Compito del 1 febbraio 2017

Describe what are the aims of Exploratory Data Analysis and present the main numerical summaries for bivariate data.

Il principale scopo dell'analisi esplorativa è quello di visualizzare graficamente e numericamente il dataset a disposizione alla ricerca di pattern, idee e ipotesi da confermare attraverso le successive analisi formali o che inizialmente non erano state prese in considerazione.

Principali indici per variabili bivariate:

- **covarianza (campionaria)** s_{xy}: misura la tendenza delle due variabili X e Y ad esser in correlazione lineare. Se $s_{xy} = 0$, allora non c'è alcuna correlazione lineare tra le due variabili.
- **coefficiente di correlazione di Pearson**: misura di correlazione lineare standardizzata tra variabili normali (o, almeno, simmetriche)
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$
$r_{xy} \in [-1, 1]$. Se $|r_{xy}| \approx 1$ allora c'è una forte correlazione lineare tra le variabili.
- **coefficiente di correlazione di rango di Spearman** e **coefficente di correlazione τ di Kendall**: misurano la correlazione tra due variabili in caso di distribuzioni asimmetriche o correlazioni di tipo non-lineare.
- **tabelle di contingenza** per le variabili categoriali.

Define the multiple linear regression model and highlight the basic assumptions.

In un modello di regressione lineare multiplo il valore della variabile risposta $Y = (Y_1, \dots, Y_n)$ (n è il numero di osservazioni) dipende da due o più regressori $X_1, \dots, X_p$ definito come

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i_2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Assunzioni del modello:

- I termini di errore sono **normalmente distribuiti** ($\epsilon_i \sim N(0, \sigma^2)$) e gli errori delle singole unità sono indipendenti tra di loro (ipotesi di normalità e omoschedasticità dei residui)
- Le singole variabili di risposta Y_i sono:
  - i.i.d.
  - normalmente distribuite
  - con varianza costante σ^2
- valore medio definito come combinazione lineare dei regressori (ipotesi di linearità)
$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 xi_2 + \cdots + \beta_p x_{ip}$$

Describe the least squares estimators for the regression parameters and define a suitable estimator for the variance parameter.

Il metodo dei minimi quadrati serve a stimare il valore dei parametri di regressione per un modello di regressione lineare semplice. Minimizza la somma dei quadrati dei residui, ossia le differenze tra i dati osservati e i valori stimati dal modello.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{s_{xy}^2}{s_x^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$\hat{\alpha}$ e $\hat{\beta}$ sono stimatori **consistenti** e **non distorti**.

Uno stimatore per la varianza di Y è il **residual standard error** elevato al quadrato, che consiste nella somma dei quadrati dei residui diviso il numero di osservazioni - 2 (i gradi di libertà).

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2}$$

Discuss the usefulness of the fitted regression model for inferential and prediction purposes.

Nella statistica inferenziale i modelli di regressione servono a stimare la relazione $Y = f(X)$ sulla base dei dati a disposizione.

Nella statistica predittiva i modelli di regressione servono, dato un insieme di input, ad ottenere una stima (o un intervallo di stima) dei valori associati utilizzando un predittore, che può essere considerato una "black box"

Define the confidence intervals for both the regression parameters and the regression line and specify the prediction interval for a future response variable.

Dato un livello γ di confidenza:

$$\alpha = \left[\hat{\alpha} \pm t_{n-2,\frac{1-\gamma}{2}} SE(\hat{\alpha})\right]$$

t_{n-2} è la distribuzione t di Student con $n-2$ gradi di libertà, $SE(\cdot)$ lo standard error.

La formula è analoga per il parametro β e per la linea di regressione $\mu_0 = \hat{\alpha} + \hat{\beta}x_0$.

