UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models

2425-2-E4102B084-E4102B085M

Obiettivi formativi

Il modulo di modelli statistici intende sviluppare le conoscenze teoriche e applicative circa i modelli lineare classico, di regressione lineare multipla e di regressione logistica multipla.

Conoscenza e comprensione
Lo studente viene introdotto ai concetti sottostanti i modelli statistici e le relative assunzioni. Impara ad utilizzare i modelli attraverso il loro impiego con dati reali e simulati. Impara ad interpretare i risultati e a verificare la sostenibilità del modello. Vengono trattati aspetti di analisi grafica, e analisi computazionale utilizzando la notazione matriciale.
Capacità di applicare conoscenza e comprensione
Il corso sviluppa le competenze per l'analisi dei dati aventi natura multivariata e provenienti da varie fonti informative: contesti aziendali, economici, biologici, fisici, medici, astronomici, ambientali, sociali e sportivi. Lo studente approfondisce le competenze nell'utilizzo della semantica del software R per le analisi di statistica descrittiva multivariata e per l'applicazione dei modelli di regressione lineare multipla, lineare classico, di regressione logistica. Lo studente impara a creare dei report dove illustra le analisi effettuate e commenta i risultati ottenuti.
Il corso permette allo studente di acquisire gli elementi di base di teoria e di applicazione dei modelli statistici e si qualifica come indispensabile sia per il successivo percorso universitario di formazione professionale nella scienza dei dati che per eventuali contesti lavorativi.

Contenuti sintetici

Il corso è suddiviso in quattro parti:

1. IL MODELLO STATISTICO. Nell'introduzione viene spiegato il concetto di analisi statistica e di modello statistico Vengono poi introdotti alcuni concetti fondamentali necessari per lo svolgimento del corso

2. REGRESSIONE LINEARE MULTIPLA. Vengono analizzate le metodologie per risolvere una regressione multipla descrittiva. In particolare vengono analizzati il metodo di stima dei minimi quadrati, i criteri di bontà di adattamento e per la scelta di un modello di regressione rispetto a un altro. Si valutano la presenza di multicollinearità ed outliers che possono inficiare la bontà di un'analisi.
3. IL MODELLO LINEARE CLASSICO In chiave inferenziale viene proposto il modello lineare classico con le sue sei assunzioni e l'ipotesi di normalità degli errori. Si verifica, come gli stimatori campionari dei minimi quadrati posseggono proprietà ottimali per studiare i parametri della popolazione . Inoltre, sotto l'ipotesi di normalità degli errori si costruiscono test di ipotesi sulla significatività dei singoli parametri, e relativi intervalli di confidenza . Si descrivono a questo punto i criteri per la scelta dei modelli tra i diversi possibili e l'uso dei modelli a fini esplicativi
   4)ESTENSIONE DEL MODELLO LINEARE CLASSICO SI propone il modello con variabili esplicative qualitative e miste, le principali trasformazioni delle variabili e si analizza in particolare il modello di regressione logistica

Programma esteso

Lo studente viene introdotto allo studio della regressione multipla e del modello lineare classico nonché delle principali trasformazioni di variabili ad esso inerente quali il modello logistico. Ciò avviene non solo in chiave teorica ma anche applicativa con grandissima attenzione alla risoluzione di problemi basati su dati reali e simulati mediante R.
Il corso è suddiviso in quattro parti:

1. IL MODELLO STATISTICO Nell'introduzione viene spiegato il concetto di analisi statistica e di modello statistico nelle sue differenze con il modello matematico. Vengono quindi analizzate le differenze tra modelli descrittivi e inferenziali, causazione e associazione, modello vero e proprio e machine learning. Vengono poi introdotti alcuni concetti fondamentali: la classificazione delle diverse tipologie di caratteri, le più elementari rappresentazioni grafiche, la rappresentazione matriciale dei dati, il significato e la formulazione dell'indice di correlazione tra caratteri quantitativi. Infine, sono accennate le fasi della costruzione di un modello: costruzione delle ipotesi, specificazione, raccolta dati, stima dei parametri, criteri della bontà di adattamento, previsione.
2. REGRESSIONE LINEARE MULTIPLA Vengono analizzate in profondità le fasi della costruzione di un modello a partire dalla sua specificazione In particolare si studia la regressione lineare multipla e la stima dei suoi parametri, dei criteri della bontà di adattamento, della previsione Viene introdotta la funzione di regressione lineare multipla anche in termini matriciali e si esplicitano le assunzioni sottostanti. In particolare vengono analizzati il metodo di stima dei minimi quadrati e i criteri di bontà di adattamento e per la scelta di un modello di regressione ad esso inerente. Si valutano la presenza di multicollinearità ed outlier che possono inficiare la bontà di una analisi.
3. IL MODELLO LINEARE CLASSICO In chiave inferenziale viene proposto il modello lineare classico con le sue sei assunzioni e il loro significato. Successivamente dopo aver illustrato la distribuzione di Gauss bivariata e multivariata e le relative proprietà viene presentata l'ipotesi di normalità degli errori. Si propone a questo punto il modello lineare classico in chiave campionaria e ci si chiede come le stime campionarie possono darci informazioni adeguate degli ignoti valori della popolazione Si verifica, a questo punto che gli stimatori dei minimi quadrati posseggono le proprietà ottimali della correttezza, consistenza efficienza. Inoltre sotto l'ipotesi di normalità degli errori permettono di costruire test di ipotesi sulla significatività dei singoli parametri, di gruppi, di essi, del modello nel suo complesso. Si possono infine costruire intervalli di confidenza per verificare con che probabilità ed entro che limiti giacciono i veri valori dei parametri nella popolazione . Si descrivono a questo punto i criteri per la scelta dei modelli tra i diversi possibili Infine si analizza il significato dei modelli a fini previsivi mostrando le differenze rispetto dopo uso dei modelli a fini esplicativi
4. ESTENSIONE DEL MODELLO LINEARE CLASSICO SI propone il modello con variabili esplicative qualitative e miste. Si presentano le principali trasformazioni delle variabile riposta ed esplicative possibili e necessarie in particolari contesti suggeriti dalle esigenze di ricerca e dalla natura dei dati. Si analizza in

particolare il modello di regressione logistica generale con i suoi metodi di stima dei parametri e di incertezza associata alla stima tramite gli errori standard, l'interpretazione dei coefficienti stimati e l'utilizzo del modello per scopi fini previsionali, il concetto e la misura di odds e odds ratio

