Introduzione al Machine Learning

Manuel Di Agostino
manuel.diagostino@studenti.unipr.it

Appunti del corso tenuto dal **Prof. Francesco Morandin**

Università degli Studi di Parma
Anno Accademico 2025/2026

Indice

Disclaimer

Nota

I seguenti appunti sono stati generati a partire dal materiale didattico e dalle note ufficiali del corso 'Introduzione al Machine Learning'. Il contenuto è stato riorganizzato, formattato e parzialmente rielaborato con l'ausilio di AI al fine di creare un documento coeso e ottimizzato per lo studio.

Si raccomanda di fare sempre riferimento al materiale originale fornito dal docente per garantire la massima accuratezza.

1 Ripasso Proprietà Variabili Aleatorie

1.1 Definizioni Generali

Ripassiamo i concetti fondamentali di Funzione di Ripartizione Cumulativa (CDF), Funzione di Densità di Probabilità (PDF) per variabili continue, e Funzione di Massa di Probabilità (PMF) per variabili discrete.

Definizione 1.1: Funzione di Ripartizione Cumulativa (CDF)

La CDF, indicata con $F_X(t)$, descrive la probabilità che una variabile aleatoria (v.a.) X assuma un valore minore o uguale a t.

$$F_X(t) = P(X \leq t)$$

La CDF è una funzione non decrescente con valori compresi tra 0 e 1. Per una v.a. continua, la CDF è una funzione continua, mentre per una v.a. discreta è una funzione a gradini.

Definizione 1.2: PDF (per v.a. continue) e PMF (per v.a. discrete)

- La **Funzione di Densità di Probabilità (PDF)**, $f(t)$, per una v.a. continua, è la derivata della CDF. L'area sottesa alla curva della PDF tra due punti a e b rappresenta la probabilità che la variabile cada in quell'intervallo.

$$f(t) = F'(t) \quad \text{e} \quad P(a < X \leq b) = \int_a^b f(t)\,dt = F(b) - F(a)$$

- La **Funzione di Massa di Probabilità (PMF)**, $\varphi_X(k)$, per una v.a. discreta, dà la probabilità esatta che la variabile assuma il valore k.

$$\varphi_X(k) = P(X = k) = F_X(k) - F_X(k-1)$$

La CDF può essere ottenuta sommando i valori della PMF.

$$F_X(k) = \sum_{j \leq k} \varphi_X(j)$$

Nota 1.1: Calcolo di probabilità tramite CDF

Ecco un riassunto delle formule per calcolare la probabilità di un evento usando la CDF:

- **Coda sinistra (v.a. continue e discrete):** Probabilità che X sia minore o uguale ad a.

$$P(X \leq a) = F_X(a)$$

- **Coda destra (v.a. continue):** Probabilità che X sia maggiore di b.

$$P(X > b) = 1 - F_X(b)$$

- **Coda destra (v.a. discrete):** Probabilità che X sia maggiore o uguale a b.

$$P(X \geq b) = 1 - F_X(b-1)$$

- **Intervallo (v.a. continue):** Probabilità che X sia compreso tra a e b.

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

- **Intervallo (v.a. discrete, estremi esclusi):** Probabilità che X sia strettamente compreso tra a e b.

$$P(a < X < b) = P(X \leq b-1) - P(X \leq a) = F_X(b-1) - F_X(a)$$

1.2 Inversione della Formula

Un uso molto comune della CDF è il calcolo del **quantile**, ovvero l'inversione della formula per trovare il valore x corrispondente a una data probabilità cumulata.

Esempio 1.1: Trovare un quantile da una probabilità

Supponiamo di voler trovare il valore x per cui la probabilità che una variabile con distribuzione Gamma $\gamma(\alpha, \beta)$ sia minore o uguale a x sia almeno del 5%.

1. **Impostare la disequazione:**

$$P(\text{gamma}(\alpha, \beta) \leq x) \geq 0.05$$

2. **Esprimere tramite CDF:**

$$F_{\text{gamma}}(x) \geq 0.05$$

3. **Applicare la funzione inversa della CDF (F_g^{-1}):** Poiché la CDF e la sua inversa sono funzioni crescenti, il verso della disuguaglianza non cambia.

$$F_g^{-1}(F_g(x)) \geq F_g^{-1}(0.05)$$

4. **Risolvere per x:**

$$x \geq F_g^{-1}(0.05)$$

La conclusione è che tutti i valori di x maggiori o uguali al quantile al 5% della distribuzione Gamma soddisfano la richiesta. In software come Excel, questo valore si calcola con la funzione `GAMMA.INV(0.05, alpha, beta)`.

1.3 Inversione con Due Incognite

Spesso si cerca un intervallo $[a, b]$ tale per cui la probabilità che una variabile aleatoria cada al suo interno sia pari a un valore prefissato (es. 90% o 95%), tipicamente per costruire intervalli di confidenza.

$$P(X \in [a, b]) = 1 - \alpha$$

Ad esempio, se vogliamo $P(a \leq X \leq b) = 90\%$, la probabilità totale nelle code (a sinistra di a e a destra di b) deve essere del 10%. Poiché ci sono infinite coppie (a, b) che soddisfano questa condizione, si usano dei criteri per scegliere una soluzione unica.

Nota 1.2: Criteri per la scelta dell'intervallo

I tre criteri più comuni sono:

- **Intervallo Canonico (o delle code uguali):** La probabilità residua α viene divisa equamente tra le due code. Se $\alpha = 10\%$, si pone il 5% sulla coda sinistra e il 5% sulla destra.
$$a = F_X^{-1}(0.05) \quad \text{e} \quad b = F_X^{-1}(0.95)$$
- **Intervallo Arbitrario:** La probabilità α viene suddivisa in modo arbitrario, purché la somma sia corretta. Ad esempio, si potrebbe avere una coda sinistra con il 7% di probabilità e una destra con il 3%.
- **Intervallo di ampiezza minima:** Si cerca l'intervallo $[a, b]$ che, a parità di probabilità interna, ha la larghezza $(b - a)$ più piccola possibile. Per una distribuzione unimodale continua, questo si ottiene quando la funzione di densità ha la stessa altezza agli estremi.
$$f(a) = f(b)$$

Nota 1.3: Caso delle distribuzioni simmetriche

Se la distribuzione di probabilità è simmetrica e unimodale (es. Normale, t di Student), l'intervallo **canonico** (a code uguali) coincide con l'intervallo di **ampiezza minima**.

1.4 Proprietà di Simmetria e Correzione per la Continuità

Nota 1.4: Simmetria della CDF per distribuzioni simmetriche

Per una distribuzione simmetrica attorno a zero, come la Normale standard $\mathcal{N}(0, 1)$, la CDF $\Phi(x)$ ha le seguenti proprietà:

$$\Phi(-x) = 1 - \Phi(x) \implies \Phi(x) + \Phi(-x) = 1$$

Questa proprietà si estende anche alla sua funzione inversa (la funzione quantile):

$$\Phi^{-1}(\alpha) = -\Phi^{-1}(1 - \alpha)$$

Ciò significa che il quantile di livello α è l'opposto del quantile di livello $1 - \alpha$.

Nota 1.5: Correzione per la Continuità

Quando si approssima una v.a. discreta (che assume valori interi) con una v.a. continua, si introduce un errore. La **correzione per la continuità** serve a ridurre questo errore. Per calcolare $P(X \leq k)$ dove X è una v.a. discreta, usando la CDF continua F come approssimazione, è più accurato calcolare F nel punto $k + 0.5$.

$$F_X(k) = P(X_{\text{discreta}} \leq k) \approx F_{\text{continua}}(k + 0.5)$$

Questo perché $k + 0.5$ è il punto medio tra k e $k + 1$, fornendo una stima migliore del valore della funzione a gradini discreta.

1.5 Generazione di v.a. con Legge Data

Il metodo dell'**Inverse Transform Sampling** (Campionamento tramite Inversione della Trasformata) permette di generare numeri casuali da qualsiasi distribuzione di probabilità di cui sia nota la CDF. Si basa sulla trasformazione integrale di probabilità.

Proposizione 1.1: Metodo dell'Inverse Transform Sampling

Sia F la CDF di una distribuzione target. Per generare un campione X da questa distribuzione:

1. Si genera un numero casuale U da una distribuzione Uniforme in $[0, 1]$.
2. Si calcola $X = F^{-1}(U)$, dove F^{-1} è la funzione quantile (l'inversa della CDF).

La variabile aleatoria X così generata avrà come distribuzione proprio quella descritta da F.

Esempio 1.2: Generazione da una distribuzione Esponenziale

Vogliamo generare un campione da una distribuzione Esponenziale di parametro λ.

1. **CDF:** $F(t) = 1 - e^{-\lambda t}$ per $t \geq 0$.
2. **Inversa della CDF:** Poniamo $y = 1 - e^{-\lambda t}$ e risolviamo per t.

$$1 - y = e^{-\lambda t} \implies \log(1 - y) = -\lambda t \implies t = -\frac{1}{\lambda} \log(1 - y)$$

   Quindi $F^{-1}(y) = -\frac{1}{\lambda} \log(1 - y)$.

3. **Generazione:** Generiamo $U \sim \text{Unif}(0, 1)$ e calcoliamo:

$$X = F^{-1}(U) = -\frac{1}{\lambda} \log(1 - U)$$

Dimostrazione 1.1: Giustificazione del metodo

Vogliamo dimostrare che la CDF di $X = F^{-1}(U)$ è proprio F. Sia $F_X(t)$ la CDF di X.

$$F_X(t) = P(X \leq t) = P(F^{-1}(U) \leq t)$$

Poiché F è una funzione crescente, possiamo applicarla a entrambi i lati della disuguaglianza senza cambiarne il verso:

$$P(F^{-1}(U) \leq t) = P(U \leq F(t))$$

Dato che $U \sim \text{Unif}(0, 1)$, per definizione la sua CDF è $F_U(y) = P(U \leq y) = y$. Sostituendo $y = F(t)$, otteniamo:

$$P(U \leq F(t)) = F(t)$$

Abbiamo quindi dimostrato che $F_X(t) = F(t)$, confermando che la variabile X generata ha la distribuzione desiderata.

□

1.6 CDF Empirica e Diagramma Q-Q

Quando non si conosce la vera distribuzione di un set di dati, si possono usare strumenti empirici per stimarla e visualizzarla.

Definizione 1.3: Funzione di Ripartizione Cumulativa Empirica (ECDF)

Dato un campione di n osservazioni $x_1, \ldots, x_n$, la **CDF Empirica** $\hat{F}_n(t)$ è la proporzione di osservazioni nel campione che sono minori o uguali a t.

$$\hat{F}_n(t) = \frac{\#\{i : x_i \leq t\}}{n}$$

Graficamente, è una funzione a gradini che aumenta di $1/n$ in corrispondenza di ogni dato osservato. La ECDF è una stima della vera CDF (sconosciuta) della popolazione da cui il campione è stato estratto.

Diagramma Q-Q (Quantile-Quantile) Il diagramma Q-Q è uno degli strumenti grafici più efficaci per confrontare la distribuzione dei dati con una distribuzione teorica. L'idea è quella di plottare i quantili campionari (i dati ordinati) contro i quantili teorici della distribuzione di riferimento.

Se i dati seguono la distribuzione teorica, i punti sul grafico si allineano lungo una retta. Per verificare l'ipotesi di normalità (o gaussianità), si costruisce il grafico plottando i punti:

$$\left(\Phi^{-1}\left(\frac{i - 0.5}{n} \right), x_{(i)} \right)$$

dove $x_{(i)}$ è l'i-esimo dato ordinato e Φ^{-1} è la funzione quantile della Normale standard.

Nota 1.6: Derivazione Matematica e Interpretazione

La linearità del grafico Q-Q per dati Normali deriva da una relazione matematica precisa.

Relazione Fondamentale. Supponiamo che i nostri dati x_i seguano una distribuzione Normale $\mathcal{N}(\mu, \sigma^2)$. Allora, ogni punto può essere espresso in termini di un quantile p_i e della funzione quantile inversa della Normale standard, Φ^{-1}.

$$p_i = \Phi\left(\frac{x_i - \mu}{\sigma}\right) \iff x_i = \mu + \sigma\Phi^{-1}(p_i)$$

Questa equazione mostra che ogni quantile dei nostri dati, x_i, è una funzione lineare del quantile corrispondente di una Normale standard, $\Phi^{-1}(p_i)$.

Costruzione del Grafico. Per costruire il grafico, confrontiamo i quantili campionari (i nostri dati ordinati) con i quantili teorici.

- **Asse Y (Quantili Campionari):** Utilizziamo i dati osservati e ordinati:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

- **Asse X (Quantili Teorici):** Dobbiamo calcolare i quantili teorici corrispondenti. Se i dati fossero Normali, le probabilità cumulate $p_{(i)}$ associate a ogni $x_{(i)}$ sarebbero uniformemente distribuite. Approssimiamo queste probabilità con le posizioni di plotting:

$$p_{(i)} \approx \frac{i - 0.5}{n}, \quad \text{per } i = 1, \ldots, n$$

  I quantili teorici della Normale standard sono quindi:

$$z_i = \Phi^{-1}\left(\frac{i - 0.5}{n}\right)$$

Interpretazione. Il diagramma Q-Q plotta i punti $(z_i, x_{(i)})$. Se l'ipotesi di normalità è vera, allora, sostituendo nell'equazione fondamentale, otteniamo:

$$x_{(i)} \approx \mu + \sigma\Phi^{-1}(p_{(i)}) \approx \mu + \sigma z_i$$

I punti del grafico $(z_i, x_{(i)})$ dovrebbero quindi giacere approssimativamente sulla retta $y = \mu + \sigma z$. Una deviazione sistematica da questa retta indica che i dati non seguono una distribuzione Normale.

2 Leggi di Variabili Aleatorie Importanti

2.1 Distribuzione Gaussiana (o Normale)

La distribuzione Gaussiana è una delle più importanti in statistica, descrivendo molti fenomeni naturali che risultano dalla somma di numerosi piccoli effetti indipendenti.

Definizione 2.1: Distribuzione Gaussiana

Una variabile aleatoria X segue una distribuzione Gaussiana con media μ e varianza σ^2, indicata con $X \sim \mathcal{N}(\mu, \sigma^2)$, se la sua funzione di densità di probabilità (PDF) è:

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

Nota 2.1: Teorema del Limite Centrale (TLC)

Il TLC afferma che le grandezze casuali che sono la **somma** di tanti piccoli contributi indipendenti tendono ad avere una distribuzione Normale.

Proposizione 2.1: Proprietà della Gaussiana

- **Chiusura per trasformazioni lineari:** Se $X \sim \mathcal{N}(\mu, \sigma^2)$, allora la trasformazione lineare $a + bX$ segue ancora una distribuzione Normale:
$$a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$$
- **Riproducibilità:** La somma di due v.a. Gaussiane *indipendenti* è ancora una v.a. Gaussiana. Se $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ sono indipendenti:
$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$
- **CDF Canonica:** La CDF di una qualsiasi Normale può essere ricondotta a quella della Normale standard $\Phi(t) = F_{\mathcal{N}(0,1)}(t)$:
$$F_{\mathcal{N}(\mu,\sigma^2)}(x) = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

2.2 Distribuzione Lognormale

La Lognormale descrive grandezze che derivano dal *prodotto* di molti fattori casuali.

Definizione 2.2: Distribuzione Lognormale

Una variabile aleatoria X è Lognormale se il suo logaritmo naturale è una v.a. Normale. Se $Y = \log X \sim \mathcal{N}(\mu, \sigma^2)$, allora si scrive $X \sim \text{Lognorm}(\mu, \sigma^2)$. In altre parole, è l'esponenziale di una Gaussiana:

$$X = e^Y, \quad \text{con } Y \sim \mathcal{N}(\mu, \sigma^2)$$

Nota 2.2: Origine della Lognormale

Una versione "moltiplicativa" del TLC afferma che le grandezze che sono il **prodotto** di tanti piccoli contributi indipendenti tendono ad avere una distribuzione Lognormale. Esempi tipici sono redditi, patrimoni e dimensioni di frammenti.