L'intervallo di previsione per una variabile di risposta $Y_0 = \mu_0 + \varepsilon_0$ dato un nuovo valore x_0 vale

$$\left[\hat{Y_0} \pm t_{n-2,\frac{1-\gamma}{2}} SE(\hat{Y_0})\right]$$

Con $SE(\hat{Y_0}) = \sqrt{\hat{\sigma^2} + SE(\hat{\mu_0})^2}$

Compito del 13 febbraio 2017

Describe the purpose of an interval estimation procedure.

Gli intervalli di confidenza offrono delle stime molto più accurate sul reale valore del parametro ricercato rispetto a una semplice stima puntuale.

Give the right statistical interpretation of an observed 95% confidence interval for an interest parameter.

Un osservazione di un intervallo casuale, basato su un campione statistico, con probabilità del 95% che nell'intervallo di confidenza sia contenuto il reale valore del parametro.

Present a simple application regarding the estimation of a population mean.

Da un data set di n osservazioni su una variabile y di interesse. Da queste osservazioni calcolo:

- media campionaria $\bar{y}$
- mediana $y_{0.5}$
- varianza campionaria (corretta) S^2
- deviazione standard $S = \sqrt{S^2}$
- standard error della media $SEM = \frac{S}{\sqrt{n}}$

Dato un livello di confidenza α l'intervallo di confidenza per $\bar{y}$ sarà

$$\left[\bar{y} \pm t(n-1)_{\frac{1-\alpha}{2}} SEM\right]$$

List some useful steps in the model fitting procedure.

1. Esamina la distribuzione delle variabili esplicative e della variabile risposta. Cerca distribuzioni asimmetriche e gli outlyers.
2. Esamina gli scatterplot di tutte le variabili esplicative e della variabile risposta.
3. Notare gli intervalli di ciascuna delle variabili del grafico a dispersione, considerando se variano sufficientemente per influenzare la variabile di risposta e se ciascuna delle variabili esplicative è misurata accuratamente.
4. Nel caso in cui uno scatterplot suggerisca dei pattern non lineari, considera l'uso di trasformate
5. In caso di distribuzioni asimmetriche si consiglia di traformare la variabile risposta
6. Le coppie di variabili esplicative con un'alta correlazione tali da sembrare fornire le stesse informazioni dovrebbero essere analizzati. Le informazioni di base possono suggerire quale delle due conservare.

Recall the main statistical indices and procedures for model assessment and model selection.

- Coefficiente di determinazione R^2 e la sua versione corretta: misura quanto la variabilità della variabile risposta è coperta dal modello.
- Procedure basate sui test F per la selezione di modelli annidati
  - test di ANOVA
  - Si parte dal modello con tutti i regressori, ad ogni iterazione si scarta (o si trasforma) un regressore, eliminando dal modello il regressore con il p-value più alto, finchè tutti i regressori hanno il p-value sotto una certa soglia (backward selection)
  - Si parte da un modello semplice a un regressore e ad ogni iterazione si aggiunge il regressore con il p-value del test F più basso, finchè il valore del test F converge (forward selection)
- Statistiche AIC e BIC:
$$AIC = n \log\left(\frac{\sum_{i=1}^{n} \hat{\varepsilon_i^2}}{n}\right) + 2p + const$$
BIC sostituisce $2p$ con $\log(n) \cdot p$, penalizzando i modelli con molti parametri
- Statistica di Mallow C_p:
$$C_p = n \log\left(\frac{\sum_{i=1}^{n} \hat{\varepsilon_i^2}}{n}\right) + 2p - n$$
- Anche per queste statistiche si possono implementare dei metodi di forward e stepback selection (simili a quelli visti in precedenza)

Compito del 15 febbraio 2018

Describe the purpose of a point estimation procedure.

Dato un modello statistico parametrico con un parametro sconosciuto θ e una serie di daty y una procedura di stima puntuale ha lo scopo di trovare i valori per θ più consistenti con i dati y.

List the main property of an estimator and define the standard error.