Prerequisiti

Si richiede di aver superato gli esami degli insegnamenti propedeutici: Statistica I, Analisi Matematica I, Algebra Lineare, Calcolo delle Probabilità. Per una più agevole comprensione dei contenuti del corso è fortemente consigliato conoscere le nozioni di inferenza statistica impartite al corso di Statistica II

Metodi didattici

Tutte le lezioni si svolgono in laboratorio informatico: Sono previste lezioni frontali riguardanti la parte di teoria. Già in quest'ambito la parte di teoria viene affiancata allo sviluppo di applicazioni che riguardano dati multivariati riferiti a casi di studio sia reali che simulati e a diversi ambiti applicativi: numerosissimi sono gli output su dati reali e simulati presentati negli ambienti R con l'ausilio di RMarkdown. Nelle esercitazioni pratiche si apprendono le procedure necessarie e i codici necessari per svolgere in autonomia gli esercizi. Infatti, con l'ausilio di R nell'ambiente RStudio e dell'interfaccia RMarkdown lo studente impara il relativo linguaggio di programmazione e crea documenti riproducibili. Durante le esercitazioni lo studente viene incoraggiato a riconoscere la problematica dell'esercizio, e a individuare la metodologia più adatta, oltre che ad applicare le analisi e commentare i risultati.
Saranno a disposizione sulla pagina e-learning degli studenti prima di ogni lezione le slides e la parti della dispensa inerenti gli argomenti presentati.

Modalità di verifica dell'apprendimento

L'esame è in forma scritta. Non sono previste prove intermedie. Le seguenti modalità di verifica dell'apprendimento sono valide sia per gli studenti frequentanti le lezioni in presenza che non frequentanti.
L'esame si svolge in laboratorio. Lo studente dovrà rispondere a due quesiti teorici tra un insieme di domande predeterminate che conoscerà già all'inizio del corso. Occorre argomentare la risposta in termini comprensibili ed esaurienti riportando le dimostrazioni richieste. Il punto di riferimento per le risposte sono le slides e la dispensa: ovviamente si possono riportare le conoscenze acquisite dai libri consigliati. Si devono riportare formule e grafici: se risulta difficoltoso si possono scrivere su foglio con penna e poi scannerizzarlo. La lunghezza richiesta delle risposte dipenderà dalla domanda: si suggeriscono risposte che non superino i quattro fogli dattiloscritti in calibri 12 interlinea 1.5 (12000 battute spazi inclusi).
La seconda parte dell'esame consterà in un esercizio pratico su dati reali o simulati forniti dal docente mediante l'uso di pacchetti statistici. Gli strumenti statistici che dovrà utilizzare saranno quelli appresi al corso. Nell'elaborato tutti i grafici e gli output dovranno essere opportunamente commentati, sia da un punto di vista teorico, sia rispetto all'applicazione in esame. Lo svolgimento avviene tramite l'ambiente R. Lo studente potrà utilizzare i codici delle esercitazioni durante l'esame. Tali codici verranno forniti il giorno della prova.

La seconda parte dell'esame consterà in un esercizio pratico su dati reali o simulati forniti dal docente mediante l'uso di pacchetti statistici. Gli strumenti statistici che dovrà utilizzare saranno quelli appresi al corso. L' elaborato dovrà comprendere commenti dettagliati rispetto ai codici impiegato e ai risultati ottenuti. Lo svolgimento avviene tramite l'ambiente R. Lo studente potrà utilizzare i codici delle esercitazioni durante l'esame. Tali codici verranno forniti il giorno della prova.

Testi di riferimento

I principali testi di riferimento sono
• Spinelli, D. , Vittadini G.(2023) course slides
• Pennoni, F. Spinelli D, Vittadini G. (2023). Dispensa di Analisi Statistica Multivariata –Modulo Modelli Statistici-parte di teoria e applicazioni con R e SAS. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.
• - Baltagi B. H. (2008), Econometrics, fourth Edition, Springer Berlin
• Faraway, J. J. (2014). Linear models in R, Second Edition, Chapman & Hall, CRC Press.
• - Freund, R. J., Wilson, W. J., and Sa, P. (2006), Regression Analysis: Statistical Modeling of a Response Variable, 2nd edition, Academic Press
• Johnson, R. A., and Wichern, D. W. (2002). Applied multivariate statistical analysis, Pearson Education International, Prentice-Hall.
• Hastie, T., D. & Tibshirani, R. (2013). An introduction to statistical learning, New York, Springer.
• Littell, R. C., Freund, R. J., and Spector, P. C. (2002), SAS for Linear Models, 4th Edition, Cary, NC: SAS Institute Inc.
• Manual SAS/STAT 15.1
• Nolan, D., & Lang, D. T. (2015). Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving. Chapman & Hall, CRC Press.
• R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
• Wooldridge, J. M. (2015). Introductory econometrics: A modern approach. Cengage learning.

Periodo di erogazione dell'insegnamento

II Semestre, III Ciclo

Lingua di insegnamento

Italiana

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