Proposizione 2.2: Proprietà della Lognormale

- **Asimmetria:** È una distribuzione asimmetrica con una coda lunga a destra, adatta a modellare quantità che non possono essere negative ma possono assumere valori molto grandi.
- **Analisi dei dati:** Per analizzare dati lognormali, è pratica comune calcolarne il logaritmo e trattare i dati trasformati come Normali.
- **Media e Mediana:** A differenza della Normale, media e mediana non coincidono.

$$E(X) = e^{\mu + \frac{\sigma^2}{2}}, \quad \text{Mediana}(X) = e^{\mu}$$

2.3 Distribuzione Esponenziale

L'Esponenziale è la distribuzione di base per i tempi di attesa.

Definizione 2.3: Distribuzione Esponenziale

Una v.a. T segue una distribuzione Esponenziale di parametro $\lambda > 0$, indicata con $T \sim \text{Expo}(\lambda)$, se la sua PDF è:

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$

Il parametro λ è chiamato **tasso** o **intensità**, e rappresenta il numero medio di eventi nell'unità di tempo.

Proposizione 2.3: Proprietà dell'Esponenziale

- **Versione continua della Geometrica:** Rappresenta il tempo di attesa per un evento che ha la stessa "probabilità" infinitesima di avvenire in ogni istante.
- **Assenza di memoria:** La probabilità di attendere ancora non dipende da quanto si è

già atteso. Formalmente:

$$P(T > a + b \mid T > a) = P(T > b)$$

- **Applicazioni:** Modella tempi di attesa per eventi improvvisi e imprevedibili come guasti, telefonate, decadimenti radioattivi.

2.4 Distribuzione Gamma

La distribuzione Gamma è una generalizzazione dell'Esponenziale e della Chi-Quadro, molto flessibile per modellare tempi di attesa.

Definizione 2.4: Distribuzione Gamma

Una v.a. X segue una distribuzione Gamma definita da un parametro di forma $\alpha > 0$ e un parametro di tasso $\lambda > 0$. Si indica $X \sim \text{Gamma}(\alpha, \lambda)$. La sua PDF è:

$$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, \quad t > 0$$

dove $\Gamma(\alpha)$ è la funzione Gamma di Eulero. Una parametrizzazione alternativa usa un parametro di scala $\beta = 1/\lambda$.

Proposizione 2.4: Proprietà della Gamma

- **Riproducibilità:** La somma di v.a. Gamma indipendenti con lo stesso tasso λ è ancora una Gamma. Se $X \sim \text{Gamma}(\alpha_1, \lambda)$ e $Y \sim \text{Gamma}(\alpha_2, \lambda)$ sono indipendenti:

$$X + Y \sim \text{Gamma}(\alpha_1 + \alpha_2, \lambda)$$

- **Scalatura:** Se $X \sim \text{Gamma}(\alpha, \lambda)$, allora $cX \sim \text{Gamma}(\alpha, \lambda/c)$.
- **Media e Varianza:**

$$E(X) = \frac{\alpha}{\lambda}, \quad \text{Var}(X) = \frac{\alpha}{\lambda^2}$$

Nota 2.3: Casi Particolari della Gamma

- **Esponenziale:** Per $\alpha = 1$, la Gamma diventa un'Esponenziale: $\text{Gamma}(1, \lambda) \equiv \text{Expo}(\lambda)$.
- **Erlang:** Se $\alpha = n$ è un intero, la distribuzione si chiama Erlang ed è la somma di n v.a. $\text{Expo}(\lambda)$ indipendenti.
- **Chi-Quadro:** È un altro caso speciale, fondamentale in statistica.

2.5 Distribuzione Chi-Quadro

Definizione 2.5: Distribuzione Chi-Quadro

Una v.a. W segue una distribuzione Chi-Quadro con k gradi di libertà, $W \sim \chi^2(k)$, se è la somma dei quadrati di k v.a. Normali standard indipendenti.

$$W = \sum_{i=1}^{k} Z_i^2, \quad \text{dove } Z_i \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

È un caso particolare della Gamma: $\chi^2(k) \equiv \text{Gamma}(k/2, 1/2)$.

Nota 2.4: Proprietà della Chi-Quadro

Derivando dalla Gamma, si ottiene:

$$E(W) = k, \quad \text{Var}(W) = 2k$$

Inoltre, poiché $Z^2 \sim \chi^2(1)$, si ha che $\chi^2(1) \equiv \text{Gamma}(1/2, 1/2)$.

2.6 Il Processo di Poisson

Il processo di Poisson descrive il verificarsi di eventi casuali nel tempo.

Definizione 2.6: Processo di Poisson

Un processo di Poisson è una sequenza di eventi istantanei tali che i tempi di inter-arrivo tra un evento e il successivo sono variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) come un'Esponenziale di tasso λ.

Proposizione 2.5: Componenti del Processo di Poisson

- **Tempi di inter-arrivo T_i:** Il tempo tra l'evento $i-1$ e l'evento i.

$$T_i \sim \text{Expo}(\lambda) \text{ i.i.d.}$$

- **Tempo di arrivo dell'n-esimo evento S_n:** È la somma dei primi n tempi di inter-arrivo.

$$S_n = \sum_{i=1}^{n} T_i \sim \text{Gamma}(n, \lambda)$$

- **Numero di eventi in $[0,t]$, N_t:** Il conteggio degli eventi fino a un certo istante t. Segue una distribuzione di Poisson.

$$N_t \sim \text{Pois}(\lambda t)$$

2.7 Distribuzione Binomiale

Definizione 2.7: Distribuzione Binomiale

Una variabile aleatoria discreta X segue una distribuzione Binomiale con parametri n (un intero positivo) e $p \in [0,1]$, indicata con $X \sim \text{Bin}(n,p)$, se la sua funzione di massa di probabilità (PMF) è:

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

Proposizione 2.6: Proprietà della Binomiale

- Il caso con $n = 1$ è detto distribuzione **Bernoulliana**.
- La Binomiale rappresenta il numero di successi in n prove indipendenti, ognuna con la stessa probabilità di successo p. È la somma di n v.a. di Bernoulli i.i.d..
- **Riproducibilità:** La somma di v.a. Binomiali indipendenti con lo stesso parametro p è ancora una v.a. Binomiale. Se $X_i \sim \text{Bin}(n_i, p)$ sono indipendenti:

$$\sum_{i=1}^{m} X_i \sim \text{Bin}\left(\sum_{i=1}^{m} n_i, p\right)$$

- **Media e Varianza:**

$$E(X) = np, \quad \text{Var}(X) = np(1-p)$$

2.8 Distribuzione di Poisson

Definizione 2.8: Distribuzione di Poisson

Una v.a. discreta X segue una distribuzione di Poisson di parametro $\nu > 0$, indicata con $X \sim \text{Pois}(\nu)$, se la sua PMF è:

$$P(X=k) = \frac{\nu^k e^{-\nu}}{k!}, \quad k = 0, 1, 2, \ldots$$

Nota 2.5: Relazione con la Binomiale

La distribuzione di Poisson è il limite della Binomiale per n grande e p piccolo. In pratica, se $p \ll 1$, allora $\text{Bin}(n,p) \approx \text{Pois}(np)$. Per questo motivo, è usata per contare il numero di successi (eventi rari) in scenari con un gran numero di prove, come il numero di gol in una partita o il numero di iscritti a un corso.

Proposizione 2.7: Proprietà della Poisson

- **Riproducibilità:** La somma di v.a. di Poisson indipendenti è ancora una v.a. di Poisson. Se $X_i \sim \text{Pois}(\nu_i)$ sono indipendenti:

$$\sum_{i=1}^{m} X_i \sim \text{Pois}\left(\sum_{i=1}^{m} \nu_i\right)$$

- **Media e Varianza:** A differenza della Binomiale, media e varianza coincidono.

$$E(X) = \nu, \quad \text{Var}(X) = \nu$$

2.9 Distribuzione Uniforme

Definizione 2.9: Distribuzione Uniforme Continua

Una v.a. X segue una distribuzione Uniforme sull'intervallo $[a, b]$, con $a < b$ reali, se la sua PDF è costante su quell'intervallo e zero altrove.

$$f_X(t) = \frac{1}{b-a}, \quad \text{per } a < t < b$$

La funzione `rand()` nei linguaggi di programmazione genera tipicamente campioni da $\text{Unif}(0, 1)$.

Proposizione 2.8: Proprietà dell'Uniforme

- **Media e Varianza:**

$$E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

- È una classe chiusa per trasformazioni lineari.

2.10 Distribuzione Beta

Definizione 2.10: Distribuzione Beta

Una v.a. X segue una distribuzione Beta con parametri di forma $\alpha, \beta > 0$, se la sua PDF è:

$$f_X(t) = c_p t^{\alpha-1}(1-t)^{\beta-1}, \quad \text{per } 0 < t < 1$$

È una distribuzione molto flessibile, definita su un intervallo limitato.

Nota 2.6: Interpretazione della Beta

- Il caso speciale Beta$(1, 1)$ corrisponde alla distribuzione Unif$(0, 1)$.
- **Statistica d'ordine:** La distribuzione Beta(m, n) è la distribuzione della m-esima variabile più piccola tra $m + n - 1$ v.a. Unif$(0, 1)$ indipendenti.

2.11 Distribuzione t di Student

Definizione 2.11: Distribuzione t di Student

La distribuzione t di Student con k gradi di libertà, $t(k)$, è definita operativamente dal rapporto tra una v.a. Normale standard e la radice di una Chi-Quadro indipendente, divisa per i suoi gradi di libertà. Se $Z \sim \mathcal{N}(0, 1)$ e $W \sim \chi^2(k)$ sono indipendenti, allora:

$$T = \frac{Z}{\sqrt{W/k}} \sim t(k)$$

Proposizione 2.9: Proprietà della t di Student

- Ha una forma a campana simile alla Normale, ma con code più "pesanti".
- Per $k \to \infty$, la distribuzione $t(k)$ converge alla Normale standard $\mathcal{N}(0, 1)$.
- La media è $E(T) = 0$ (per $k > 1$).

2.12 Distribuzione F di Fisher

Definizione 2.12: Distribuzione F di Fisher

La distribuzione F di Fisher con (m, n) gradi di libertà, $F(m, n)$, è definita operativamente come il rapporto tra due v.a. Chi-Quadro indipendenti, ciascuna divisa per i propri gradi di libertà. Se $W_1 \sim \chi^2(m)$ e $W_2 \sim \chi^2(n)$ sono indipendenti, allora:

$$F = \frac{W_1/m}{W_2/n} \sim F(m, n)$$

Nota 2.7: Uso della Distribuzione F

È usata principalmente per confrontare due varianze campionarie. I parametri m e n sono i gradi di libertà del numeratore e del denominatore, rispettivamente.

2.13 Distribuzione Uniforme Discreta

Definizione 2.13: Distribuzione Uniforme Discreta

Una v.a. X segue una distribuzione Uniforme Discreta se può assumere n valori, $\{1, 2, \ldots, n\}$, ciascuno con la stessa probabilità. L'esempio classico è il lancio di un dado a n facce.

$$P(X = i) = \frac{1}{n}, \quad \text{per } i = 1, 2, \ldots, n$$

Proposizione 2.10: Proprietà dell'Uniforme Discreta

- **Non è riproducibile:** La somma di due o più v.a. uniformi discrete indipendenti non è più uniforme. La sua distribuzione tende a una forma a campana (analogo discreto del TLC).
- **Media e Varianza:**

$$E(X) = \frac{n+1}{2}, \quad \text{Var}(X) = \frac{n^2 - 1}{12}$$

2.14 Processo di Bernoulli e Distribuzioni Associate

Molte distribuzioni discrete di base emergono dal **Processo di Bernoulli**, l'analogo a tempo discreto del Processo di Poisson.

Definizione 2.14: Processo di Bernoulli

Un processo di Bernoulli è una sequenza di prove o esperimenti indipendenti, ciascuno con due soli esiti possibili: "successo" (con probabilità p) o "insuccesso" (con probabilità $1 - p$).

Nota 2.8: Variabili aleatorie in un Processo di Bernoulli

All'interno di un processo di Bernoulli si possono definire diverse variabili aleatorie di interesse:

- **Numero di successi N_n:** Il conteggio dei successi in n prove. Segue una distribuzione $\text{Bin}(n, p)$.
- **Tempo del primo successo T_1:** Il numero di prove necessarie per ottenere il primo successo. Segue una distribuzione $\text{Geom}(p)$.
- **Tempo dell'm-esimo successo S_m:** Il numero di prove necessarie per ottenere l'm-esimo successo. Segue una distribuzione $\text{NegBin}(m, p)$.

2.15 Distribuzione Geometrica

Definizione 2.15: Distribuzione Geometrica

Una v.a. X segue una distribuzione Geometrica di parametro $p \in (0, 1]$ se la sua PMF è:

$$P(X = k) = p(1 - p)^{k-1}, \quad k = 1, 2, 3, \dots$$

Rappresenta il numero di prove necessarie per ottenere il primo successo in un processo di Bernoulli.

Proposizione 2.11: Proprietà della Geometrica

- È la versione discreta della distribuzione Esponenziale.
- **Media e Varianza:**

$$E(X) = \frac{1}{p}, \quad \text{Var}(X) = \frac{1 - p}{p^2}$$

2.16 Distribuzione Binomiale Negativa

Questa distribuzione generalizza la Geometrica al caso di r successi.

Definizione 2.16: Distribuzione Binomiale Negativa

Esistono due definizioni comuni per la Binomiale Negativa $\text{NegBin}(r, p)$:

1. **Numero di prove:** X è il numero totale di prove per ottenere r successi. La sua PMF è:

$$P(X = k) = \binom{k-1}{r-1} p^r (1 - p)^{k-r}, \quad k = r, r + 1, \dots$$

2. **Numero di insuccessi:** $\tilde{X}$ è il numero di insuccessi che avvengono prima di ottenere r successi. La sua PMF, usata spesso in software come SciPy, è:

$$P(\tilde{X} = k) = \binom{k+r-1}{k} p^r (1 - p)^k, \quad k = 0, 1, 2, \dots$$

Proposizione 2.12: Proprietà della Binomiale Negativa

- Se $r = 1$, si ottiene la distribuzione Geometrica.
- **Riproducibilità:** La somma di v.a. Binomiali Negative indipendenti con lo stesso p è ancora una Binomiale Negativa. Se $X_i \sim \text{NegBin}(r_i, p)$ sono indipendenti:

$$\sum X_i \sim \text{NegBin}\left(\sum r_i, p\right)$$

- **Media e Varianza (per la def. 1):**

$$E(X) = \frac{r}{p}, \quad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$$

3 Simulazione Monte Carlo

3.1 Il Problema: Valori Attesi e Probabilità Complesse

Spesso in statistica e machine learning ci si imbatte in variabili aleatorie la cui complessità rende difficile o impossibile calcolare analiticamente il loro valore atteso o la probabilità di un evento.

Esempio 3.1: Tempo di superamento di una soglia

Consideriamo una successione di variabili aleatorie $U_1, U_2, \ldots$ indipendenti e identicamente distribuite come Unif$(0,1)$. Definiamo le somme parziali $S_n = \sum_{i=1}^{n} U_i$. Vogliamo studiare la variabile T, che rappresenta il "tempo" (numero di passi) necessario per superare una soglia $a > 0$:

$$T := \inf\{n : S_n \geq a\}$$

La distribuzione di T è complicata e dipende da a. Come possiamo stimare quantità come il suo valore atteso $E(T)$ o la probabilità $P(T \leq c)$?

Altri esempi includono il calcolo dell'entropia di una distribuzione, la stima del valore atteso di trasformazioni non lineari di variabili (es. $E(\sqrt{Z})$ con $Z \sim \mathcal{N}$), o il calcolo delle probabilità degli esiti di un processo complesso come una battaglia nel gioco "Risiko". In tutti questi casi, una soluzione analitica è impraticabile e si ricorre a metodi numerici o stocastici.

3.2 Soluzione Stocastica: Il Metodo Monte Carlo

Definizione 3.1: Metodo Monte Carlo

Il metodo Monte Carlo è una tecnica computazionale che permette di ottenere risultati numerici approssimati utilizzando campionamenti casuali. Per stimare il valore atteso $E(X)$ di una v.a. X:

1. Si generano N campioni indipendenti $X_1, X_2, \ldots, X_N$ dalla distribuzione di X.
2. Si stima $E(X)$ con la media campionaria $\bar{X}$.

$$E(X) \approx \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Per la Legge dei Grandi Numeri (LGN), questa stima converge al valore vero per $N \to \infty$.