Uno stimatore $\hat{\theta}$ (inteso come variabile casuale) per un parametro θ dovrebbe essere:

- Consistente: $E(\hat{\theta}) = \theta$
- Con varianza bassa.

- Non distorto: $\hat{\theta} \xrightarrow{p} \theta$
- Asintoticamente normale per gli stimatori di massima verosimiglianza

Lo standard error (SE) è la radice quadrata del **mean square error** (MSE), ed è una misura dell'accuratezza della stima.

Per stimatori non distorti, lo standard error è la radice quadrata della varianza.

Present a simple application regarding the estimation of a proportion.

Stimare la probabilità di successo p per una variabile casuale $Y = Y_1, \ldots, Y_n$ con $Y_i \sim Ber(p)$.

Uno stimatore $\hat{p}$ non distorto e consistente per p è la media campionaria

$$\bar{Y} = \frac{\text{n° successi}}{n}$$

Lo SE associato è

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Define the one-way and the two-way analysis of variance models and highlight the basic assumptions.

I modelli ANOVA servono a confrontare le medie di più gruppi, in particolare come la media della variabile risposta dipenda dal livello dei regressori di tipo categoriale (uno nel caso del one-way ANOVA, due nel caso del two-way ANOVA, ma è generalizzabile).

Il modello statistico prevede le seguenti assunzioni:

- i termini di errore sono $\varepsilon_{ij} \sim N(0, \sigma^2)$ i.i.d.
- l'j-esimo valore del gruppo i-esimo $Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$ con μ la media globale, τ_i la deviazione dalla media globale dell'i-esimo gruppo

Describe the statistical tests on the main effects and on the interaction effect of the factors on the mean response.

L'ipotesi nulla per questi test è che tutti gli effetti introdotti per ogni gruppo sia uguale a 0, ossia che tutti i dati provengano dalla stessa origine a prescindere dal gruppo di appartenenza.

Confrontando delle opportune somme di quadrati e gradi di libertà, si definisce un opportuno test F.

Nel caso in cui l'interazione tra due gruppi sia significativa sul valore della media, non ha molto senso investigare anche gli effetti delle due variabili separatamente, perchè c'è un alta probabilità che entrambi la influenzino.

Compito del 11 giugno 2018

Present a simple application regarding the estimation of a population variance.

Uno stimatore consistente e non distorto per la varianza è la **varianza campionaria (corretta)**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

In caso in cui i dati di partenza siano i.i.d. e $Y_i \sim N(\mu, \sigma^2)$, allora

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Con due campioni X e Y i.i.d. di dimensioni n_X e n_Y, per comparare le due varianze si utilizza il rapporto tra le varianze due campionarie corrette

$$\frac{S_X^2}{S_Y^2}$$

Se $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$ allora

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n_X - 1, n_Y - 1)$$

Introduce and discuss the topic of regression models with non-Gaussian response variables.

Nel caso in cui le variabili risposta non siano di tipo gaussiano (binomiali, Bernoulli, Poisson), si utilizzano i modelli di regressione lineare generalizzati (es. i modelli logistici).

Dal punto di vista del predittore lineare sono identici ai modelli gaussiani, con l'aggiunta di una funzione chiamata **link** che serve a convertire i valori dalla scala di Y alla scala del predittore.

Nei modelli di regressione generalizzati non esiste la varianza, ma la si generalizza con la *devianza* (che ha un ruolo simile alla somma dei quadrati) e non si utilizza il metodo dei minimi quadrati per stimare i parametri di regressione (si utilizza la stima di massima verosimiglianza o i metodi Bayesiani)

Consider the case of a Bernoulli distributed response and define the logistic regression model.

