UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models

2021-2-E4102B084-E4102B085M

Obiettivi formativi

Il modulo di modelli statistici intende sviluppare le conoscenze teoriche e applicative circa il modello 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 l'utilizzo dei modelli attraverso l'impiego di 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 dei software R e SAS sia per le analisi di statistica descrittiva multivariata che per l'applicazione del modello di regressione lineare multipla e del modello 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

Lo studente viene richiamato allo schema concettuale dell'inferenza statistica e alle analisi grafiche multivariate ancorché all'utilizzo dei coefficienti di correlazione totali e parziali come misure di associazione.

Viene introdotta la funzione di regressione lineare multipla nel caso di tre variabili e si esplicitano le assunzioni sottostanti. Viene spiegato il metodo di stima dei minimi quadrati e le proprietà principali degli stimatori dei parametri del modello. Si illustra la distribuzione di Gauss bivariata e multivariata e le relative proprietà vengono enunciate sia a livello teorico che con esempi applicativi su dati reali e simulati.

Si considera il modello di regressione lineare multipla a fini esplicativi e presivisi. Si illustra come valutare il modello considerando i seguenti aspetti: gli indici di adattamento, la scelta del numero di variabili esplicative, le analisi grafiche dei residui, ed i criteri d'informazione. Si discute e si valuta la presenza di multicollinearità e si accenna ai metodi di linearizzazione. Ci considera il concetto di odds e odds ratio e si introduce il modello di regressione logistica generale, i metodi di stima dei parametri e l'interpretazione dei coefficienti stimati.

Nelle prime tre settimane di corso gli esempi su dati reali e simulati vengono svolti nell'ambiente R con l'ausilio di RMarkdown per integrare codice e output. In questo modo lo studente apprende anche ad effettuare analisi riproducibili. Nelle ultime settimane viene spiegato l'utilizzo delle procedure SAS sia in riferimento alle analisi preliminari dei dati sia per l'adattamento del modello di regressione lineare multipla e di regressione logistica.

Programma esteso

Il corso viene introdotto accennando all'impianto concettuale dell'inferenza statistica e alle differenze tra causazione e associazione. Si richiamano le tipologie dei caratteri, la rappresentazione matriciale dei dati e l'indice di correlazione tra caratteri quantitativi.

Il modello di regressione lineare multipla viene prima introdotto come funzione di regressione che coinvolge tre variabili sia nella notazione estesa che nella notazione matriciale. Si richiama la scomposizione della devianza totale, nel caso di tre variabili esplicative si richiama il metodo dei minimi quadrati. Vengono illustrate le proprietà degli stimatori in base alle assunzioni del modello e l'inferenza sui coefficienti di regressione viene presentata sia per il singolo parametro che per coppie di parametri attraverso la determinazione degli intervalli di confidenza congiunti.

Si introduce la distribuzione di Gauss bivariata e multivariata. Si illustra il metodo per ottenere delle realizzazioni simulate da entrambe le distribuzioni attraverso i vettori delle medie e la matrice di varianza-covarianza. Si utilizzano i grafici a dispersione a due e a tre dimensioni e le curve di livello per la distribuzione bivariata insieme all'ellissoide di concentrazione.

Si descrivono le principali analisi diagnostiche riguardanti i residui. Si introduce il criterio d'informazione Bayesiano e le tecniche di stepwise selection per la selezione delle variabili esplicative. Si accenna al problema della multicollinearità e viene introdotto l'indice d'inflazione della varianza. Viene illustrato l'utilizzo del modello ai fini predittivi ed i concetti di training set e testing set. Vengono illustrate le previsioni sia la risposta riferita ad una singola unità sia per il valore medio della risposta.

Nel corso si introducono anche gli aspetti seguenti: *i)* il metodo di stima della massima verosimiglianza; ii) trasformazione delle variabili; *iii*) variabili esplicative categoriali; *iv*) modelli con ordini di interazione tra variabili esplicative; *v*) odds e odds ratio; *vi*) variabile risposta categoriale con riferimento al modello generale di regressione logistica multipla.

Gli argomenti trattati a livello teorico sono affiancati dall'illustrazione di numerose applicazioni su dati reali e simulati che vengono sviluppate tramite l'ambiente statistico R, Rstudio utilizzando il marcatore di testo RMarkdown per sviluppare analisi riproducibili. Nelle ultime due settimane di corso lo studente impara anche la

semantica del software SAS per le analisi descrittive e per la stima del modello di regressione lineare multipla e logistica attraverso le procedure proc sgscatter, proc reg, proc glm, proc glmselect.