3.3 Intervalli di Confidenza per le Stime

Una stima puntuale ($\bar{X}$) non basta; un intervallo di confidenza ci dà una misura dell'incertezza della nostra stima.

Proposizione 3.1: Intervallo di Confidenza per la Media

Un intervallo di confidenza per $E(X)$ a un livello $1-\alpha$ è dato da:

$$E(X) \in \bar{X} \pm q\frac{S_X}{\sqrt{N}}$$

dove $\bar{X}$ è la media campionaria, S_X è la deviazione standard campionaria, N è la dimensione del campione, e q è il quantile appropriato (es. $q \approx 1.96$ per un livello di confidenza del 95%). Se l'intervallo risulta troppo ampio, si può ridurre la sua ampiezza aumentando N.

3.4 Stima di Probabilità

Il metodo Monte Carlo può essere usato anche per stimare una probabilità $P(A)$. La tecnica consiste nel ricondurre il calcolo della probabilità a quello di un valore atteso.

Nota 3.1: Probabilità come Valore Atteso

Si definisce una variabile aleatoria indicatore (o di Bernoulli) $\mathbb{I}_A$ tale che:

$$\mathbb{I}_A = \begin{cases} 1 & \text{se l'evento A si verifica} \\ 0 & \text{altrimenti} \end{cases}$$

Il suo valore atteso è esattamente la probabilità di A: $E(\mathbb{I}_A) = P(A)$.

Per stimare $P(A)$, si eseguono N simulazioni e si calcola la frequenza relativa f_c con cui A si è verificato. Questa frequenza è la media campionaria delle v.a. indicatore ed è la nostra stima della probabilità.

$$P(A) \approx \hat{p} = f_c = \frac{\#\{\text{volte in cui A si è verificato}\}}{N}$$

L'intervallo di confidenza per la probabilità p diventa quindi:

$$p \in \hat{p} \pm q\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

3.5 Soluzione Numerica (Alternativa alla Simulazione)

Nota 3.2: Metodi Numerici vs. Monte Carlo

Per i problemi visti in precedenza, esistono approcci numerici deterministici che si contrappongono alla soluzione stocastica Monte Carlo. La loro fattibilità, tuttavia, dipende strettamente dalla natura del problema.

- **Integrale Numerico:** Per calcolare un valore atteso definito da un integrale, come nel caso di $E(\sqrt{Z})$ per $Z \sim \mathcal{N}(\mu, \sigma^2)$, si può ricorrere alla quadratura numerica. Questa tecnica approssima l'area sottesa alla curva della funzione integranda sommando le aree di un gran numero di rettangoli o altri poligoni semplici.

- **Approssimazione di Somme Infinite:** Per calcolare quantità come l'entropia, che richiedono una somma su infiniti termini, si adotta un approccio di troncamento. La somma viene calcolata solo fino a un certo punto, fermandosi quando i termini successivi diventano così piccoli da non contribuire in modo significativo al risultato finale.
- **Esaustione dei Casi:** Per problemi con uno spazio degli esiti finito e discreto, come la battaglia nel gioco "Risiko", è teoricamente possibile calcolare le probabilità esatte enumerando tutti i casi possibili. Nel caso di 3 dadi contro 3, questo comporterebbe l'analisi di $6^6 = 46556$ combinazioni.
- **Casi Inattuabili:** Problemi ad alta dimensionalità, come il calcolo del "tempo di superamento di una soglia" (esempio Esempio 3.1), sono spesso impossibili da risolvere con questi metodi deterministici. In questi scenari, la simulazione Monte Carlo rimane l'approccio più efficace e, a volte, l'unico praticabile.

4 Vettori Aleatori e Machine Learning

4.1 Apprendimento Supervisionato (Supervised Learning)

Definizione 4.1: Apprendimento Supervisionato

Nell'apprendimento supervisionato, l'obiettivo è predire una o più variabili di output a partire da un insieme di variabili di input (dette predittori).

Esempio 4.1: Casi d'uso

Alcune applicazioni pratiche includono:

- Prevedere le risorse necessarie (tempo, soldi, energia) per una commessa o un progetto.
- Prevedere il peso corporeo di un soggetto basandosi su altre misure fisiche.
- Classificare il contenuto di un'immagine, assegnandola a una categoria specifica.

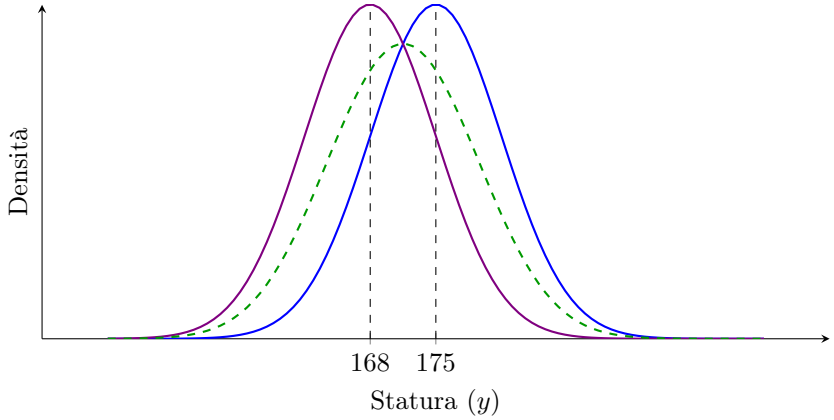

Figura 1: Illustrazione della distribuzione della statura Y condizionata dal sesso X. Le curve continue sono le densità condizionate per maschi ($\mu = 175$) e femmine ($\mu = 168$). La curva tratteggiata è la densità marginale di Y.

Il concetto fondamentale è imparare la relazione che lega l'input X all'output Y. Ad esempio, si può modellare la relazione tra il sesso di una persona (X) e la sua statura (Y). In questo caso, X è una variabile discreta (es. 0 per maschio, 1 for femmina) e Y è una variabile continua. Dopo aver appreso il modello, per un dato input si ottiene la distribuzione dell'output. Per $X = 0$ (maschio), la statura Y potrebbe seguire una distribuzione $\mathcal{N}(175, 7^2)$, mentre per $X = 1$ (femmina), $Y \sim \mathcal{N}(168, 7^2)$. Questa relazione è descritta dalla distribuzione condizionata $f_{Y|X}(y|x)$.

Nota 4.1: Indipendenza delle Variabili

Se una variabile di input non ha dipendenza con la variabile di output (es. colore degli occhi rispetto alla statura), la predizione per l'output Y non sarà condizionata da tale input, ma risulterà in una distribuzione mista (mixture).

4.2 Apprendimento non Supervisionato (Unsupervised Learning)

Definizione 4.2: Apprendimento non Supervisionato

Nell'apprendimento non supervisionato, non ci sono variabili di output predefinite. L'obiettivo è esplorare i dati per capire la loro distribuzione multivariata e scoprire pattern o strutture intrinseche.

Esempio 4.2: Clusterizzazione di Cellule

Un'applicazione tipica è la clusterizzazione di dati biologici. Ad esempio, partendo da un dataset dove le righe sono cellule e le colonne sono l'espressione di migliaia di geni, l'obiettivo è:

- Trovare le relazioni tra le variabili (i geni).
- Scoprire come si raggruppano le cellule in base a questi pattern.
- Cercare di identificare e dare un significato a questi cluster.

4.3 La Matrice di Covarianza

Definizione 4.3: Matrice di Covarianza

Dato un vettore aleatorio $X = (X_1, \ldots, X_m)$, la sua matrice di covarianza, indicata con Σ o $C(X)$, è una matrice $m \times m$ i cui elementi rappresentano la covarianza tra le componenti del vettore. L'elemento (i, j) della matrice è definito come:

$$\Sigma_{ij} := \mathrm{Cov}(X_i, X_j)$$

Proposizione 4.1: Proprietà della Matrice di Covarianza

- È una matrice **simmetrica**, poiché $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$.
- Sulla diagonale principale si trovano le **varianze** delle singole componenti, $\Sigma_{ii} = \mathrm{Var}(X_i)$, che sono sempre non negative.
- Se le componenti $X_1, \ldots, X_m$ sono **indipendenti** tra loro, la matrice di covarianza è **diagonale**, in quanto tutte le covarianze tra variabili diverse sono nulle.

Nota 4.2: Ripasso sulla Covarianza

L'operatore covarianza $\mathrm{Cov}(\cdot, \cdot)$ è bilineare e ha le seguenti proprietà:

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
- Se X e Y sono indipendenti, allora $\mathrm{Cov}(X, Y) = 0$.
- La covarianza con una costante è zero: $\mathrm{Cov}(X, \mathrm{cost}) = 0$.

5 Trasformazioni Lineari di Vettori Aleatori

5.1 Definizione

Una trasformazione lineare è una funzione che mappa un vettore da uno spazio a un altro tramite operazioni di rotazione, scalatura e traslazione.

Definizione 5.1: Trasformazione Lineare di un Vettore Aleatorio

Dato un vettore aleatorio $X \in \mathbb{R}^m$, una trasformazione lineare $g : \mathbb{R}^m \to \mathbb{R}^k$ è definita come:

$$Y = g(X) = \alpha + BX$$

dove $\alpha \in \mathbb{R}^k$ è un vettore di costanti (traslazione) e $B \in M_{k,m}$ è una matrice di costanti (rotazione, scalatura, deformazione).

Nota 5.1: Effetto Geometrico

L'effetto di una trasformazione lineare sulla distribuzione di un vettore aleatorio può essere visualizzato geometricamente:

- Il vettore α causa una **traslazione** della distribuzione nello spazio.
- La matrice B applica una **rotazione** e una **deformazione**. Se la matrice B è diagonale, l'effetto è una semplice scalatura indipendente su ciascuna componente.

5.2 Trasformazione di Media e Covarianza

Quando si applica una trasformazione lineare a un vettore aleatorio, anche la sua media e la sua matrice di covarianza si trasformano secondo regole precise.

Proposizione 5.1: Trasformazione della Media

Sia X un vettore aleatorio con media $\mu_X = E(X)$. La media del vettore trasformato $Y = \alpha + BX$ è:

$$\mu_Y = \alpha + B\mu_X$$

Dimostrazione 5.1

La dimostrazione segue dalla definizione di prodotto matrice-vettore e dalla linearità del valore atteso applicata a ogni componente.
Consideriamo la componente i-esima del vettore Y:

$$Y_i = \alpha_i + [BX]_i = \alpha_i + \sum_{j=1}^{m} B_{ij} X_j$$

Calcoliamo il valore atteso di Y_i per ottenere la componente i-esima della media μ_Y:

$$\begin{aligned}
[\mu_Y]_i &= E[Y_i] \\
&= E\left[\alpha_i + \sum_{j=1}^{m} B_{ij} X_j\right] \\
&= E[\alpha_i] + E\left[\sum_{j=1}^{m} B_{ij} X_j\right] && \text{(per linearità di } E\text{)} \\
&= \alpha_i + \sum_{j=1}^{m} B_{ij} E[X_j] && (B_{ij} \text{ costanti}) \\
&= \alpha_i + \sum_{j=1}^{m} B_{ij} [\mu_X]_j
\end{aligned}$$

L'ultima espressione, $\alpha_i + \sum_{j=1}^{m} B_{ij}[\mu_X]_j$, è per definizione la componente i-esima del vettore $\alpha + B\mu_X$. Poiché $[\mu_Y]_i = [\alpha + B\mu_X]_i$ vale per ogni componente i, l'uguaglianza tra i vettori è dimostrata.

□

Proposizione 5.2: Trasformazione della Matrice di Covarianza

Sia X un vettore aleatorio con matrice di covarianza $\Sigma_X = C(X)$. La matrice di covarianza del vettore trasformato $Y = \alpha + BX$ è:

$$\Sigma_Y = B\Sigma_X B^{\mathsf{T}}$$

Dimostrazione 5.2

Il termine di traslazione α non influenza la covarianza. Calcoliamo l'elemento (i, j) della matrice Σ_Y, ricordando che $Y_i = \alpha_i + \sum_k B_{ik} X_k$:

$$\begin{aligned}
[\Sigma_Y]_{ij} &= \mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}\left(\alpha_i + \sum_k B_{ik} X_k, \alpha_j + \sum_h B_{jh} X_h\right) \\
&= \mathrm{Cov}\left(\sum_k B_{ik} X_k, \sum_h B_{jh} X_h\right) && \text{(per la bilinearità della covarianza)} \\
&= \sum_k \sum_h B_{ik} B_{jh} \mathrm{Cov}(X_k, X_h) \\
&= \sum_k \sum_h B_{ik} [\Sigma_X]_{kh} B_{jh}
\end{aligned}$$

Questa espressione corrisponde esattamente all'elemento (i, j) del prodotto matriciale $B\Sigma_X B^{\mathsf{T}}$.

□

5.3 Il Coefficiente di Correlazione Lineare

Definizione 5.2: Coefficiente di Correlazione Lineare

Date due variabili aleatorie X e Y con varianze finite e non nulle, il loro coefficiente di correlazione lineare (di Pearson), denotato con $\rho(X, Y)$, è definito come:

$$\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Nota 5.2: Proprietà della Correlazione

- È una misura adimensionale normalizzata della covarianza, con valori sempre compresi nell'intervallo $[-1, 1]$.
- Misura la forza e la direzione della **relazione lineare** tra due variabili.
- Un valore vicino a $+1$ indica una forte correlazione positiva, vicino a -1 una forte correlazione negativa, e vicino a 0 un'assenza di correlazione lineare.

5.4 Trasformazioni Lineari Frequenti

Alcune trasformazioni lineari sono usate così spesso nel pre-processing dei dati da meritare una menzione speciale.

5.4.1 Centrare il Vettore

L'obiettivo di questa trasformazione è spostare la distribuzione dei dati in modo che la sua media sia il vettore nullo.

- **Trasformazione:** Si sottrae il vettore delle medie μ_X dal vettore aleatorio X.

$$Y = X - \mu_X$$

- **Forma Matriciale:** Corrisponde a $Y = \alpha + BX$ con $\alpha = -\mu_X$ e $B = I$ (matrice identità).
- **Risultato:** La nuova media è nulla, mentre la matrice di covarianza rimane invariata.

$$\mu_Y = E(Y) = 0, \quad \Sigma_Y = C(Y) = \Sigma_X$$

- **A livello di campione:** Questa operazione equivale a sottrarre da ogni dato la media della sua colonna: $Y_{ij} = X_{ij} - \bar{X}_j$.

5.4.2 Standardizzare le Varianze

L'obiettivo è scalare le componenti del vettore in modo che abbiano tutte varianza unitaria (pari a 1).

- **Trasformazione:** Si divide ogni componente X_i per la sua deviazione standard $\text{std}(X_i)$.

$$Y_i = \frac{X_i}{\text{std}(X_i)}$$

- **Forma Matriciale:** Corrisponde a $Y = BX$ dove B è una matrice diagonale contenente le inverse delle deviazioni standard:

$$B = \text{diag}(\text{std}(X_1)^{-1}, \ldots, \text{std}(X_m)^{-1})$$

- **Risultato:** La matrice di covarianza del vettore trasformato Y diventa la **matrice di correlazione** del vettore originale X. L'elemento (i, j) di $C(Y)$ è:

$$[C(Y)]_{ij} = \text{Cov}\left(\frac{X_i}{\text{std}(X_i)}, \frac{X_j}{\text{std}(X_j)}\right) \overset{\text{(per bilinearità della Cov)}}{=} \frac{\text{Cov}(X_i, X_j)}{\text{std}(X_i)\text{std}(X_j)} = \rho(X_i, X_j)$$

Standardizzazione Completa (Z-score) Questa trasformazione combina le due precedenti per ottenere un vettore le cui componenti hanno media 0 e varianza 1.

- **Trasformazione:** È l'operazione nota come calcolo dello Z-score.

$$Y_i = \frac{X_i - E(X_i)}{\text{std}(X_i)}$$

- **Forma Matriciale:** Si applicano in ordine la centratura e la standardizzazione della varianza.

$$Y = \text{diag}(\text{std}(X)^{-1})(X - \mu_X)$$

- **Risultato:** Il vettore trasformato Y ha media nulla e la sua matrice di covarianza coincide con la matrice di correlazione di X.

$$\mu_Y = 0, \quad \Sigma_Y = \rho(X)$$

6 Principal Component Analysis (PCA)

6.1 Richiami utili

Nota 6.1: Spettro di una matrice simmetrica

Visto che $\Sigma \in \mathcal{M}_{n\times n}$ è simmetrica e definita positiva (come ogni matrice di covarianza), allora:

- ammette n autovalori reali e non negativi: $\lambda_i \geq 0$;
- esiste una base ortonormale di autovettori v_i;
- gli autovettori sono ortogonali tra loro;
- $\Sigma v_i = \lambda_i v_i$ equivale a dire che $\Sigma = V\Lambda V^\mathsf{T}$.