Per le variabili da una distribuzione Bernoulliana (quindi con risposta binaria), il modello di regressione più utilizzato è quello delle *log odds* (quote), utilizzato nell'ambito delle scommesse. La funzione link detta **logit (logistic) link** è

$$f(u) = \log\left(\frac{u}{1-u}\right)$$

Data una probabilità p che un evento si realizzi e una certa quota q:

$$q = \frac{p}{1-p}, \quad \log(q) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$$

Di conseguenza dato il valore di una quota q, il valore di p è uguale a

$$p = \frac{e^{\log(q)}}{1 + e^{\log(q)}}$$

In questo modo è possibile costruire dei modelli di regressione per ottenere il valore di $\log(q)$

With regard to a fitted logistic regression model, emphasize the interpretation of the estimated regression parameter and discuss its potential application for predicting a future binary response.

Le stime dei parametri di regressione nei modelli logistici misurano il "peso" che questi regressori hanno nel definire la probabilità che un evento accada o meno (al netto dei valori dei p-value sugli stessi regressori).

Se un parametro ha valore positivo, allora quel regressore ha un'influenza positiva sulla probabilità (più aumenta il valore del regressore, più aumenta la probabilità che l'evento accada), altrimenti quel regressore ha un'influenza negativa (più aumenta il valore del regressore, più la probabilità che l'evento accada cala).

In particolare l'esponenziale dei coefficienti fornisce l'incremento (o il decremento) della quota di risposta all'aumentare del valore del regressore.

Ovviamente non è sufficiente analizzare i singoli regressori, ma anche la correlazione tra essi, altrimenti si corre il rischio di trarre conclusioni errate o superficiali (*confounding phenomenon*).

Compito del 4 febbraio 2019

Define the Gaussian distribution and describe its usefulness in statistical applications.

La distribuzione gaussiana (o normale) $X \sim N(\mu, \sigma^2)$ ha le seguenti caratteristiche:

- distribuzione **continua** con supporto $\mathbb{R}$
- parametri $\mu \in \mathbb{R}$ e $\sigma^2 > 0$
- valore atteso $E(X) = x_{0.5} = \mu$ e varianza $Var(X) = \sigma^2$
- funzione di densità di probabilità

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- chiusa rispetto a trasformazioni lineari $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$

Si tratta della distribuzione più importante nell'ambito della Statistica e Probabilità, in quanto è:

- il risultato del teorema centrale del limite
- viene utilizzata per modellare un gran numero di fenomeni continui (anche attraverso l'uso di traformazioni).

Discuss the case in which the explanatory variables are factor, with particular regard to the codification using dummy variables.

Nel caso in cui i regressori siano fattori si utilizzano i modelli ANOVA.

La variabile risposta (o la sua trasformata) deve avere varianza costante.

Per includere un fattore con h livelli in un modello, si introducono nel modello $h-1$ variabili binarie (dummy), ognuna con il proprio coefficiente di regressione, il quale serve a esprimere il main effect del fattore sulla risposta (lo scostamento tra la media globale e la media del gruppo).

Uno dei livelli è impostato come referenza, per cui gli effetti degli altri livelli sono misurati a partire da quest'ultimo.

Consider the situation with both factors and numerical explanatory variables, focusing on the particular case of models admitting different simple regression lines.

Nel caso in cui si debba fare una regressione che coinvolga sia variabili categoriali che numeriche, si utilizza l'analisi della varianza e i test F (o t) per scegliere sostanzialmente tra due tipi di modelli:

- modelli di regressione in cui la differenza tra un livello e l'altro è solo nell'intercetta (quindi le linee di regressione sono parallele)
- modelli che includono anche l'interazione tra le variabili categoriali e quelle numeriche (moltiplicando la variabile dummy con la variabile numerica), ottenendo quindi linee di regressione diverse per ogni categoria

Compito del 21 febbraio 2019

Present a simple application regarding the estimation of the difference of the means of two independent populations.

