     0.017809
    x1                0.17176    0.032771      5.2413   3.4512e-07
    x2              -0.037151    0.059654     -0.62278      0.53401
    x3                -0.2479      0.1186     -2.0902     0.037638
    x4               -0.64999     0.26926      -2.414     0.016514
    x5                0.56882     0.10703      5.3145   2.412e-07
    x6                0.21433     0.17378      1.2334      0.21863
    x7                0.46143     0.16682      2.7661    0.0061075
```

Per tutti (intercetta inclusa)

- Se il p value è >0.05 => non posso dire nulla
- Se il p value è <0.05 => il coefficiente è significativamente diverso da zero

Per i coefficienti che <u>non</u> sono l'intercetta, <u>inoltre:</u>

- Se il p value è >0.05 => non posso dire nulla (come prima)
- Se il p value è <0.05 => posso dire che la variabile corrispondente è significativamente associata alla variabile dipendente*

** nel senso della correlazione lineare di Pearson e solo se tutte le variabili sono indipendenti (non solo non collineari)*
[BONUS: mi piaceva scrivere, per una volta, l'interpretazione corretta; per l'esame non serve]

Funzione `lasso` (consultare l'help per maggiori informazioni)

Traccia di utilizzo (in rosso il codice; in nero il "testo libero")

```
lambda_range =  un vettore di candidati lambda
[beta_hat, info] = lasso(X_no_intercept, Y, ...
        'Intercept', true, ...  -- la solita intercetta
        'Standardize', true, ...  -- per standardizzare prima di regolarizzare (v. teoria)
        'Alpha',  1 vuol dire "L1", un numero molto vicino a 0, "L2",  ...
        'lambda', lambda_range, ...
        'CV',  numero di fold della K-fold CV)
```

0.2227	0.2226	0.2225	0.2223	0.2220
0.0302	0.0300	0.0296	0.0292	0.0286
-0.1856	-0.1855	-0.1854	-0.1852	-0.1850
-0.4001	-0.3991	-0.3977	-0.3958	-0.3934
0.4901	0.4903	0.4905	0.4907	0.4910
0.1917	0.1919	0.1922	0.1927	0.1933
0.4531	0.4527	0.4521	0.4514	0.4504
-0.1642	-0.1629	-0.1612	-0.1590	-0.1560
-0.0514	-0.0512	-0.0508	-0.0504	-0.0497
-0.0401	-0.0393	-0.0384	-0.0371	-0.0354
0.3332	0.3325	0.3315	0.3301	0.3284
-2.4960	-2.4952	-2.4942	-2.4929	-2.4912

$\hat{\beta}$ per il quarto valore in lambda_range

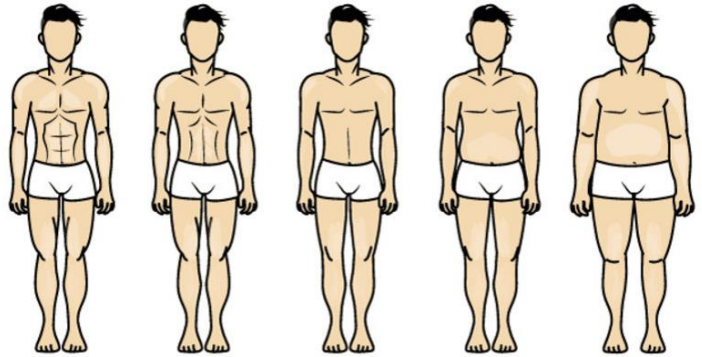
1x1 struct with 12 fields

Field	Value
Intercept	*1x50 double*
Lambda	*1x50 double*
Alpha	1
DF	*1x50 double*
MSE	*1x50 double*
PredictorNames	*0x0 cell*
UseCovariance	1
SE	*1x50 double*
LambdaMinMSE	0.2121
Lambda1SE	0.4942
IndexMinMSE	20
Index1SE	23

NB: come si intuisce dalla presenza del parametro `Alpha` e dal fatto che, se questo vale 1, stiamo regolarizzando L1, mentre, se vale circa 0, stiamo regolarizzando L2, la funzione `lasso`, in realtà, implementa il metodo elastic net.