6.2 Definizione

La **Principal Component Analysis (PCA)** è una trasformazione lineare che serve a decorrelare le componenti di un dataset multidimensionale. È comunemente utilizzata per la riduzione della dimensionalità, il pre-processing e la visualizzazione.

Definizione 6.1: Componenti principali

La PCA consiste nel trovare una base ortonormale nello spazio dei dati tale che:

- le nuove coordinate (*componenti principali*) siano incorrelate tra loro;
- la prima componente abbia la massima varianza possibile;
- ogni successiva componente massimizzi la varianza residua, mantenendosi ortogonale alle precedenti.

6.3 Interpretazione geometrica

La PCA applica una **rotazione dello spazio** dei dati centrati (cioè con media nulla), allineando gli assi principali con le direzioni di massima varianza. Se $X \in \mathbb{R}^{n\times d}$ è il dataset centrato:

$$Y = V^\mathsf{T} X$$

dove $V \in \mathbb{R}^{d\times d}$ è la matrice degli autovettori di $\Sigma = \mathrm{Cov}(X)$. Le nuove variabili Y sono scorrelate e ordinate per varianza decrescente.

6.4 Autovalori e autovettori della covarianza

Teorema 6.1: Teorema spettrale per la matrice di covarianza

Sia $\Sigma \in \mathbb{R}^{d\times d}$ una matrice di covarianza, cioè reale, simmetrica e definita positiva. Allora esistono:

- una base ortonormale di autovettori $v_1, \dots, v_d \in \mathbb{R}^d$;
- autovalori reali non negativi $\lambda_1, \dots, \lambda_d \geq 0$;

tali che:

$$\Sigma v_k = \lambda_k v_k, \qquad v_j^\mathsf{T} v_k = \delta_{jk}$$

cioè:

$$\Sigma = V \Lambda V^\mathsf{T}$$

dove $V = [v_1 \ \dots \ v_d]$ è ortogonale e $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$.

Proposizione 6.1: La matrice degli autovettori è ortogonale

Sia $\Sigma \in \mathbb{R}^{d\times d}$ una matrice simmetrica. Siano $v_1, \dots, v_d \in \mathbb{R}^d$ autovettori ortonormali di Σ, e si definisca $V := [v_1 \ \dots \ v_d] \in \mathbb{R}^{d\times d}$. Allora:

$$V^\mathsf{T} V = I \quad \text{e} \quad V V^\mathsf{T} = I$$

cioè V è una matrice ortogonale.

Dimostrazione 6.1

Osserviamo che l'elemento (i, j) della matrice $V^\mathsf{T} V$ si calcola come:

$$[V^\mathsf{T} V]_{ij} = \sum_{k=1}^{d} [V^\mathsf{T}]_{ik} [V]_{kj} = \sum_{k=1}^{d} v_{k,i} v_{k,j}$$

Notiamo che:

$$\sum_{k=1}^{d} v_{k,i} v_{k,j} = v_i^\mathsf{T} v_j = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases} = \delta_{ij}$$

Quindi:

$$V^\mathsf{T} V = I \quad \text{(matrice identità)}$$

Segue che V è ortogonale, quindi $V^\mathsf{T} = V^{-1}$. Da questo deduciamo anche $V V^\mathsf{T} = I$, e quindi:

$$V V^\mathsf{T} = V V^{-1} = I$$

In conclusione, V è una rotazione (o riflessione), cioè una trasformazione ortogonale dello spazio. □

Proposizione 6.2: Diagonalizzazione spettrale: $\Sigma V = V\Lambda$

Sia $\Sigma \in \mathbb{R}^{d\times d}$ una matrice simmetrica, e siano $v_1, \dots, v_d$ i suoi autovettori ortonormali associati agli autovalori $\lambda_1, \dots, \lambda_d$. Costruiamo:

$$V := [v_1\ v_2\ \dots\ v_d], \quad \Lambda := \operatorname{diag}(\lambda_1, \dots, \lambda_d)$$

Allora:

$$\Sigma V = V\Lambda$$

Dimostrazione 6.2

Versione vista a lezione Verifichiamo che $\Sigma V = V\Lambda$ calcolando il generico elemento (i, j) di entrambi i membri.
A sinistra:

$$[\Sigma V]_{ij} = \sum_k \Sigma_{ik} V_{kj} = \sum_k \Sigma_{ik} (v_j)_k = [\Sigma v_j]_i$$

Poiché v_j è autovettore di Σ, abbiamo:

$$\Sigma v_j = \lambda_j v_j \quad \Rightarrow \quad [\Sigma v_j]_i = \lambda_j (v_j)_i = \lambda_j V_{ij}$$

A destra:

$$[V\Lambda]_{ij} = \sum_k V_{ik}\Lambda_{kj} = V_{ij}\lambda_j = \lambda_j V_{ij}$$

Poiché i due membri coincidono elemento per elemento:

$$[V\Lambda]_{ij} = \sum_k V_{ik}\Lambda_{kj} \overset{(\Lambda_{ij}=0 \forall k \neq j)}{=} V_{ij}\lambda_j = \lambda_j V_{ij}$$

Alternativa Abbiamo già verificato che:

$$\Sigma = V\Lambda V^\mathsf{T}, \quad V^\mathsf{T}V = VV^\mathsf{T} = I$$

Moltiplicando ambo i membri della prima equazione per V otteniamo:

$$\begin{aligned} \Sigma &= V\Lambda V^\mathsf{T} \\ \Sigma V &= V\Lambda(V^\mathsf{T}V) \\ \Sigma V &= V\Lambda \end{aligned}$$

□

6.5 PCA e decorrelazione

Proposizione 6.3: Decorrelazione delle componenti tramite PCA

Sia $X \in \mathbb{R}^{d \times n}$ un dataset centrato con matrice di covarianza $\Sigma = \mathrm{Cov}(X)$. Sia $V \in \mathbb{R}^{d \times d}$ una matrice ortogonale composta dagli autovettori di Σ, e sia:

$$Y = V^\mathsf{T} X$$

la trasformazione PCA. Allora:

$$\mathrm{Cov}(Y) = \Lambda$$

dove Λ è la matrice diagonale degli autovalori di Σ.

Dimostrazione 6.3

Poiché $Y = V^\mathsf{T} X$, la matrice di covarianza di Y è:

$$\mathrm{Cov}(Y) = \mathrm{Cov}(V^\mathsf{T} X)$$

Usando la Proposizione 5.2 sulla variazione della covarianza in una trasformazione lineare otteniamo:

$$\mathrm{Cov}(Y) = V^\mathsf{T} \, \Sigma \, (V^\mathsf{T})^\mathsf{T} = V^\mathsf{T} \Sigma V$$

Poiché $\Sigma = V \Lambda V^\mathsf{T}$, allora:

$$\mathrm{Cov}(Y) = V^\mathsf{T}(V \Lambda V^\mathsf{T})V = (V^\mathsf{T} V)\Lambda(V^\mathsf{T} V) = I \Lambda I = \Lambda$$

Quindi $\mathrm{Cov}(Y)$ è diagonale e coincide con la matrice degli autovalori di Σ, cioè le varianze delle componenti principali.

□

6.6 Scelte pratiche nella PCA: standardizzazione o no?

Due approcci standard alla PCA. Ci sono due modalità comuni per eseguire la PCA:

1. Centrare i dati rispetto alla media, poi eseguire la PCA sulla matrice di covarianza $\Sigma = \mathrm{Cov}(X)$ e infine standardizzare.
2. Centrare i dati rispetto alla media, standardizzare ogni variabile (cioè trasformarla in una variabile con media 0 e varianza 1), poi fare la PCA sulla matrice di correlazione ed eventualmente standardizzare di nuovo subito dopo.

Questi due approcci corrispondono a:

- PCA su Σ: privilegia le direzioni di massima varianza assoluta;
- PCA su matrice di correlazione: privilegia le direzioni di massima varianza relativa, indipendente dall'unità di misura.

Nota 6.2: Approccio standardizzato

Applicare la PCA dopo aver standardizzato equivale a diagonalizzare la matrice di correlazione $\rho(X)$, ovvero $\mathrm{Cov}(Z)$ dove Z è la versione standardizzata di X.

Unità di misura e varianza. La matrice di covarianza $\Sigma = \mathrm{Cov}(X)$ è influenzata dall'unità di misura delle variabili: se una variabile ha un'unità molto più grande, avrà anche una varianza più grande, e quindi tenderà a dominare le componenti principali.

Nota 6.3: Influenza delle unità di misura

Se le variabili hanno unità di misura molto diverse (es. altezza in cm, peso in kg), conviene standardizzare prima della PCA. Altrimenti la componente principale potrebbe riflettere solo la scala di una variabile.

Rappresentazione grafica dei due approcci. Nella seconda riga viene applicata la standardizzazione prima della PCA: il risultato finale (dopo PCA) è visivamente diverso. In entrambi i casi si ottiene una matrice di covarianza diagonale, ma le componenti principali sono diverse.

Nota 6.4: Quando usare la standardizzazione

Se non si ha una chiara ragione per dare peso a una variabile più che ad un'altra (es. tutte hanno importanza comparabile), allora è consigliabile usare l'approccio con standardizzazione.

7 Analisi Fattoriale (Factor Analysis)

L'Analisi Fattoriale è una tecnica statistica utilizzata per ridurre il numero di variabili osservate in un numero inferiore di variabili latenti chiamate **fattori**. Mentre la PCA riduce la dimensionalità conservando la varianza, la Factor Analysis cerca di spiegare le correlazioni tra le variabili attraverso fattori latenti, e può essere vista come un'estensione della PCA.

Definizione 7.1: Fattori e variabili osservate

Nel modello di Analisi Fattoriale, le variabili osservate $X_1, X_2, \ldots, X_m$ sono espresse come una combinazione lineare di fattori latenti $F_1, F_2, \ldots, F_k$ più un errore $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$:

$$X_i = \lambda_{i1}F_1 + \lambda_{i2}F_2 + \cdots + \lambda_{ik}F_k + \epsilon_i$$

dove λ_{ij} è il **factor loading** che esprime la relazione tra la variabile osservata X_i e il fattore F_j.

7.1 Factor Loadings

I **factor loadings** λ_{ij} rappresentano il peso di ciascun fattore latente F_j sulla variabile osservata X_i. Questi valori mostrano quanto ciascun fattore contribuisce alla varianza della variabile osservata. Un alto factor loading indica che la variabile è fortemente correlata con il fattore.

Esempio 7.1: Factor loadings

Nel caso di due fattori latenti F_1 e F_2, i fattori di carico potrebbero essere:

$$F_1 = 1.1X_1 + 0.8X_2, \quad F_2 = 1.1X_1 + 0.8X_2$$

Ciò significa che X_1 e X_2 sono fortemente influenzati da entrambi i fattori, con pesi $\lambda_{11} = 1.1$, $\lambda_{12} = 0.8$, e così via.

7.2 Obiettivo dell'analisi fattoriale

L'obiettivo dell'Analisi Fattoriale è quello di ridurre il numero di variabili osservate $X_1, X_2, \ldots, X_m$ in k fattori $F_1, F_2, \ldots, F_k$, dove $k < m$, cercando di mantenere la maggior parte della varianza. L'analisi si concentra nel trovare i fattori latenti che meglio spiegano le correlazioni tra le variabili.

Nota 7.1: Riduzione dimensionale nella Factor Analysis

La riduzione di dimensione in Factor Analysis non è come nella PCA, dove si cerca di massimizzare la varianza, ma si cerca di spiegare le correlazioni tra le variabili attraverso un numero ridotto di fattori.

7.3 Quando usare l'Analisi Fattoriale?

L'Analisi Fattoriale è utile quando:

- Le variabili originali sono fortemente correlate tra loro;
- Si vuole ridurre la dimensionalità dei dati senza perdere troppe informazioni;
- Le variabili sono influenzate da un numero ridotto di fattori latenti.

Nota 7.2: Fattori "schiacciati"

Quando il numero di variabili osservate m è grande e ci sono molte componenti di Y con varianza piccola, l'Analisi Fattoriale è spesso più adatta rispetto alla PCA, che potrebbe perdere troppe informazioni in presenza di molte variabili "schiacciate" (ovvero con bassa varianza).

7.4 Relazione con la PCA

L'Analisi Fattoriale può essere vista come una generalizzazione della PCA. Mentre la PCA si concentra nel massimizzare la varianza, la Factor Analysis cerca di spiegare la varianza condivisa tra le variabili attraverso fattori latenti. Quindi, la PCA può essere considerata come un caso particolare di Factor Analysis, dove tutti i fattori sono assunti ortogonali e indipendenti.

7.5 Riduzione dimensionale e scelta del numero di componenti

Quando m (il numero di variabili) è grande, esistono molte componenti principali con varianza piccola. La PCA cerca di ridurre la dimensione del vettore X mantenendo la varianza totale e riducendo la complessità del modello. La somma delle varianze prima e dopo la trasformazione è costante e conserva la varianza totale:

$$\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_m) = \mathrm{Var}(Y_1) + \cdots + \mathrm{Var}(Y_m) = \text{varianza totale}$$

Questa relazione implica che la traccia della matrice di covarianza Σ è uguale alla somma degli autovalori di Σ, ovvero:

$$\mathrm{tr}(\Sigma) = \mathrm{tr}(\Lambda)$$

dove Λ è la matrice diagonale degli autovalori.

7.5.1 Distribuzione degli autovalori

Tipicamente, gli autovalori $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ sono ordinati in modo decrescente. Questo ordine riflette la quantità di varianza spiegata da ciascuna componente principale. I componenti principali con autovalori maggiori spiegano una porzione maggiore della varianza totale.

Un approccio comune è quello di utilizzare un grafico dei cosiddetti **cambiamenti di pendenza** (come lo scree plot) per determinare il numero di componenti principali da mantenere. Un cambiamento significativo nella pendenza suggerisce il numero ottimale di componenti da considerare: Nel grafico sopra, le prime 3 componenti sembrano spiegare la maggior parte della varianza.

7.6 Quando standardizzare i dati

Quando i dati hanno unità di misura diverse, è importante standardizzarli prima di applicare la PCA. Questo è cruciale, poiché le variabili con unità più grandi tenderanno a dominare la varianza, influenzando fortemente le componenti principali.

Se dopo la PCA i dati non vengono standardizzati, si ottiene una rotazione dei dati senza alcuna scalatura. In questo caso, la PCA non fornirà una riduzione della dimensione che tiene conto della varianza relativa di ciascuna variabile, ma solo una rotazione rispetto alla distribuzione dei dati. Di conseguenza, la distanza tra i punti sarà influenzata solo dalla loro distribuzione e non dalla varianza delle variabili.

Nota 7.3: Standardizzazione dopo PCA

Standardizzare i dati prima della PCA è fondamentale per ridurre il rischio di amplificare il rumore nelle componenti con bassa varianza. La standardizzazione aiuta a bilanciare l'influenza di variabili con scale diverse.

7.7 Effetto della standardizzazione

Quando i dati vengono standardizzati dopo la PCA, la varianza di ciascuna componente principale è distribuita in modo più uniforme, evitando che variabili con bassa varianza distorcano i risultati. Il grafico sottostante mostra l'effetto della standardizzazione:

Nel grafico: - μ e Σ indicano i dati originali con la loro media e covarianza, - $0, \Sigma$ indica i dati centrati ma non standardizzati, - $0, I$ rappresenta i dati dopo standardizzazione.

Nota 7.4: Importanza della standardizzazione

La standardizzazione dopo la PCA è essenziale quando le variabili hanno scale diverse, poiché permette una comparazione equa tra le variabili e riduce l'influenza di quelle con una varianza maggiore.

