UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models and R

2627-2-E4103B010-E4103B010-1

Obiettivi formativi

L'obiettivo del corso è duplice: richiamare concetti di programmazione in R e fornire gli strumenti statistici fondamentali per analizzare congiuntamente più variabili misurate sullo stesso insieme di unità statistiche. Al termine del corso, le competenze acquisite permetteranno a studenti e studentesse di:

- utilizzare il linguaggio di programmazione R per condurre analisi dei dati univariati e multivariati
- leggere, comprendere, scrivere ed eseguire script in R, sia propri che prodotti da altri;
- analizzare le relazioni tra variabili mediante modelli di regressione lineare, sia da un punto di vista teorico sia attraverso l'applicazione pratica in R.

Contenuti sintetici

Il corso fornisce un'introduzione al linguaggio di programmazione R, includendo elementi di programmazione, statistica descrittiva, probabilità e inferenza statistica. La seconda parte del corso è dedicata alla specificazione, stima, verifica e diagnostica di modelli statistici lineari.

Programma esteso

Prima parte:

- **Richiami di programmazione in R:** introduzione agli oggetti e alle loro classi (vettori, booleani, matrici, data.frame, liste); strutture condizionali e cicli; funzioni; rivisitazione dei principali argomenti di statistica descrittiva univariata e bivariata applicata a dataset, incluse rappresentazioni grafiche e loro personalizzazioni; principali funzioni per lavorare con variabili casuali.

Seconda parte:

- **Variabili casuali multidimensionali:** funzione di densità e di ripartizione congiunte; marginalizzazione; distribuzioni condizionate; momenti marginali e condizionati; vettore delle medie e matrice di varianze e covarianze; v.c. Normale multivariata e sue proprietà.

- **Specificazione del modello:** fasi per la specificazione di un modello statistico; classificazione dei modelli.

- **Modello di regressione lineare semplice:** assunzioni; interpretazione dei parametri; stima dei parametri (minimi quadrati e massima verosimiglianza); proprietà degli stimatori; teorema di Gauss-Markov; indice di determinazione lineare R^2.

- **Verifica ed utilizzo del modello:** verifiche di ipotesi sul valore di un singolo coefficiente; verifica di ipotesi sulla bontà del modello; utilizzo del modello per fare previsione puntuale ed intervallare.

- **Modelli polinomiali:** definizione di un modello di regressione polinomiale; determinazione del grado massimo di un polinomio di regressione e relativa definizione della funzione di regressione; scomposizione della devianza residua in devianza residua riducibile ed irriducibile.

- **Diagnostica del modello:** metodi per valutare le assunzioni relative alla struttura del modello, agli errori e all'assenza di osservazioni inusuali.

- **Modello di regressione lineare multiplo:** specificazione del modello in forma matriciale e sue assunzioni; interpretazione dei parametri; stima dei parametri (minimi quadrati e massima verosimiglianza); proprietà degli stimatori; teorema di Gauss-Markov; indice di determinazione multipla R^2.

- **Verifica di un sistema di ipotesi lineare:** teoria generale e casi particolari.

- **Variabili qualitative:** inserimento di variabili qualitative nel modello mediante l'inserimento di dummy; interazioni.

- **Multicollinearità:** approfondimento sulla multicollinearità e definizione di indici adatti a rilevarne gli impatti

- **Modelli di regressione vincolati:** inserimento di vincoli lineari (puntuali e matriciali) in un modello di regressione. Dimostrazione della forma dello stimatore dei minimi quadrati vincolato e relative proprietà inferenziali.

- **Selezione del modello:** contributo assoluto e relativo di una variabile esplicativa; indice di determinazione parziale (IDP); selezione criterion-based delle variabili esplicative tramite approccio backward, forward e stepwise; AIC e BIC.

- **Casi particolari:** Utilizzo del modello lineare per la definizione di metodologie standard, quali il test T per l'uguaglianza della media di due popolazioni ed il test ANOVA.

Prerequisiti

Questa attività formativa deve essere preceduta dal superamento degli esami di Algebra lineare, Analisi Matematica I, Calcolo delle Probabilità e Statistica I. Si consiglia inoltre la conoscenza degli argomenti trattati nel corso di Statistica II.

Metodi didattici

Il corso è erogato in italiano e prevede lezioni frontali sia in aula sia in laboratorio informatico.
Le lezioni in aula sono mirate all'approfondimento delle conoscenze teoriche dello studente sugli argomenti del Corso ed alla loro formalizzazione. Nelle lezioni svolte in laboratorio informatico si trattano gli aspetti di implementazione dei modelli su dati reali e simulati utilizzando il software R.

In particolare, **Modelli Statistici e R** prevede un totale di **73 ore** di lezioni frontali svolte in modalità erogativa in presenza, ciascuna formata da blocchi da 2 o 3 ore, molte delle quali saranno svolte in un'aula informatica. In aggiunta, saranno erogate delle attività di tutorato a supporto di studenti e studentesse.

Modalità di verifica dell'apprendimento

Per superare l'insegnamento di Analisi Statistica Multivariata è necessario ottenere una valutazione maggiore o uguale a 18 in entrambe le parti che compongono l'insegnamento (Modelli Statistici e R (9 CFU) ed Analisi Esplorativa (6 CFU)). Il voto finale è determinato dalla media (ponderata con i rispettivi CFU) dei voti riportati nelle prove parziali.

Per la parte di **Modelli Statistici e R (9 CFU)**:

- la prova è scritta e articolata in 3 o 4 esercizi. Tali esercizi includono domande teoriche, esercizi di programmazione, classici esercizi scritti e/o analisi di dati reali e prevede anche l'uso di R.
- Sono previste due prove in itinere: la prima (sessione invernale) riguardante la programmazione in R e la seconda riguardante i modelli lineari (sessione primaverile). Durante ogni altra sessione, la prova sarà formata esclusivamente dalla prova intera.
- Durante la prova non è ammesso l'uso di testi o altro materiale, con l'esclusione dei codici che verranno messi a disposizione dal docente all'inizio della prova.
- Studenti e studentesse, così come il docente, possono richiedere (a seguito di una valutazione sufficiente nella prova scritta) una prova orale facoltativa, riguardante l'intero programma da 9 CFU.

Testi di riferimento

Prima parte: Modelli Statistici e R (9 CFU)

- Appunti delle lezioni forniti dal docente
- Albert, J. & M. Rizzo (2012). *R by Example*. Springer.
- Stefano Leonardi. Impariamo R - Un'introduzione facile al linguaggio. Città Studi edizioni.
- Venables, W. N., Smith D. M. & the R Core Team (2021). *An Introduction to R*.
- M. Grigoletto, F. Pauli, L. Ventura, Modello lineare, teoria e applicazioni con R. Giappichelli, 2017
- J. Fox. Applied regression analysis and generalized linear models, third edition. Sage.
- Piccolo, D. (2010), Statistica, Terza edizione, Il Mulino.

Periodo di erogazione dell'insegnamento

- **Modelli Statistici e R (9 CFU) :** II periodo del I semestre; I periodo del II semestre

Lingua di insegnamento

Italiano

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