CONTESTO DELL'ESERCITAZIONE E DATI (esteso rispetto al lab 7)

Dataset di misure antropometriche per la predizione della % di grasso corporeo (`bodyfat_extended.mat`).



Dati di 252 uomini descritti da 14 variabili

1. BodyFat (in %) sarà la nostra variabile dipendente
2. Age (età in anni, years)
3. Weight (peso in libbre, lbs)
4. Height (altezza in pollici, inches)
5. Neck (circonferenza del collo in cm)
6. Chest (circonferenza del petto in cm)
7. Hip (circonferenza dei fianchi in cm)
8. Thigh (circonferenza della coscia in cm)
9. Knee (circonferenza del ginocchio in cm)
10. Ankle (circonferenza della caviglia in cm)
11. Bicep (circonferenza del bicipite in cm)
12. Forearm (circonferenza dell'avambraccio in cm)
13. Wrist (circonferenza del polso in cm)

Adattato da https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset

ESERCIZIO 1 - PARTE 1: REGRESSIONE LINEARE (proposto)

- Caricare i dati
- Considerare la prima variabile (**`BodyFat`**) come variabile dipendente e le altre 13 come variabili indipendenti.
- Usare la funzione **`fitlm`** per effettuare la stima "automatica" dei parametri del modello di regressione
  - Bonus: confrontare con i risultati ottenuti a mano

ESERCIZIO 1 - PARTE 2: REGRESSIONE LINEARE REGOLARIZZATA L1 (svolto)

- Stimare i parametri di una regressione regolarizzata L1 sugli stessi dati (**`lasso`**)
  - Lo spazio di ricerca per lambda è equispaziato logaritmicamente tra 10^{-3} e 10^{3} e consta di 50 possibili candidati lambda (**`logspace`**)
  - Considerare il lambda ottimo come quello che minimizza l'MSE medio su una K-fold cross-validation con $K = 5$ fold (**`'CV', 5`**)
  - Come da teoria, standardizzare i coefficienti prima di regolarizzare (**`'Standardize', true`**)
  - NB: per fare la regressione L1, bisogna utilizzare il name-value pair **`'Alpha', 1`**
- Dire quanti parametri diversi da 0 rimangono in corrispondenza del valore ottimo di lambda

ESERCIZIO 1 - PARTE 3: REGRESSIONE LINEARE REGOLARIZZATA L2 (proposto)

- Ripetere il punto 2 (sempre funzione **`lasso`**), ma con la regolarizzazione L2
  - Bisogna mettere un valore "piccolo" per **`Alpha`**, diciamo 10^{-32}

ESERCIZIO 1 - PARTE 4: CONFRONTO TRA MODELLI (proposto)

- Calcolare AIC, BIC e R^2 adjusted per i modelli non regolarizzato, regolarizzato L1 e regolarizzato L2.

- Confrontarli e trarre le conclusioni del caso

Suggerimento: il "problema", qui, è calcolare la predizione $\hat{Y}$ nei diversi casi, perché `fitlm` e `lasso` danno oggetti in uscita leggermente diversi.

In particolare:

- per `fitlm` invocato con `'Intercept', true`, l'intercetta è il primo elemento del vettore dei beta stimati; gli altri beta stimati seguono nello stesso vettore
  - `Y_hat = X_no_intercept*beta_hat(2:end) + beta_hat(1)` (con la solita nomenclatura)
- per `lasso` invocato con `'Intercept', true`, l'intercetta è in un campo del secondo argomento di uscita; gli altri beta stimati sono nella colonna corrispondente al lambda ottimo del primo argomento di uscita
  - `Y_hat = X_no_intercept*beta_hat_best + intercept_best`
    - dove `beta_hat_best` e `intercept_best` si trovano così (v. spiegazione in aula; v. soluzioni)

```
[mse_min, i_lambda_best] = min(info.MSE);
lambda_best = lambda_range(i_lambda_best);
beta_hat_best = beta_hat(:, i_lambda_best);
intercept_best = info.Intercept(i_lambda_best);
```