8 Massima verosimiglianza

8.1 Stimatori

Definizione 8.1: Campione casuale

Un *campione casuale* è una collezione di variabili aleatorie i.i.d.:

$$X_1, X_2, \ldots, X_n$$

Le realizzazioni osservate sono i dati: $x_1, x_2, \ldots, x_n$.

Ad esempio, possiamo raccogliere le altezze o le pressioni sanguigne di un gruppo di individui (Esempio 8.1). Questi dati rappresentano realizzazioni di variabili aleatorie, che modellano fenomeni osservabili in un contesto sperimentale o reale.

Esempio 8.1: Pressione sanguigna

Raccolgo le seguenti misurazioni di pressione da un gruppo di studenti:

$$115, \quad 135, \quad 127, \quad 148, \quad 126$$

Posso ipotizzare che:

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \mu < 200 \text{ con alta probabilità}$$

Definizione 8.2: Statistica inferenziale

Lo scopo della statistica inferenziale è quello di trarre conclusioni sulla *legge delle* X_i a partire dai dati osservati $x_1, \ldots, x_n$.

Per realizzare questo obiettivo, utilizziamo funzioni dei dati osservati che riassumono l'informazione rilevante contenuta nel campione. Queste funzioni sono chiamate **statistiche**.

Definizione 8.3: Statistica

Una **statistica** è una variabile aleatoria che è funzione deterministica del campione, ovvero:

$$T = f(X_1, X_2, \ldots, X_n)$$

In particolare, quando una statistica viene usata per fornire una stima di un parametro incognito del modello (ad esempio, la media o la varianza della popolazione), essa prende il nome di **stimatore**.

Informalmente, uno **stimatore** di un parametro θ è una statistica $\hat{\theta} = f(X_1, \ldots, X_n)$ la cui legge è in qualche modo concentrata attorno a θ.

Definizione 8.4: Stimatore

Uno **stimatore** è una funzione dei dati che serve a stimare un parametro ignoto (es. la media μ) della distribuzione da cui proviene il campione.

Tra i più comuni stimatori, troviamo ad esempio:

- La **media campionaria**:
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- La **mediana campionaria**, ovvero $X_{(n/2)}$

Nel caso normale (gaussiano), media e mediana coincidono. In generale, no.

Definizione 8.5: Stimatore consistente

Uno **stimatore consistente** di un parametro θ è una famiglia di statistiche $\hat{\theta}_n$, con $n = 1, 2, \ldots$, tale che:
$$\hat{\theta}_n \xrightarrow[P]{n\to\infty} \theta \quad \text{(convergenza in probabilità)}$$
Cioè:
$$\forall \varepsilon > 0, \quad P(|\hat{\theta}_n - \theta| > \varepsilon) \xrightarrow{n\to\infty} 0$$
Operativamente, è sufficiente che:
$$\mathbb{E}[\hat{\theta}_n] \xrightarrow{n\to\infty} \theta \quad \text{e} \quad \mathrm{Var}(\hat{\theta}_n) \xrightarrow{n\to\infty} 0$$

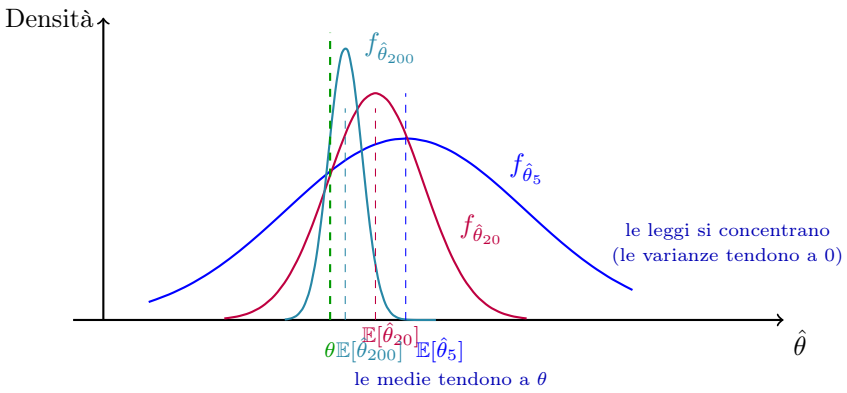

Figura 2: Illustrazione della consistenza: la distribuzione di $\hat{\theta}_n$ si concentra attorno a θ all'aumentare di n.

Definizione 8.6: Stimatore corretto

Uno **stimatore corretto** di un parametro θ è una statistica $\hat{\theta}$ per cui:

$$\mathbb{E}[\hat{\theta}] = \theta$$

In caso contrario, lo stimatore è detto **distorto**, e la differenza

$$\mathbb{E}[\hat{\theta}] - \theta$$

è chiamata **bias** (errore sistematico).

Nota 8.1: Media campionaria

Ricordiamo che:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad \mathbb{E}[\bar{X}] = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

La media campionaria $\bar{X}$ è **consistente e corretta**.

Nota 8.2: Varianza campionaria

La **varianza campionaria** è definita come:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

La **deviazione standard campionaria** è:

$$S_X = \sqrt{S_X^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Entrambe le statistiche S_X^2 e S_X sono **consistenti** per σ^2 e σ, rispettivamente. Tuttavia:

$$\mathbb{E}[S_X^2] = \sigma^2 \quad \text{(corretto)} \qquad \text{ma} \qquad \mathbb{E}[S_X] < \sigma \quad \text{(distorto)}$$

In particolare, se $X_i \sim \mathcal{N}(\mu, \sigma^2)$ si dimostra che:

$$\mathrm{Var}(S_X^2) = \frac{2\sigma^4}{n-1} \quad \Rightarrow \quad S_X^2 \text{ è consistente}$$

mentre S_X è distorto perché:

$$\mathbb{E}[{S_X}^2] = \sigma^2, \quad 0 < \mathrm{Var}(S_X) = \mathbb{E}[{S_X}^2] - \mathbb{E}[S_X]^2 \quad \Longrightarrow \quad \mathbb{E}[S_X]^2 < \sqrt{\mathbb{E}[{S_X}^2]} = \sigma^2$$
$$\Longrightarrow \quad \mathbb{E}[S_X] < \sigma$$

Uno degli approcci più diffusi per stimare i parametri incogniti di un modello statistico è il metodo della **massima verosimiglianza**. L'idea di base è semplice: tra tutti i possibili valori del parametro, scegliamo quello che rende i dati osservati più "probabili".

8.2 Definizione

Se il modello prevede una densità (o massa) di probabilità parametrica $f(x;\theta)$, e abbiamo a disposizione un campione di osservazioni $x_1, \dots, x_n$, allora consideriamo la funzione di verosimiglianza come una funzione del parametro θ, e ne cerchiamo il massimo.

Definizione 8.7: Massima verosimiglianza

Dato un campione i.i.d. $X_1, \dots, X_n \sim f(x;\theta)$, la **funzione di verosimiglianza** è definita come:

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

Lo **stimatore di massima verosimiglianza** (MLE) è il valore del parametro che massimizza la funzione di verosimiglianza:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta; x_1, \dots, x_n)$$

Nota 8.3: Il parametro θ

Normalmente, il parametro θ è un vettore $\theta = (\theta_1, \theta_2, \dots, \theta_k)$.

In pratica, invece della funzione L, si lavora spesso con la *log-verosimiglianza*:

$$\ell(\theta) = \log L(\theta; x_1, \dots, x_n) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

Poiché il logaritmo è strettamente crescente, il valore che massimizza $\ell(\theta)$ è lo stesso che massimizza $L(\theta)$, ma l'analisi è spesso più semplice con somme piuttosto che prodotti.

Esempio 8.2: MLE per la distribuzione Gamma

Supponiamo che $X_1, \dots, X_n$ siano variabili aleatorie indipendenti e distribuite secondo la legge Gamma(α, β), con densità:

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0$$

1. Funzione di verosimiglianza. Per un campione osservato $x_1, \dots, x_n$, la funzione di verosimiglianza è:

$$L(\alpha,\beta; x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i} = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\beta x_i}$$

2. Log-verosimiglianza. Per semplificare il calcolo si prende il logaritmo:

$$\ell(\alpha, \beta) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i$$

3. Massimizzazione. Deriviamo rispetto a β e poniamo la derivata pari a zero:

$$\frac{\partial \ell}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \hat{\beta} = \frac{n\alpha}{\sum x_i} = \frac{\alpha}{\bar{x}}$$

Dove $\bar{x} = \frac{1}{n} \sum x_i$ è la media campionaria.
4. Derivata rispetto ad α. Risulta più complicata, e coinvolge la funzione digamma $\psi(\alpha) = \frac{d}{d\alpha} \log \Gamma(\alpha)$:

$$\frac{\partial \ell}{\partial \alpha} = n \log \beta - n\psi(\alpha) + \sum \log x_i$$

Questa equazione si risolve numericamente (es. metodo di Newton-Raphson) per ottenere $\hat{\alpha}$.

Conclusione. Gli stimatori MLE per Gamma(α, β) sono dati da:

$$\hat{\beta} = \frac{\hat{\alpha}}{\bar{x}}, \quad \hat{\alpha} \text{ ottenuto risolvendo: } \log \hat{\beta} - \psi(\hat{\alpha}) + \frac{1}{n} \sum \log x_i = 0$$

8.3 Esempi notevoli

8.3.1 Distribuzione esponenziale

Consideriamo un campione casuale $X_1, \ldots, X_n \sim \text{expo}(\lambda)$, con densità:

$$f(x; \lambda) = \text{expo}(\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0$$

Esempio 8.3: MLE per la legge esponenziale

La log-verosimiglianza associata al campione è:

$$\ell(\lambda) = \sum_{i=1}^{n} \log f(x_i; \lambda) = \sum_{i=1}^{n} (\log \lambda - \lambda x_i) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

La funzione è concava, poiché la derivata seconda è negativa:

$$\ell''(\lambda) = -\frac{n}{\lambda^2} < 0$$

Derivando la log-verosimiglianza e ponendo uguale a zero otteniamo:

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{n}{\sum x_i} = (\bar{x})^{-1}$$

Conclusione: lo stimatore di massima verosimiglianza per λ è:

$$T = \left(\bar{X}\right)^{-1}$$

Questo esempio mostra che, nel caso esponenziale, la media campionaria $\bar{X}$ è uno stimatore corretto della media teorica $\frac{1}{\lambda}$. Poiché $\bar{X}$ è corretta per la media, il suo inverso T è uno stimatore corretto per λ — e coincide con l'MLE.

8.3.2 Legge uniforme

Consideriamo un campione casuale $X_1, \dots, X_n \sim \text{unif}(0, a)$, con densità:

$$f(x; a) = \text{unif}(0, a) = \frac{1}{a}$$

Esempio 8.4: MLE per la legge uniforme

La funzione di verosimiglianza è:

$$L(a) = \prod_{i=1}^{n} f(x_i; a) = \begin{cases} \frac{1}{a^n} & \text{se } a \geq \max x_i \\ 0 & \text{altrimenti} \end{cases}$$

La log-verosimiglianza è:

$$\ell(a) = \log L(a) = \begin{cases} -n \log a & \text{se } a \geq \max x_i \\ -\infty & \text{altrimenti} \end{cases}$$

Poiché $\ell(a)$ decresce in a, per massimizzarla bisogna scegliere il valore minimo ammissibile:

$$\hat{a}_{\text{MLE}} = \max x_i$$

Conclusione: lo stimatore MLE per a è:

$$Y := \max_{i=1,\dots,n} X_i$$

Questo stimatore è *consistente* ma *distorto*. Infatti:

$$\mathbb{E}(Y) = \frac{n}{n+1} a \quad \Rightarrow \quad Y \text{ sottostima sistematicamente } a$$

Correzione del bias:

$$Y_{\text{adj}} := \frac{n+1}{n} Y = \frac{n+1}{n} \max X_i$$

Questo nuovo stimatore Y_{adj} è **corretto** e anch'esso consistente.

Nota 8.4: Bias degli MLE

Lo stimatore di massima verosimiglianza è spesso *consistente* ma *non corretto*. In molti casi, è possibile applicare una correzione (bias correction) per ottenere uno stimatore corretto mantenendo la consistenza.

8.3.3 Distribuzione normale

Consideriamo il caso classico in cui $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, con entrambi i parametri μ e σ^2 ignoti.

Esempio 8.5: MLE per la legge normale

La densità della normale è:

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

La log-verosimiglianza per un campione è:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(x_i; \mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Calcolo degli MLE:

- Derivando rispetto a μ e ponendo uguale a zero:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \quad \Rightarrow \quad \hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Derivando rispetto a σ^2 e ponendo uguale a zero:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Conclusione: gli stimatori di massima verosimiglianza sono:

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \mathbb{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Osservazioni:

- $\bar{X}$ è **consistente e corretto** come stimatore di μ
- $\mathbb{S}^2$ è **consistente ma distorto** come stimatore di σ^2
- Lo stimatore corretto per la varianza è:

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

che viene preferito nella pratica quando si desidera correggere il bias.

8.3.4 Distribuzione di Bernoulli

Nel caso più semplice di variabile binaria, supponiamo che $X_1, \dots, X_n \sim \text{Bern}(p)$, dove $p \in (0,1)$ è la probabilità di successo.

Esempio 8.6: MLE per la legge di Bernoulli

La funzione di massa di probabilità della variabile Bernoulli è:

$$f(x;p) = p^x(1-p)^{1-x}, \quad x \in \{0,1\}$$

Data un'osservazione $x_1, \dots, x_n$, la funzione di verosimiglianza è:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{O_1}(1-p)^{O_0}$$

dove:

$$O_1 = \sum_i x_i \qquad \text{(numero di successi)}$$

$$O_0 = n - O_1 \qquad \text{(numero di insuccessi)}$$

La log-verosimiglianza è:

$$\ell(p) = O_1 \log p + O_0 \log(1-p)$$

Derivando e ponendo uguale a zero otteniamo:

$$\ell'(p) = \frac{O_1}{p} - \frac{O_0}{1-p} = 0 \quad \Rightarrow \quad \hat{p} = \frac{O_1}{n} = \bar{x}$$

Conclusione: lo stimatore MLE di p è la media campionaria:

$$\hat{p}_{\text{MLE}} = \bar{X}$$

Questo stimatore è sia **corretto** che **consistente**.

8.3.5 Distribuzione multinomiale

Supponiamo di osservare n campioni indipendenti da una distribuzione **multinomiale categoriale** su m categorie, con probabilità:

$$\boldsymbol{p} = (p_1, p_2, \dots, p_m), \quad \text{dove } \sum_{j=1}^{m} p_j = 1$$

Ogni osservazione è un vettore *one-hot* $b(i) = (b_1(i), \dots, b_m(i))$, dove esattamente una componente è 1 (e tutte le altre 0).

Esempio 8.7: MLE per la legge multinomiale

La funzione di massa di probabilità calcola la probabilità che l'osservazione i cada nella categoria k; in altre parole, la probabilità di osservare:

$$P\left(b(i) = (0, 0, \ldots, \overset{\text{pos. } k}{1}, \ldots, 0)\right) = p_1^{x_1(i)} p_2^{x_2(i)} \cdots p_m^{x_m(i)}$$

ossia:

$$f(x; \boldsymbol{p}) = \prod_{j=1}^{m} p_j^{b_j(i)}$$

A questo punto, definiamo:

$$O_j := \sum_{i=1}^{n} b_j(i) \tag{1}$$

il numero di volte in cui è stata osservata la categoria j.
La funzione di verosimiglianza è:

$$L(p_1, \ldots, p_m) = \prod_{i=1}^{n} \prod_{j=1}^{m} p_j^{b_j(i)}$$

Espandiamo il logaritmo della verosimiglianza:

$$\begin{aligned}
\ell(p_1, \ldots, p_m) &= \log L(p_1, \ldots, p_m) \\
&= \log \left(\prod_{i=1}^{n} \prod_{j=1}^{m} p_j^{b_j(i)} \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \log \left(p_j^{b_j(i)} \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} b_j(i) \log p_j
\end{aligned}$$

Invertendo l'ordine delle somme otteniamo:

$$\begin{aligned}
\ell(p_1, \ldots, p_m) &= \sum_{j=1}^{m} \left(\sum_{i=1}^{n} b_j(i) \right) \log p_j \\
&= \sum_{j=1}^{m} O_j \log p_j \qquad \text{[per l'Equazione (1)]}
\end{aligned}$$

Questa funzione va massimizzata sotto il vincolo:

$$\sum_{j=1}^{m} p_j = 1$$

Soluzione: introducendo un moltiplicatore di Lagrange per il vincolo, si ottiene:

$$\hat{p}_j = \frac{O_j}{n}$$

Conclusione: lo stimatore MLE per ciascuna probabilità p_j è:

$$\hat{p}_j = \pi_j := \frac{O_j}{n}$$

Questo corrisponde alla frequenza relativa con cui la categoria j è stata osservata.