1. Calcolo le medie campionarie delle due popolazioni $\bar{X}$ e $\bar{Y}$ e stimo il valore della differenza come $\bar{X} - \bar{Y}$

2. Calcolo lo standard error della differenza SED utilizzando la **varianza campionaria aggregata corretta (pooled sample variance)** S_p^2

$$S_p^2 = \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2}{n_X + n_Y - 2}$$

$$SED = \sqrt{SEM_X^2 + SEM_Y^2} = S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

Compito del 28 gennaio 2020

Describe the purpose of a (parametric) hypothesis testing procedure.

I test statistici parametrici si basano sulla specifica di un modello statistico parametrico (con parametro θ) e hanno lo scopo di verificare se una particolare ipotesi sul parametro θ è consistente con i dati a disposizione.

I test non-parametrici invece utilizzano delle assunzioni più deboli, sopratutto sulla distribuzione di probabilità della variabile casuale in analisi.

Define the notions of significance level, critical region and p-value.

- il livello di significatività α di un test, solitamente all'1% o al 5%, è la probabilità che il test conduca a un errore di tipo I (rifiutare l'ipotesi nulla quando è vera).
- la regione critica è l'insieme dei valori della distribuzione della statistica di test per cui l'ipotesi nulla viene rifiutata a favore di quella alternativa.
- il p-value è la probabilità (sotto l'ipotesi nulla) di ottenere risultati ugualmente o meno compatibili, di quelli osservati durante il test, con la suddetta ipotesi. In pratica aiuta a capire se la differenza tra il valore osservato e quello ipotizzato è dovuto alla casualità introdotta dal campionamento oppure se tale differenza è statisticamente significativa.

Solitamente se il valore del p-value è inferiore al livello di significatività, allora l'ipotesi nulla viene rigettata.

Present a simple application concerning the testing on the equality of the means of two independent populations.

Le ipotesi sono:

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Se le osservazioni provengono da una v.c. con distribuzione normale con varianza sconosciuta, si utilizza il t test, che vale, sotto H_0

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

Se la varianza σ^2 è conosciuta si usa lo z test, che vale, sotto H_0

$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Dato un livello di significatività α la regione critica per il t test è

$$R_\alpha = \{y : |t| \geq t_{n-1;\alpha/2}\}$$

Il p-value

$$p = 2\min\{P_{H_0}(T \leq t), P_{H_0}(T \geq t)\}$$

Discuss the crucial point of selecting the explanatory variables in multiple linear regression models.

1. Partire da informazioni conosciute sul fenomeno in questione (ad esempio, ci viene indicato da un esperto che una certa variabile va necessariamente inserita in un modello)
2. Avere a disposizione un buon numero di dati (almeno 10 volte il numero di covariate)
3. Avere una visualizzazione grafica delle variabili in gioco (es. matrice scatterplot)
4. Evitare covariate che forniscono informazioni ridondanti o sovrapposte
5. Attenzione alle correlazioni spurie! (correlazioni casuali o dovute a variabili nascoste)
6. Utilizzare tecniche di selezione semi-automatica delle covariate come lasso, booting, least angle regression, oppure l'Analisi delle Componenti principali (ridurre il numero di covariate sostituendole con delle combinazioni lineari delle stesse)

Discuss the problem of multicollinearity and consider the potential remedies.

Ci possono essere situazioni in cui le covariate portano contenuti informativi sovrapposti (c'è un legame tra le covariate).

Per misurare queste correlazioni si utilizza il **fattore di inflazione della varianza** che vale per la variabile x_j

$$VIF_j = \frac{1}{1 - R_j^2}$$

Dove R_j^2 è il coefficiente R^2 del modello di regressione ottenuto utilizzando x_j come variabile risposta e le altre variabili esplicative come regressori.

Compito del 18 febbraio 2020

Describe what are the aims of Exploratory Data Analysis and present the main graphical summaries for describing the relationship between different types (namely, categorical and numerical) of variables.

Il principale scopo dell'analisi esplorativa è quello di visualizzare graficamente e numericamente il dataset a disposizione alla ricerca di pattern, idee e ipotesi da confermare attraverso le successive analisi formali o che inizialmente non erano state prese in considerazione.