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 consigliato aver già sostenuto l'esame di inferenza statistica.

Metodi didattici

Sono previste lezioni frontali riguardanti la parte di teoria, queste vengono affiancate da esercitazioni pratiche. Tutte le lezioni si svolgono in laboratorio informatico: la parte di teoria viene affiancata allo sviluppo di applicazioni a problemi concreti relativi a dati multivariati sia reali che simulati e riferiti a svariati ambiti applicativi. Sono inoltre previste delle lezioni di tutoraggio affinché lo studente possa essere coadiuvato nello svolgimento degli esercizi assegnati settimanalmente.

Durante le esercitazioni con l'ausilio di R nell'ambiente RStudio e dell'interfaccia RMarkdown lo studente impara il relativo linguaggio di programmazione e crea documenti riproducibili. Lo studente impara inoltre l'utilizzo del software SAS per le analisi dei dati e la stima dei parametri dei modelli statistici. Viene incentivato l'apprendimento cooperativo. Durante delle esercitazioni lo studente viene incoraggiato a: riconoscere la problematica dell'esercizio, individuare la metodologia più adatta, applicare le analisi e commentare i risultati.

Durante il periodo di emergenza covid le lezioni si svolgeranno in modalità da remoto (lezioni videoregistate) con incontri periodici in videoconferenza tramite piattaforma webex e con in presenza secondo le calendarizzazioni previste dall'ateneo e che verranno rese note nella pagina del corso.

Modalità di verifica dell'apprendimento

L'esame è in forma scritta con orale facoltativo, non sono previste prove intermedie. Le seguenti modalità di verifica dell'apprendimento riguardano sia gli studenti che frequentano le lezioni sia coloro che non possono essere presenti alle lezioni. L'esame scritto ha durata complessiva di un'ora e trenta minuti e si svolge presso il laboratorio informatico. Lo studente deve rispondere ai punti dell'esercizio utilizzando il computer. Questi riguardano sia la parte di teoria che l'applicazione delle analisi descrittive e dei modelli di regressione lineare multipla o logistica utilizzando dati reali o simulati forniti dal docente. Lo studente predispone un elaborato con commenti dettagliati rispetto al codice impiegato ed i risultati ottenuti oltre che la teoria illustrata nei principali aspetti. Lo svolgimento avviene tramite l'ambiente R oppure tramite il software SAS. Lo studente può disporre del materiale fornito durante il corso e del codice illustrato durante le lezioni e le esercitazioni. La prova permette la verifica delle nozioni teoriche e della capacità di comprensione del problema applicativo nonché di risoluzione dello stesso tramite l'analisi dei dati. Permette di valutare inoltre la capacità comunicativa tramite la creazione di un report.

La prova orale è facoltativa e riguarda sia la teoria che le applicazioni. Può essere richiesta da coloro che hanno un esito di almeno 18/30 alla prova scritta al momento della pubblicazione degli esiti. Questi sono pubblicati sulla pagina di e-learning dedicata al corso.

Durante il periodo di emergenza Covid-19 la modalità di esame dipende dalle disposizioni ateneo.

Testi di riferimento

Il materiale didattico è costituito principalmente dalle dispense redatte dal docente riguardanti sia la parte teorica che le applicazioni. Tutto il materiale è disponibile nella pagina della piattaforma e-learning dell'ateneo dedicata al corso. Il docente pubblica al termine di ogni lezione anche le slides, i programmi di calcolo, gli esercizi, i dati, e le soluzioni di ogni lezione. Nella stessa pagina sono pubblicati alcuni testi d'esame.

Durante il periodo di emergenza Covid-19 nella pagina del corso vengono anche pubblicate le videoregistrazioni delle lezioni.

I principali testi di riferimento sono elencati nella bibliografia delle dispense, tra gli altri si segnalano i seguenti:

Faraway, J. J. (2014). *Linear models in R,* Second Edition, Chapman & Hall, CRC 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.

Nolan, D., & Lang, D. T. (2015). *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*. Chapman & Hall, CRC Press.

Pennoni, F. (2020). *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.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

SAS/STAT 9.4. PROC SGSCATTER, PROC CORR, PROC REG, PROC GLM, PROC GLMSELECT,*User's guide*, SAS Institute, 2012.

Periodo di erogazione dell'insegnamento

II Semestre, III Ciclo: febbraio - aprile 2020.

Lingua di insegnamento

Il corso viene erogato in lingua italiana. Gli studenti Erasmus possono utilizzare il materiale didattico in Inglese e possono richiedere al docente che la prova d'esame sia svolta in lingua inglese.