8.3.6 Applicazioni al Machine Learning

Nel contesto del **machine learning**, abbiamo una situazione più complessa: ogni osservazione è composta da un **input** x e da un **output** Y, e assumiamo che la distribuzione di Y dipenda da x.

Esempio motivante. Supponiamo che l'altezza Y di una persona dipenda dal genere $x \in \{0, 1\}$. Possiamo modellare questa dipendenza come segue:

$$\begin{cases} x = 0 \Rightarrow Y \sim \mathcal{N}(175, 7^2) \\ x = 1 \Rightarrow Y \sim \mathcal{N}(168, 7^2) \end{cases}$$

In generale, assumiamo che la distribuzione condizionata $Y \mid x$ appartenga a una famiglia parametrica, ad esempio una distribuzione normale o multinomiale.

Distribuzione Multinomiale Condizionata. Supponiamo ora che l'output Y sia discreto su m classi, e che $Y \mid x \sim \text{Multinomial}(1; p_1(x), \ldots, p_m(x))$, dove le probabilità $p_j(x)$ dipendono dall'input x. Indichiamo con $\vec{\alpha}$ il vettore di parametri del modello (es. pesi di una rete neurale), e assumiamo che:

$$p_j = p_j(x; \vec{\alpha})$$

La funzione di log-verosimiglianza per n osservazioni diventa:

$$\ell(\vec{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{m} b_j(i) \log p_j(x_i; \vec{\alpha})$$

dove $b_j(i)$ è la codifica one-hot della classe osservata per l'osservazione i.

Equivalentemente, indicando con $y(i)$ la classe osservata per x_i, otteniamo la forma più compatta:

$$\ell(\vec{\alpha}) = \sum_{i=1}^{n} \log p_{y(i)}(x_i; \vec{\alpha})$$

in quanto soltanto la classe $j = y(i)$ contribuisce alla somma lungo l'asse j.

Osservazioni pratiche.

- Le funzioni $p_j(x;\vec{\alpha})$ sono spesso reti neurali o modelli statistici complessi;
- Il vettore $\vec{\alpha}$ contiene tutti i parametri del modello, che possono anche essere milioni;
- La forma della funzione di verosimiglianza rimane invariata, ma la sua ottimizzazione richiede metodi numerici;
- In pratica, si utilizza un **ottimizzatore iterativo** (come la discesa del gradiente) per massimizzare $\ell(\vec{\alpha})$.

Nota 8.5: Likelihood e deep learning

Nel deep learning, la stima di massima verosimiglianza corrisponde alla minimizzazione della *cross-entropy loss* tra le etichette osservate e le probabilità predette dal modello $p_j(x;\vec{\alpha})$.

Tabella 1: Riepilogo degli stimatori di massima verosimiglianza (MLE)

Distribuzione	**Parametri**	**MLE**
Bernoulli	p	$\hat{p} = \bar{X}$
Binomiale $\text{Bin}(n,p)$	p	$\hat{p} = \frac{k}{n}$ (con $k = \sum X_i$)
Esponenziale	λ	$\hat{\lambda} = \frac{1}{\bar{X}}$
Normale	μ, σ^2	$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$
Uniforme $\text{Unif}(0,a)$	a	$\hat{a} = \max X_i$
Gamma $\text{Gamma}(\alpha,\beta)$	α, β	Nessuna formula chiusa;
		$\hat{\beta} = \frac{\hat{\alpha}}{\bar{X}}, \quad \hat{\alpha}$ risolta numericamente
Multinomiale	$\boldsymbol{p} = (p_1, \ldots, p_m)$	$\hat{p}_j = \frac{O_j}{n}$ per ogni j

8.4 Legame con la Cross-Entropy Loss

In problemi di classificazione, la funzione di log-verosimiglianza coincide (a meno di un segno) con la **cross-entropy loss**. Questo vale sia nel caso classico della distribuzione multinomiale semplice, sia nel caso condizionato tipico del machine learning.

Caso 1: Distribuzione multinomiale classica

Siano $Y_1, \ldots, Y_n$ variabili i.i.d. che seguono una distribuzione categoriale con vettore di probabilità $p = (p_1, \ldots, p_m)$. Definiamo:

$$q_j := \frac{o_j}{n}$$

la frazione di dati che hanno categoria j. Allora la **cross-entropy loss** è:

$$\begin{aligned} H(q,p) &= -\sum_{j=1}^{m} q_j \log p_j \\ &= -\frac{1}{n}\sum_{j=1}^{m} o_j \log p_j \\ &= -\frac{1}{n}\ell(p_1,\dots,p_m) \end{aligned}$$

Interpretazione. Minimizzare la cross-entropy equivale a massimizzare la verosimiglianza. La stima MLE dei p_j è quindi la stessa che minimizza $H(q,p)$.

Caso 2: Distribuzione multinomiale condizionata

Nel machine learning, le probabilità p_j dipendono da un input x_i e da un vettore di parametri $\vec{\alpha}$. Indichiamo:

$$p_j(i) := p_j(x_i;\vec{\alpha})$$

dove $p_j(i)$ è la probabilità che il modello assegna alla classe j dato l'input x_i. Ogni etichetta $y(i)$ è rappresentata da un vettore one-hot $b(i) \in \{0,1\}^m$.

La log-verosimiglianza per n osservazioni è:

$$\ell(\vec{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{m} b_j(i)\log p_j(x_i;\vec{\alpha})$$

Nota 8.6: Differenza rispetto al caso non condizionato

La differenza principale rispetto al caso non condizionato è che ora le probabilità p_j dipendono da un input x_i e da un vettore di parametri $\vec{\alpha}$ e non è possibile semplificare le due sommatorie per ottenere i vari o_j.

La cross-entropy loss media è quindi:

$$\begin{aligned} -\frac{1}{n}\ell(\vec{\alpha}) &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} b_j(i)\log p_j(x_i;\vec{\alpha}) && [\text{def. di } \ell(\vec{a})] \\ &= \frac{1}{n}\sum_{i=1}^{n} H\big(b(i),\ p(x_i;\vec{\alpha})\big) && \left[\sum_j b_j(i) = 1\right] \end{aligned}$$

Nota 8.7: Cross-entropy nel machine learning

In classificazione supervisionata, la cross-entropy rappresenta la media delle entropie incrociate tra le etichette (codificate in one-hot) e le distribuzioni predette dal modello. Minimizzarla equivale a migliorare la probabilità assegnata alle classi corrette.

8.4.1 Logits e Cross-Entropia

Consideriamo la funzione logit definita come segue:

$$p_j(x; \alpha, \beta_j) \to \mathbb{R}, \quad x \in \mathbb{R}^d$$

dove $\alpha \in \mathbb{R}^m$ e $\beta_j \in \mathbb{R}^m$ sono parametri del modello. La funzione logit è data da:

$$\text{logits} \to Y = b + Wx \in \mathbb{R}^m$$

dove $W \in \mathbb{R}^{m \times d}$ è la matrice dei pesi, e $b \in \mathbb{R}^m$ è il bias. L'output Y deve essere trasformato in probabilità tramite la funzione softmax:

$$\hat{p}_j = \frac{e^{Y_j}}{\sum_{i=1}^{m} e^{Y_i}}$$

dove la softmax è applicata elemento per elemento per ottenere le probabilità previste per ogni classe.

A questo punto, possiamo definire la cross-entropia come segue:

$$H(q, p) = -\sum_{j=1}^{m} q_j \log(\hat{p}_j)$$

dove q_j è la distribuzione target (ad esempio, un vettore one-hot), e $\hat{p}_j$ è la probabilità predetta dalla softmax. In alternativa, possiamo scrivere la cross-entropia in termini di logits:

$$H(q, p) = -\sum_{j=1}^{m} q_j \log\left(\frac{e^{Y_j}}{\sum_{i=1}^{m} e^{Y_i}}\right)$$

Questa espressione lega direttamente la funzione di perdita con i logits calcolati. La cross-entropia è utilizzata per ottimizzare il modello, rendendo le probabilità previste il più possibile simili alla distribuzione di probabilità target.

8.5 Mean Squared Error Loss (MSE)

La **Mean Squared Error Loss (MSE)** è una funzione di perdita ampiamente utilizzata per i modelli di regressione, particolarmente quando i dati sono distribuiti secondo una **distribuzione gaussiana**. Essa misura la media dei quadrati delle differenze tra i valori osservati e quelli predetti dal modello.

Definizione 8.8: MSE

La **MSE** per n osservazioni è definita come:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

dove y_i è il valore osservato e $\hat{y}_i$ è il valore predetto dal modello.

Distribuzione dei dati

Nel contesto dei dati Gaussiani, supponiamo che $Y(i)$ sia distribuito come $Y(i) \sim \mathcal{N}(\mu(i), \sigma^2(i))$, con $\mu(i)$ e $\sigma(i)$ dipendenti dal parametro $x(i)$.

La funzione di verosimiglianza per n osservazioni è quindi:

$$\ell(\mu(i), \sigma(i)) = \sum_{i=1}^{n}\left[-\log(\sqrt{2\pi}) - \log(\sigma(i)) - \frac{(y(i)-\mu(i))^2}{2\sigma(i)^2}\right]$$

Dove:

- $\mu(i)$ è la media prevista,
- $\sigma(i)$ è la deviazione standard prevista per il dato i.

8.5.1 I casi

- **Caso 1: Nessuna dipendenza dall'input (Esempio 8.5)**: Se non c'è dipendenza da $x(i)$, la funzione di log-verosimiglianza diventa:

  $$\ell(\mu, \sigma) = C - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y(i)-\mu)^2$$

  Risolvendo, otteniamo lo stimatore per la media e la deviazione standard:

  $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y(i), \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y(i)-\hat{\mu})^2$$

- **Caso 2: Modello regressivo con varianza costante (*omoschedastico*)**: Se la deviazione standard $\sigma(i)$ è costante, e i parametri $\mu(i)$ dipendono da $x(i)$, possiamo assumere una relazione parametrica. Ad esempio, un modello lineare potrebbe essere usato per descrivere:

  $$\sigma(i) = \sigma, \quad \mu(i) = \nu(x(i); \vec{\alpha})$$

La funzione di log-verosimiglianza diventa:

$$\ell(\sigma, \vec{\alpha}) = C - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y(i) - \nu(x(i); \vec{\alpha}))^2$$

Introduciamo la funzione **SE** (*Squared Error*) che calcola la differenza al quadrato degli argomenti:

$$\mathrm{SE}(y, z) = (y - z)^2$$

Questo permette di riscrivere la log-verosimiglianza in funzione della MSE (Definizione 8.8):

$$\mathrm{loss}_{\mathrm{MSE}}(\vec{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{SE}(y(i), \nu(x(i); \vec{\alpha})) = \frac{1}{n} \sum_{i=1}^{n} (y(i) - \nu(x(i); \vec{\alpha}))^2$$

$$\Rightarrow \ell(\sigma, \vec{\alpha}) = C - n \log \sigma - \frac{1}{2\sigma^2} n \, \mathrm{loss}_{\mathrm{MSE}}(\vec{\alpha}) = C - n(\log \sigma + \frac{1}{2\sigma^2} \mathrm{loss}_{\mathrm{MSE}}(\vec{\alpha}))$$

Da qui si ottiene lo stimatore $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\mathrm{loss}_{\mathrm{MSE}}(\vec{\alpha})}$$

L'ottimizzazione degli altri parametri segue dalla minimizzazione della $\mathrm{loss}_{\mathrm{MSE}}$.

Nota 8.8: $\mathrm{loss}_{\mathrm{MSE}}$

La funzione $\mathrm{loss}_{\mathrm{MSE}}$ è semplicemente definita come lo **scarto quadratico medio** tra i valori osservati e quelli predetti dal modello.

- **Caso 3: Dipendenza da** x: Se i parametri $\mu(i)$ e $\sigma(i)$ dipendono da $x(i)$, possiamo definire la funzione di regressione come segue:

$$\mu(i) = \nu(x(i); \vec{\alpha}), \quad \sigma(i) = \tau(x(i); \vec{\alpha})$$

Dove $\nu(x(i); \vec{\alpha})$ è il modello di regressione e $\tau(x(i); \vec{\alpha})$ è la variabilità associata.

Nota 8.9: Stima di $\vec{\alpha}$

Questa funzione viene ottimizzata usando metodi iterativi, come la discesa del gradiente, per trovare i migliori parametri.

9 Regressione Lineare Semplice

La **regressione lineare semplice** è uno dei modelli più basilari in statistica e machine learning. In questo caso, si cerca di modellare la relazione tra una variabile dipendente Y e una variabile indipendente X.

9.1 Modello e parametri

Il modello della regressione lineare semplice assume che la variabile dipendente Y sia una combinazione lineare della variabile indipendente X, più un errore ϵ. In altre parole:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

dove:

- Y_i è la variabile dipendente (osservazione), che si assume essere normalmente distribuita:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2),$$

- X_i è la variabile indipendente, che viene trattata come una variabile deterministica (osservata),
- β_0 è l'intercetta del modello,
- β_1 è la pendenza della retta di regressione,
- ϵ_i è l'errore, che si assume essere normalmente distribuito con media zero e varianza costante σ^2:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Il modello descrive una **retta di regressione** che approssima la relazione tra X e Y.

Definizione 9.1: Parametri del modello

I parametri del modello di regressione lineare sono β_0 (intercetta), β_1 (pendenza), e σ^2 (varianza dell'errore).

Nota 9.1: Assunzioni del modello

Il modello di regressione lineare semplice assume che:

- La relazione tra X e Y sia lineare.
- Gli errori ϵ_i siano indipendenti e normalmente distribuiti con media zero e varianza costante (σ^2).
- La variabile dipendente Y segue una distribuzione normale con media $\mu = \beta_0 + \beta_1 X_i$ e varianza costante σ^2: $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$.
- La variabile indipendente X è deterministica, mentre la variabile dipendente Y è casuale.

Nota 9.2: Baricentro della regressione

Il **baricentro** della regressione è il punto medio dei dati osservati, dato dalla media di X e Y. In termini matematici, il baricentro $(\overline{x}, \overline{y})$ è dato da:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

9.2 Stima dei parametri tramite Maximum Likelihood (MLE)

I parametri β_0, β_1, e σ^2 possono essere stimati utilizzando il metodo della massima verosimiglianza (MLE). La funzione di log-verosimiglianza è:

$$\ell(\beta_0, \beta_1, \sigma) = C - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Per trovare i parametri che massimizzano la log-verosimiglianza, possiamo derivare rispetto ai parametri e risolvere il sistema di equazioni. I parametri stimati risultano:

Definizione 9.2: Stima della pendenza

Lo stimatore per la pendenza $\hat{\beta}_1$ nel modello di regressione lineare è dato dalla seguente espressione:

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$$

dove:

- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ è la media dei valori di x,
- $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ è la media dei valori di y,
- $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ è la media dei prodotti $x_i y_i$,
- $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ è la media dei quadrati dei valori di x.

Questa formula fornisce la stima ottimale della pendenza $\hat{\beta}_1$ nel caso di regressione lineare semplice.

Definizione 9.3: Stima dell'intercetta

La stima per l'intercetta $\hat{\beta}_0$ è data dalla formula:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$

Dove:

- $\overline{x}$ è la media dei valori di X,
- $\overline{y}$ è la media dei valori di Y,

- $\hat{\beta}_1$ è la pendenza della retta di regressione. ★

La retta di regressione $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ è costruita in modo che minimizzi la somma dei quadrati degli errori tra i valori osservati e quelli predetti. Un risultato interessante della regressione lineare è che questa retta passa sempre per il **baricentro** il dei dati, dato dalle medie $\overline{x}$ e $\overline{y}$.