I principali grafici per confrontare variabili numeriche e categoriali sono i boxplot e gli strip plot: su un asse viene messa la variabile categoriale, sull'altro asse quella numerica. Per ogni livello della variabile categoriale vengono riportati i valori (sottoforma di punti nel caso degli strip plot, oppure sottoforma di box nei boxplot) e servono a confrontare l'andamento dei dati in base al gruppo di appartenenza.

Define the simple linear regression model and recall the t test on the nullity of the slope parameter, discussing its role in evaluating the model adequacy.

Un modello di regressione lineare mette in correlazione la variabile risposta con una variabile esplicativa (detta regressore) secondo la seguente formula:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Il test t sulla nullità del parametro β utilizza come ipotesi nulla $H_0 : \beta = 0$ e ipotesi alternativa $H_1 : \beta \neq 0$. Nel caso di regressioni lineari semplici il p-value di questo test serve a verificare la significatività del regressore preso in esame sul valore della variabile risposta.

Define the one-way analysis of variance model and describe the statistical test on the effect of the factor on the mean response

One-way ANOVA è un insieme di tecniche per confrontare le medie di gruppi di dati in base al livello di un fattore (di solito varabili categoriali).

Il modello statistico alla base di ANOVA è un caso particolare di regressione lineare:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Dove y_{ij} è l'osservazione j-esima appartenente all'i-esimo gruppo, μ è la media globale, ε_{ij} l'errore random, τ_i è la differenza tra la media globale e la media dell'i-esimo gruppo (effetto trattamento). Le osservazioni dell'i-esimo gruppo $Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$ sono indipendenti.

Sul valore di τ_i è definito un test statistico così definito:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$$

$$H_1 : \tau_i \neq 0 \text{ per almeno un gruppo}$$

E si svolge comparando due stime della varianza:

$$\hat{\sigma_0^2} = \frac{\sum_{i=1}^{a} n(\bar{y_i} - \bar{y})^2}{a-1} \quad \hat{\sigma^2} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y_i})^2}{an-a}$$

Con il test F che vale, sotto H_0

$$F = \frac{\hat{\sigma_0^2}}{\hat{\sigma^2}} \sim F(a-1, an-a)$$

Compare the regression model and the ANOVA model when the levels of the factor are quantitative.

Nei modelli ANOVA il test statistico sull'effetto trattamento ignora il fatto che i livelli siano quantitativi.

In generale, in questi casi, i "classici" modelli di regressione sono preferibili rispetto ai modelli ANOVA, perchè permettono (ad esempio) di poter interpolare su livelli successivi e forniscono p-value più piccoli (in media)

Varie ed eventuali

(Cap 3a) Linea dei minimi quadrati pesata

Per stimare i parametri di regressione con il metodo dei minimi quadrati, può essere necessario introdurre dei "pesi" w_i per ogni osservazione. La funzione da minimizzare diventa

$$\min \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 w_i$$

(Cap 3a) Paradosso di Simpson

Si riferisce al fatto che le relazione tra due variabili può cambiare quando i dati sono partizionati in sottogruppi e/o quando vengono introdotte nuove variabili che prima erano nascoste. In pratica avere dati troppo aggregati può portare a conclusioni errate!

(Cap 3b) Test di indipendenza chi-quadro

Test statistico per stabilire se due variabili categoriali sono indipendenti tra loro. In pratica si testa se per tutte le combinazioni possibili la probabilità sia pari al prodotto delle probabilità marginali (ipotesi nulla), attraverso la seguente statistica:

$$X^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2((r-1)\cdot(c-1))$$

$$e_{ij} = \frac{n_i n_j}{n}$$

sotto ipotesi nulla e per grandi tabelle di contingenza ($e_{ij} \leq 5$ per ogni cella).