Dimostrazione 9.1: Baricentro appartiene alla regressione lineare

Per $x = \overline{x}$, la retta di regressione assume il valore:

$$\hat{y}(\overline{x}) = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$$

Sostituendo $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ nella formula della retta otteniamo:

$$\hat{y}(\overline{x}) = \left(\overline{y} - \hat{\beta}_1 \overline{x}\right) + \hat{\beta}_1 \overline{x} = \overline{y}$$

□

Dimostrazione 9.2: Calcolo degli stimatori β_0 e β_1

Partiamo dalla funzione di perdita MSE, che nel caso della regressione lineare è definita come:

$$\text{loss}_{\text{MSE}}(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

dove $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ sono i valori predetti dal modello, y_i sono i valori osservati, e x_i sono i valori delle variabili indipendenti.
Espandendo questa espressione otteniamo:

$$\text{loss}_{\text{MSE}}(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Ora, per trovare i parametri $\hat{\beta}_0$ e $\hat{\beta}_1$, dobbiamo minimizzare questa funzione di perdita rispetto a $\hat{\beta}_0$ e $\hat{\beta}_1$. Questo può essere fatto derivando la funzione rispetto a ciascun parametro e ponendo le derivate uguali a zero.
Cominciamo derivando la funzione di perdita rispetto a $\hat{\beta}_0$:

$$\frac{\partial \text{loss}_{\text{MSE}}}{\partial \hat{\beta}_0} = \frac{2}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-1)$$

Poniamo questa derivata uguale a zero:

$$0 = \frac{2}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

Poiché $\sum_{i=1}^{n} x_i = n\overline{x}$ e $\sum_{i=1}^{n} y_i = n\overline{y}$, otteniamo:

$$\begin{aligned} n\overline{y} &= n\hat{\beta}_0 + \hat{\beta}_1 n\overline{x} \\ \Leftrightarrow \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \end{aligned}$$

Adesso deriviamo la funzione di perdita rispetto a $\hat{\beta}_1$:

$$\frac{\partial \text{loss}_{\text{MSE}}}{\partial \hat{\beta}_1} = \frac{2}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) (-x_i)$$

Poniamo anche questa derivata uguale a zero:

$$\begin{aligned} 0 &= \frac{2}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) (-x_i) \\ \Rightarrow \sum_{i=1}^{n} x_i y_i &= \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Sostituendo $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, otteniamo:

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i &= \left(\overline{y} - \hat{\beta}_1 \overline{x} \right) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i y_i &= \overline{y} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Poiché $\sum_{i=1}^{n} x_i = n\overline{x}$, otteniamo:

$$\sum_{i=1}^{n} x_i y_i = n\overline{y}\overline{x} - n\hat{\beta}_1 \overline{x}^2 + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Ora, raccogliamo i termini con $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{y}\overline{x}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

Questa è la formula per $\hat{\beta}_1$.
Ora possiamo sostituire $\hat{\beta}_1$ nella formula per $\hat{\beta}_0$:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

In questo modo, abbiamo calcolato gli stimatori $\hat{\beta}_0$ e $\hat{\beta}_1$ utilizzando la minimizzazione della funzione di perdita MSE.

□

9.3 Errore Standard del modello

L'errore standard S_e della stima del modello di regressione è definito come:

$$S_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

dove $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ è il valore predetto dalla retta di regressione.

10 Teorema di Cochran

10.1 Teorema per il ML

Teorema 10.1: Versione ML del Teorema di Cochran

Supponiamo di essere nel caso MSE loss omoschedastico:

$$Y(i) \sim \mathcal{N}(\mu(i), \sigma^2)$$

con $\sigma \equiv \sigma(i)$ costante. Supponiamo inoltre:

$$\mu(i) = \gamma_1 c_1(x(i)) + \cdots + \gamma_k c_k(x(i)) = \gamma \cdot c(x(i)), \quad \gamma, c \in \mathbb{R}^k$$

cioè una combinazione lineare dei valori $c_1(x(i)), \ldots, c_k(x(i))$ con i parametri $\gamma_1, \ldots, \gamma_k$.
Gli stimatori MLE dei parametri si trovano minimizzando la loss:

$$\text{loss}_{\text{MSE}}(\gamma_1, \ldots, \gamma_k) = \frac{1}{n} \sum_{i=1}^{n} (Y(i) - \gamma \cdot c(x(i)))^2$$

equivalente a minimizzare:

$$W(\gamma) = \|Y - C(x) \cdot \gamma\|^2 \qquad \text{(norma del vettore differenza)}$$

dove $C(x)$ è una matrice in $\mathbb{R}^{n \times k}$ le cui colonne sono date dai vettori:

$$\begin{bmatrix} c_j(x(i)) \\ \vdots \\ c_j(x(i)) \end{bmatrix}, \quad i = 1, \ldots, n,\ j = 1, \ldots, k$$

Al variare di γ, l'immagine $\gamma \cdot C(x)$ è un sottospazio vettoriale di dimensione k in $\mathbb{R}^n$. Lo stimatore $\hat{\gamma}$ è la proiezione ortogonale di Y su questo sottospazio.
Si definisce la somma dei quadrati dei residui come:

$$SSR = W(\hat{\gamma})$$

Allora:

1. $\hat{\gamma}$ è uno stimatore corretto di γ e la sua formula è:

$$\hat{\gamma} = (C^T C)^{-1} C^T Y$$

2. $\frac{SSR}{\sigma^2} \sim \chi^2(n-k)$, quindi $\frac{SSR}{n-k}$ è uno stimatore corretto di σ^2;
3. $\hat{\gamma}$ e SSR sono variabili aleatorie indipendenti.

Dimostrazione 10.1: Formula per la proiezione

Dato che il modello è lineare, esiste una formula esplicita per calcolare la proiezione ortogonale $\hat{\gamma}$ di Y sul sottospazio generato dalle colonne della matrice C. La formula è:

$$\hat{\gamma} = (C^T C)^{-1} C^T Y$$

Per verificarla, minimizziamo la funzione di costo $W(\gamma)$ calcolandone le derivate parziali rispetto a ciascun parametro γ_j e ponendole a zero.

$$W(\gamma) = \|Y - C\gamma\|^2 = \sum_{i=1}^{n} (Y_i - (C\gamma)_i)^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} C_{ij}\gamma_j \right)^2$$

Calcoliamo la derivata parziale rispetto a γ_h:

$$\frac{\partial W(\gamma)}{\partial \gamma_h} = \sum_{i=1}^{n} 2 \left(Y_i - \sum_{j=1}^{k} C_{ij}\gamma_j \right) (-C_{ih}) = 0$$

Questo implica:

$$\sum_{i=1}^{n} \sum_{j=1}^{k} C_{ij} C_{ih} \hat{\gamma}_j = \sum_{i=1}^{n} Y_i C_{ih} \quad \forall h = 1, \ldots, k$$

Riscrivendo l'equazione in forma matriciale:

$$(C^T C\hat{\gamma})_h = \sum_{i,j} C^T_{hi} C_{ij} \hat{\gamma}_j = \sum_{i} C^T_{hi} Y_i = (C^T Y)_h$$

Otteniamo quindi:

$$C^T C\hat{\gamma} = C^T Y \implies \hat{\gamma} = (C^T C)^{-1} C^T Y$$

□

10.2 Teorema per le applicazioni

Teorema 10.2: Versione applicativa del Teorema di Cochran

Siano $X_1, \ldots, X_n$ variabili aleatorie indipendenti con

$$X_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

(modello omoschedastico). Supponiamo inoltre che

$$\mu = (\mu_1, \ldots, \mu_n)^T \in V \subseteq \mathbb{R}^n$$

dove V è un sottospazio vettoriale di dimensione k assegnato.
Allora valgono i seguenti risultati:

1. Lo stimatore ML di μ è la proiezione ortogonale $\pi_V(X)$ di X su V;

2. $\pi_V(X)$ è uno stimatore corretto di μ;
3. La quantità
$$W := \|X - \pi_V(X)\|^2$$
è indipendente da $\pi_V(X)$ e
$$\frac{W}{\sigma^2} \sim \chi^2(n-k).$$

Esempio 10.1: Proiezione ortogonale in $\mathbb{R}^3$

Sia $Y \in \mathbb{R}^3$ un vettore aleatorio e sia $C\hat{\gamma}$ un sottospazio piano (di dimensione $k = 2$). Il punto $\hat{Y} = C\hat{\gamma}$ rappresenta la proiezione ortogonale di Y sul piano generato da C.
Definiamo la somma dei quadrati dei residui come:

$$SSR = \|Y - C\hat{\gamma}\|^2$$

Il teorema di Cochran assicura che:

- $\hat{\gamma}$ è uno stimatore corretto di γ;
- $\hat{\gamma}$ è indipendente da SSR;
- $\frac{SSR}{\sigma^2} \sim \chi^2(n-k)$.

Esempio 10.2: Campione Gaussiano e t-Student

Consideriamo un campione i.i.d. $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$. In questo caso, il modello di media è $\mu(i) = \mu = \gamma_1 \cdot 1$, quindi $k = 1$ e il regressore è $c_1(x(i)) = 1$ per ogni i. La matrice C è un vettore colonna di n uni:

$$C = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Calcoliamo le matrici necessarie per la proiezione:

$$C^T C = \begin{pmatrix} 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = n$$

$$C^T Y = \begin{pmatrix} 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \sum_{i=1}^{n} Y_i$$

Lo stimatore di massima verosimiglianza per μ è quindi la media campionaria:

$$\hat{\mu} = \hat{\gamma}_1 = (C^T C)^{-1} C^T Y = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}$$

La somma dei quadrati dei residui (SSR) è:

$$SSR = \sum_{i=1}^{n}(Y_i - \mu(i))^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Definiamo la varianza campionaria corretta S_X^2 come:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{SSR}{n-1}$$

Per il teorema di Cochran, sappiamo che $\bar{Y}$ e S_X^2 sono indipendenti e che $\frac{SSR}{\sigma^2} = \frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2(n-1)$.
Sfruttando questi risultati, possiamo costruire una statistica t-Student. Ricordiamo la definizione operativa:

Nota 10.1: t-Student

Siano $Z \sim \mathcal{N}(0,1)$ e $W \sim \chi^2(k)$ due variabili aleatorie indipendenti. Allora la variabile

$$T = \frac{Z}{\sqrt{W/k}}$$

segue una distribuzione t-Student con k gradi di libertà, denotata con $t(k)$.

Nel nostro caso, la media campionaria standardizzata è Z:

$$\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

e la variabile W è $\frac{(n-1)S_X^2}{\sigma^2}$. Sostituendo nella formula della t-Student otteniamo:

$$\frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S_X^2/\sigma^2}{n-1}}} = \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S_X^2}{\sigma^2}}} = \frac{\bar{Y}-\mu}{S_X/\sqrt{n}} \sim t(n-1)$$

Questa statistica è fondamentale per costruire intervalli di confidenza e test di ipotesi sulla media μ quando la varianza σ^2 non è nota.

Proposizione 10.1: Distribuzione dello stimatore $\hat{\gamma}$

Nelle ipotesi del modello di regressione lineare (Teorema 10.1), lo stimatore di massima verosimiglianza $\hat{\gamma}$ segue una distribuzione Normale multivariata. È uno stimatore corretto, con valore atteso γ, e la sua matrice di covarianza è $\sigma^2(C^TC)^{-1}$. In sintesi:

$$\hat{\gamma} \sim \mathcal{N}(\gamma, \sigma^2(C^TC)^{-1})$$

Dimostrazione 10.2

Lo stimatore $\hat{\gamma}$ è una trasformazione lineare del vettore aleatorio Gaussiano $Y \in \mathbb{R}^n$.

$$\hat{\gamma} = \underbrace{(C^T C)^{-1} C^T}_{N} Y$$

dove N è una matrice deterministica $k \times n$. Dato che $Y(i) \sim \mathcal{N}(\mu(i), \sigma^2)$ sono indipendenti, il vettore Y ha distribuzione $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, dove per ipotesi $\mu = C\gamma$.
Una trasformazione lineare di un vettore Gaussiano è ancora Gaussiana. Per definirla, ne calcoliamo valore atteso e matrice di covarianza.
Valore atteso (Correttezza):

$$E[\hat{\gamma}] = E[NY] = NE[Y] = N\mu = (C^T C)^{-1} C^T (C\gamma) = (C^T C)^{-1}(C^T C)\gamma = \gamma$$

Lo stimatore $\hat{\gamma}$ è quindi corretto (unbiased).
Matrice di Covarianza:

$$\begin{aligned} \mathrm{Cov}(\hat{\gamma}) &= \mathrm{Cov}(NY) = N\mathrm{Cov}(Y)N^T = N(\sigma^2 I_n)N^T = \sigma^2 NN^T \\ &= \sigma^2 \left((C^T C)^{-1} C^T\right) \left((C^T C)^{-1} C^T\right)^T \\ &= \sigma^2 (C^T C)^{-1} C^T C ((C^T C)^{-1})^T \end{aligned}$$

Poiché $(C^T C)$ è simmetrica, anche la sua inversa lo è, quindi $((C^T C)^{-1})^T = (C^T C)^{-1}$.

$$\mathrm{Cov}(\hat{\gamma}) = \sigma^2 (C^T C)^{-1} (C^T C)(C^T C)^{-1} = \sigma^2 (C^T C)^{-1}$$

Avendo determinato media e covarianza, la distribuzione è completamente specificata. □

10.3 Applicazione: Regressione Lineare Semplice

Esempio 10.3: Regressione Lineare Semplice

Consideriamo il modello di regressione lineare semplice:

$$Y(i) \sim \mathcal{N}(\mu(i), \sigma^2) \quad \text{con} \quad \mu(i) = \beta_0 + \beta_1 x(i)$$

Questo è un modello lineare con $k = 2$ parametri, $\gamma = (\beta_0, \beta_1)^T$. Le funzioni base sono $c_1(x) = 1$ e $c_2(x) = x$. La matrice dei regressori C è:

$$C = \begin{pmatrix} 1 & x(1) \\ 1 & x(2) \\ \vdots & \vdots \\ 1 & x(n) \end{pmatrix}$$

Calcoliamo $C^T C$:

$$C^T C = \begin{pmatrix} \sum 1 & \sum x(i) \\ \sum x(i) & \sum x(i)^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & x^2 \end{pmatrix}$$

La sua inversa è:

$$(C^T C)^{-1} = \frac{1}{n(\overline{x^2} - \bar{x}^2)} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Applicando la Proposizione 10.1, la distribuzione del vettore degli stimatori $\hat{\gamma} = (\hat{\beta}_0, \hat{\beta}_1)^T$ è:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{n(\overline{x^2} - \bar{x}^2)} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \right)$$

Grazie a una proprietà fondamentale della distribuzione Normale multivariata, possiamo ricavare immediatamente le distribuzioni dei singoli stimatori (le **distribuzioni marginali**). La media di ogni stimatore è l'elemento corrispondente nel vettore delle medie, e la sua varianza è l'elemento corrispondente sulla diagonale principale della matrice di covarianza.

- Per $\hat{\beta}_0$, prendiamo il primo elemento della media (β_0) e il primo elemento sulla diagonale della matrice di covarianza:

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \overline{x^2}}{n(\overline{x^2} - \bar{x}^2)} \right)$$

- Per $\hat{\beta}_1$, prendiamo il secondo elemento della media (β_1) e il secondo elemento sulla diagonale:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{n(\overline{x^2} - \bar{x}^2)} \right)$$

La somma dei quadrati dei residui è $SSR = \sum (Y(i) - (\hat{\beta}_0 + \hat{\beta}_1 x(i)))^2$. Lo stimatore non distorto della varianza σ^2 è $S_e^2 = \frac{SSR}{n-2}$. Per il Teorema di Cochran, $\frac{SSR}{\sigma^2} \sim \chi^2(n-2)$ ed è indipendente da $\hat{\gamma}$.

11 Richiami di Inferenza Statistica: Il Test d'Ipotesi

L'inferenza statistica fornisce gli strumenti per trarre conclusioni su una popolazione partendo da dati campionari, che sono intrinsecamente affetti da variabilità casuale. Il test di ipotesi è la procedura formale per prendere decisioni in condizioni di incertezza.

Nota 11.1: Analogia: Il Tribunale della Statistica

Un test di ipotesi funziona come un processo in tribunale:

- L'**ipotesi nulla** (H_0) è l'imputato, presunto innocente (status quo).
- I **dati campionari** sono le prove presentate.
- Lo **statistico** è il giudice che valuta se le prove sono abbastanza schiaccianti da rifiutare l'ipotesi nulla.