(Cap 4) Il metodo di Box-Cox

Per scegliere la giusta trasformata per la variabile risposta (per aumentarne simmetria e normalità).

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log(y) & \text{se } \lambda = 0 \end{cases}$$

Per un parametro reale λ ottenibile attraverso diversi metodi, il più utilizzato prevede di massimizzare il valore di una funzione di log-verosimiglianza.

(Cap 5) Effetto leva

Dato una osservazione y_i, se il suo valore cambia a $y_i + \Delta_i$, il corrispondente fitted value nel modello $\hat{y}_i$ diventa $\hat{y}_i + h_{ii}\Delta_i$. h_{ii} è detto **leva** dell'osservazione. In un modello con p coefficienti e n osservazioni, sono considerate alte le leve maggiori di $2p/n$ o $3p/n$

(Cap 5) Punti influenti

In un modello un'osservazione è considerata **influente** se omettendola dal modello altera i fitted values del modello. L'influenza è una combinazione dei residui con la leva delle osservazioni.

Per misurare l'influenza di un'osservazione si utilizza la **Distanza di Cook**, che misura il cambiamento nelle stime del modello quando l'osservazione in esame viene omessa, utilizzando i residui standardizzati. Si considerano a influenza elevata i punti con distanza di Cook > 0.5

(Cap 6) Misurare la qualità di un modello

Per verificare poi l'accuratezza del modello si confrontano i valori predetti con quelli osservati, attraverso il calcolo del *training MSE*.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Tuttavia non è consigliato riutilizzare gli stessi dati per verificare l'accuratezza (tende a sovrastimare questo valore), per cui si ricorre a processi di **cross-validation** per ottenere riscontri più accurati.

(Cap 6) Regressione vs classificazione

Regressione: se la risposta è quantitativa Classificazione: se la risposta è qualitativa

(Cap 6) Classificatori

- Classificatore di Bayes: classificatore migliore. Assegna 1 se

$$P(Y_0 = 1|X_0 = x_0) > P(Y_0 = 0|X_0 = x_0)$$

  Siccome non è conosciuta la distribuzione di probabilità di Y_0 dato $X_0 = x_0$ non è utilizzabile nella pratica, ma lo usa come confronto per gli altri metodi

- Classificatore logistico: basato su modelli di regressione lineare logistica

$$P(Y_i = 1|X_i = x_i) = \frac{\exp\{x_i^T \beta\}}{1 + \exp\{x_i^T \beta\}}$$

- Linear e Quadratic discriminant analysis (LDA/QDA): si basa sulla stima della probabilità $P(X = x|Y = y)$ più facile da stimare e poi ottenere $P(Y = x|X = x)$ con il teorema di Bayes
- k-Nearest Neighbors (kNN): stima le probabiità di un punto x_0 in base al valore dei $k > 0$ punti più vicini appartenenti ai dati di training.

$$\hat{P}(Y_0 = 1|X_0 = x_0) = \frac{1}{k} \sum_{i \in \mathcal{N},} y_i$$

Questo metodo è molto efficace con predittori numerici, ma è generalizzabile su predittori categoriali.

Il valore di k definisce la flessibilità della procedura di classificazione:

- per valori piccoli di k riduce il rischio di bias ma aumenta la varianza della risposta
- per valori grandi di k si riduce la varianza, ma aumenta il rischio di incorrere in bias.

kNN è utilizzabile anche per i problemi di regressione per approssimare risposte continue

(Cap 6) Matrice di confusione

Matrice per misurare le performance predittive di un classificatore. Sulle righe si mettono i valori predetti, sulle colonne quelli osservati.

(Cap 6) Curva ROC (Receiver Operating Characteristic)

Grafico che serve a selezionare la soglia di probabilità per classificatori binari allo scopo di aumentare l'accuratezza del modello.

Si ottiene plottando il *true positive rate* (numero di valori osservati veri predetti correttamente su tutte le osservazioni vere) sul *false negative rate* = 1-*true negative rate* (numero di valori osservati falsi predetti correttamente su tutte le osservazioni false).

Più l'area tra la curva ROC e la diagonale del grafico (classificatore completamente casuale) è ampia più il classificatore è preciso.

(Cap 7) Analisi delle componenti principali (PCA)

Da un insieme di variabili di dimensione p vogliamo definire un insieme di variabili derivate chiamate componenti principali, che riassumono tutta la variabilità del dataset di partenza, senza perdere informazioni.

Siano $X_1, \ldots, X_p$ un set di variabili. Si dice **prima componente principale** la combinazione lineare normalizzata (la somma dei quadrati dei coefficienti è pari a 1) delle variabili con la varianza maggiore.

Buona prassi è standardizzare (normalizzare) le variabili prima di procedere con la PCA.

Trovato il primo vettore dei coefficienti ϕ_{i1} il risultati delle combinazioni sono detti **scores** della prima componente principale.

La procedura è poi interabile fino a ottenere $\min(n-1, p)$ componenti, ad ogni iterazione si richiede di massimizzare la varianza non correlata alle componenti precedenti.

(Cap 7) Proporzione della varianza spiegata

Aiuta a "capire" quante componenti principali calcolare.

La proporzione della varianza spiegata dalla s-esima componente è il rapporto tra la varianza campionaria degli scores e la somma delle varianze campionarie delle singole variabili di partenza che compongono la componente.

Sommando tra loro le varie proporzioni si ottiene la cumulata della varianza principale spiegata.

(Cap 7) Algoritmi di clustering

Esistono tre tipi di algoritmi di clustering:

- Hierarchical clustering: costruisce una gerarchia di cluster sui dati. Ne esistono di due tipi: agglomerative (di tipo bottom-up) e divisive (top-down) Producono un output grafico chiamato **dendrogramma**. L'appartenenza di uno o più sottogruppi a una certa gerarchia avviene generalizzando il concetto di dissimilarità tra osservazioni a quello tra gruppi (criterio di collegamento). I cluster vengono individuati tagliando il dendrogramma a una certa "altezza", ossia la distanza a cui vengono fusi due sottogruppi.
- Partitioning clustering: partiziona i dati ottimizzando l'allocazione di questi ultimi nei cluster.
  - Algoritmo *K-means*: assegno un cluster a caso per ogni osservazione e finchè gli assegnamenti non cambiano calcolo il centroide (il vettore con le coordinate del punto medio) di ogni cluster e ad ogni osservazione assegno il cluster più vicino secondo la distanza euclidea.
  - Algoritmo *K-medoids*: variante del *K-means* solo che invece di calcolare i centroidi, vengono scelti dei punti per ogni dataset
- Model-based clustering: partiziona i dati in base a dei modelli.

(Cap 7) Misure di dissimilarità

Misurano la "somiglianza" (in termini numerici) tra le osservazioni. Hanno le seguenti proprietà: non negatività, identità degli indiscernibili (vale 0 se e solo se misuro la distanza tra un punto e sé stesso), simmetria, rispetto della disuguaglianza triangolare. Alcune misure di dissimilarità tra vettori:

- Distanza euclidea
- Distanza di manhattan: somma delle differenze in valore assoluto delle singole componenti di due vettori
- Distanza massima: massimo della differenza in valore assoluto delle componenti di due vettori
- Distanza binaria/Distanza di Jaccard: per vettori binari
- Distanza di Hamming: per stringhe
- Dissimilarità di Gower

(Cap 7) Alcune problematiche sul clustering

Il risultato di un operazione di clustering dipende sostanzialmente da tre fattori:

- la standardizzazione delle variabili,
- la scelta dei criteri di dissimilarità e/o di collegamento,
- l'altezza a cui tagliare il dendrogramma e/o il numero di partizioni.

In generale i metodi di clustering non sono molto robusti alle perturbazioni sui dati (outliers)