11.1 Le Componenti Fondamentali di un Test

Definizione 11.1: Ipotesi Nulla (H_0) e Alternativa (H_1)

Ogni test si fonda su due ipotesi contrapposte e mutualmente esclusive:

- **Ipotesi Nulla (H_0)**: L'ipotesi dell'assenza di un effetto. Afferma che ogni differenza osservata è dovuta al caso. È l'ipotesi che cerchiamo di smentire (es. $H_0 : \mu = 100$).
- **Ipotesi Alternativa (H_1)**: L'ipotesi che si contrappone alla nulla. Può essere **bilaterale** (es. $H_1 : \mu \neq 100$) o **unilaterale** (es. $H_1 : \mu > 100$).

Definizione 11.2: Statistica Test e p-value

- **Statistica Test**: Un valore calcolato dai dati campionari che misura quanto questi si discostino da ciò che ci aspetteremmo se H_0 fosse vera.
- **p-value**: La probabilità di osservare un valore della statistica test altrettanto o più estremo di quello ottenuto, *assumendo che l'ipotesi nulla sia vera*. Un p-value piccolo indica che i dati osservati sono improbabili sotto H_0.

La Regola di Decisione. Si fissa una soglia di significatività α (solitamente 0.05).

- **Se p-value** $< \alpha$: Si rifiuta H_0. Il risultato è statisticamente significativo.
- **Se p-value** $\geq \alpha$: Non si rifiuta H_0. Non abbiamo prove sufficienti per smentirla.

11.2 Errori e Potenza di un Test

Nel prendere una decisione basata su un test di ipotesi, ci sono quattro possibili esiti che possono essere riassunti in una tabella di contingenza, spesso chiamata matrice di confusione del test.

		Decisione Presa	
		Non Rifiuto H_0	Rifiuto H_0
Realtà	H_0 è Vera	*Decisione Corretta*	**Errore I Tipo** (α)
	H_0 è Falsa	**Errore II Tipo** (β)	*Decisione Corretta*

Definizione 11.3: Errori di I e II Tipo e Potenza

- **Errore di I Tipo (Falso Positivo)**: Rifiutare un'H_0 vera. La sua probabilità è α.
- **Errore di II Tipo (Falso Negativo)**: Non rifiutare un'H_0 falsa. La sua probabilità è β.
- **Potenza del Test**: La probabilità di rifiutare correttamente un'H_0 falsa. È definita come **Potenza** $= 1 - \beta$.

12 Inferenza nel Modello di Regressione Lineare

Applichiamo ora i concetti di inferenza al modello di regressione lineare per valutare la significatività dei parametri stimati.

12.1 Regressione Lineare Semplice

12.1.1 Test di Ipotesi sul Coefficiente β_1

Obiettivo: Verificare se esista una relazione lineare statisticamente significativa tra la variabile di input X e la variabile di risposta Y, rispondendo alla domanda: "Y dipende davvero da X?".

Ipotesi: Le ipotesi statistiche per il parametro β_1 sono:

- **Ipotesi Nulla** H_0: $\beta_1 = 0$ (non c'è relazione lineare; la pendenza è nulla).
- **Ipotesi Alternativa** H_1: $\beta_1 \neq 0$ (esiste una relazione lineare).

Procedura di Derivazione della Statistica Test: La costruzione della statistica test parte dalla distribuzione dello stimatore B_1 e segue una serie di passaggi di standardizzazione.

1. **Distribuzione dello Stimatore:** Lo stimatore B_1 segue una distribuzione Normale la cui varianza è nota dalla teoria (si veda l'applicazione del Teorema di Cochran).

$$B_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{n(\overline{x^2} - \bar{x}^2)}\right)$$

2. **Funzione Ancillare (Normale):** Standardizzando B_1 si ottiene una variabile con distribuzione $\mathcal{N}(0,1)$, che però dipende ancora dal parametro ignoto σ.

$$\frac{B_1 - \beta_1}{\sigma}\sqrt{n(\overline{x^2} - \bar{x}^2)} \sim \mathcal{N}(0,1)$$

3. **Funzione Ancillare (t-Student):** Sostituendo σ con la sua stima S_e, la distribuzione diventa una t-Student con $n-2$ gradi di libertà.

$$\frac{B_1 - \beta_1}{S_e}\sqrt{n(\overline{x^2} - \bar{x}^2)} \sim t(n-2)$$

4. **Statistica Test Finale:** Valutando la funzione ancillare sotto l'ipotesi nulla $H_0 : \beta_1 = 0$, si ottiene la statistica test finale da calcolare con i dati.

$$T = \frac{B_1}{S_e}\sqrt{n(\overline{x^2} - \bar{x}^2)}$$

Metodi di Decisione: Una volta calcolato il valore T_{obs} della statistica test, si può procedere in due modi:

- **Regione di Accettazione (R.A.):** Si confronta $|T_{obs}|$ con un valore critico q. Se $|T_{obs}| > q$, si rifiuta H_0. Il valore di q è il quantile della distribuzione t-Student tale che $q = F^{-1}_{t(n-2)}(1-\alpha/2)$.
- **p-value:** Si calcola la probabilità di osservare un valore della statistica test altrettanto o più estremo di quello ottenuto, assumendo H_0 vera. Per un test a due code, la formula è:

  $$\text{p-value} = 2 \cdot P(T > |T_{obs}|) = 2 \cdot (1 - F_{t(n-2)}(|T_{obs}|))$$

  dove $F_{t(n-2)}$ è la funzione di ripartizione della distribuzione t-Student con $n-2$ gradi di libertà. Se il p-value è inferiore al livello di significatività α, si rifiuta H_0.

12.1.2 Intervallo di Confidenza per la Risposta Media

Obiettivo: L'obiettivo è stimare un intervallo di valori plausibili non per un singolo punto, ma per la **risposta media** $E[Y|x] = \beta_0 + \beta_1 x$, che è una funzione. Poiché la vera retta delle medie è sconosciuta, si costruisce un "intervallo tubolare" o **banda di confidenza** attorno alla retta stimata $(B_0 + B_1 x)$, all'interno del quale si ha un'elevata fiducia che si trovi la vera retta.

Nota 12.1: Stimatore vs Valore Vero

È importante distinguere tra:

- Il **valore incognito da stimare**: la funzione $\beta_0 + \beta_1 x$.
- Lo **stimatore puntuale**: la retta di regressione calcolata dai dati, $B_0 + B_1 x$.

Per costruire l'intervallo, si analizza la distribuzione di questo stimatore.

Proprietà dello Stimatore: Lo stimatore $B_0 + B_1 x$ è corretto e la sua varianza dipende da x, spiegando la forma a "clessidra" della banda di confidenza.

Dimostrazione 12.1: Calcolo di Media e Varianza dello Stimatore

Calcoliamo il valore atteso e la varianza dello stimatore $\hat{y}(x) = B_0 + B_1 x$.
Valore Atteso (Correttezza) Usando la linearità del valore atteso e sapendo che gli stimatori B_0 e B_1 sono corretti ($E[B_0] = \beta_0$, $E[B_1] = \beta_1$):

$$\begin{aligned} E[B_0 + B_1 x] &= E[B_0] + E[B_1 x] \\ &= E[B_0] + xE[B_1] \\ &= \beta_0 + \beta_1 x \end{aligned}$$

Lo stimatore è quindi corretto.
Varianza Usiamo la formula della varianza di una somma di variabili aleatorie:

$$\begin{aligned} \text{Var}(B_0 + B_1 x) &= \text{Var}(B_0) + \text{Var}(B_1 x) + 2\text{Cov}(B_0, B_1 x) \\ &= \text{Var}(B_0) + x^2 \text{Var}(B_1) + 2x\text{Cov}(B_0, B_1) \end{aligned}$$

Sostituendo i termini dalla matrice di covarianza dello stimatore $(\hat{\beta}_0, \hat{\beta}_1)$ si ottiene:

$$\text{Var}(B_0 + B_1 x) = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{n(\overline{x^2} - \bar{x}^2)} \right]$$

□

Formula Finale: Sfruttando le proprietà dello stimatore e sostituendo σ con la sua stima S_e, si costruisce una quantità pivotale basata sulla distribuzione t-Student. Manipolandola algebricamente, si ottiene l'intervallo di confidenza finale:

$$(B_0 + B_1 x) \pm q \cdot S_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{n(\overline{x^2} - \bar{x}^2)}}$$

dove $q = F^{-1}_{t(n-2)}(1 - \alpha/2)$ è il quantile critico della distribuzione t-Student per un livello di confidenza di $1 - \alpha$.

Calcolo della Varianza con Notazione Matriciale È possibile calcolare le varianze per tutti i punti del dataset in un'unica operazione. Il vettore delle risposte medie stimate è $\hat{Y} = C\hat{\beta}$, dove C è la matrice dei regressori $n \times 2$.

Per trovare la varianza di ogni componente di $\hat{Y}$, si calcola prima l'intera matrice di covarianza di $\hat{Y}$ usando la regola di trasformazione per vettori aleatori:

$$\text{Cov}(\hat{Y}) = \text{Cov}(C\hat{\beta}) = C\text{Cov}(\hat{\beta})C^T$$

Sostituendo $\text{Cov}(\hat{\beta}) = \sigma^2(C^T C)^{-1}$, otteniamo:

$$\text{Cov}(\hat{Y}) = C\left(\sigma^2(C^T C)^{-1}\right)C^T = \sigma^2\left(C(C^T C)^{-1}C^T\right)$$

Il risultato è una matrice $n \times n$, e le varianze richieste sono gli elementi sulla sua **diagonale principale**. Il vettore delle varianze per ogni $\hat{y}_i$ è quindi:

$$\begin{pmatrix} \text{Var}(\hat{y}_1) \\ \vdots \\ \text{Var}(\hat{y}_n) \end{pmatrix} = \text{diag}\left(\text{Cov}(\hat{Y})\right)$$

Dalla Diagonale all'Intervallo di Confidenza Una volta calcolata la matrice di covarianza delle risposte stimate, $\text{Cov}(\hat{Y})$, il passo finale è costruire l'intervallo per ciascuna delle n medie stimate $\hat{y}_i$.

Il i-esimo elemento della diagonale, $[\text{Cov}(\hat{Y})]_{ii}$, rappresenta la varianza della i-esima risposta media stimata, $\text{Var}(\hat{y}_i)$. Questa varianza dipende però dal parametro ignoto σ^2. Per calcolare l'intervallo, dobbiamo prima stimarla, sostituendo σ^2 con la sua stima corretta S_e^2.

Lo **standard error** per la i-esima risposta media stimata, $SE(\hat{y}_i)$, è la radice quadrata di questa varianza stimata:

$$SE(\hat{y}_i) = \sqrt{S_e^2 \cdot [C(C^T C)^{-1}C^T]_{ii}} = S_e\sqrt{[C(C^T C)^{-1}C^T]_{ii}}$$

dove $[\cdot]_{ii}$ indica l'i-esimo elemento diagonale della matrice.

L'intervallo di confidenza al livello $1-\alpha$ per la risposta media al punto x_i è infine:

$$\hat{y}_i \pm q \cdot SE(\hat{y}_i)$$

dove $q = F^{-1}_{t(n-2)}(1-\alpha/2)$ è il quantile critico della distribuzione t-Student. Applicando questa formula per ogni $i = 1, \ldots, n$, si ottengono i limiti superiore e inferiore che definiscono la banda di confidenza.

12.1.3 Intervallo di Predizione per una Osservazione Futura

Obiettivo: Per un nuovo valore di input $\tilde{x}$, predire un intervallo di valori plausibili in cui cadrà una **singola osservazione futura** $\tilde{Y}$. L'obiettivo è creare una "banda di predizione" che contenga, con alta probabilità, i futuri punti dati.

Differenza Chiave rispetto all'Intervallo di Confidenza: Questo intervallo è sempre più largo di quello di confidenza perché deve tenere conto di **due fonti di incertezza**:

1. L'incertezza sulla stima della retta di regressione (la variabilità dello stimatore $B_0 + B_1\tilde{x}$).
2. La variabilità intrinseca della singola osservazione futura $\tilde{Y}$, che fluttua casualmente attorno alla sua media vera (con varianza σ^2).

Nota 12.2: Larghezza dell'intervallo

Mentre l'intervallo di confidenza per la media si stringe all'aumentare dei dati, quello di predizione rimane sempre relativamente largo per via di questa seconda componente di incertezza.

Proprietà dell'Errore di Predizione: La costruzione dell'intervallo si basa sull'analisi dell'errore di predizione, definito come la differenza tra il valore futuro e la sua stima: $\tilde{Y} - (B_0 + B_1\tilde{x})$. Questo errore ha media zero e la sua varianza è la somma delle varianze delle due componenti di incertezza.

Dimostrazione 12.2: Calcolo di Media e Varianza dell'Errore di Predizione

Consideriamo l'errore di predizione $e_{pred} = \tilde{Y} - (B_0 + B_1\tilde{x})$.
Valore Atteso Il valore atteso dell'errore è zero.

$$\begin{aligned} E[e_{pred}] &= E[\tilde{Y} - (B_0 + B_1\tilde{x})] \\ &= E[\tilde{Y}] - E[B_0 + B_1\tilde{x}] \\ &= (\beta_0 + \beta_1\tilde{x}) - (\beta_0 + \beta_1\tilde{x}) = 0 \end{aligned}$$

Varianza Data l'indipendenza tra l'osservazione futura $\tilde{Y}$ e gli stimatori B_0, B_1, la varianza della differenza è la somma delle varianze:

$$\begin{aligned}\mathrm{Var}(e_{pred}) &= \mathrm{Var}(\tilde{Y}) + \mathrm{Var}(B_0 + B_1\tilde{x}) \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(\tilde{x}-\bar{x})^2}{n(\overline{x^2}-\bar{x}^2)}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(\tilde{x}-\bar{x})^2}{n(\overline{x^2}-\bar{x}^2)}\right]\end{aligned}$$

□

Calcolo Matriciale della Varianza di Predizione In analogia con l'intervallo di confidenza, possiamo calcolare le varianze dell'errore di predizione per tutti i punti del dataset simultaneamente. La varianza dell'errore di predizione per una futura osservazione al punto x_i è la somma di due componenti: la varianza del modello e la varianza della stima della media in quel punto.

$$\mathrm{Var}(e_{pred,i}) = \mathrm{Var}(\tilde{Y}_i) + \mathrm{Var}(\hat{y}_i) = \sigma^2 + \mathrm{Var}(\hat{y}_i)$$

Per ottenere il vettore di queste varianze per tutti gli n punti, sommiamo la varianza intrinseca σ^2 a ciascun elemento del vettore delle varianze delle risposte medie stimate.

Ricordiamo che il vettore delle varianze per $\hat{Y}$ è dato dalla diagonale della sua matrice di covarianza:

$$\begin{pmatrix}\mathrm{Var}(\hat{y}_1)\\ \vdots \\ \mathrm{Var}(\hat{y}_n)\end{pmatrix} = \mathrm{diag}\left(\sigma^2 C(C^TC)^{-1}C^T\right)$$

Il vettore delle varianze dell'errore di predizione è quindi:

$$\begin{pmatrix}\mathrm{Var}(e_{pred,1})\\ \vdots \\ \mathrm{Var}(e_{pred,n})\end{pmatrix} = \sigma^2 \cdot \mathbf{1}_n + \mathrm{diag}\left(\sigma^2 C(C^TC)^{-1}C^T\right) = \sigma^2\left(\mathbf{1}_n + \mathrm{diag}\left(C(C^TC)^{-1}C^T\right)\right)$$

dove $\mathbf{1}_n$ è un vettore colonna di n uni.

Lo **standard error** per la i-esima risposta media stimata, $SE(\hat{y}_i)$, è la radice quadrata di questa varianza stimata:

$$SE(\hat{y}_i) = \sqrt{S_e^2 \cdot (1 + [C(C^TC)^{-1}C^T]_{ii})} = S_e\sqrt{1 + [C(C^TC)^{-1}C^T]_{ii}}$$

dove $[\cdot]_{ii}$ indica l'i-esimo elemento diagonale della matrice.

$$\hat{y}_i \pm q \cdot SE(\hat{y}_i)$$

dove $q = F^{-1}_{t(n-2)}(1-\alpha/2)$ è il quantile critico della distribuzione t-Student. Applicando questa formula per ogni $i = 1, \ldots, n$, si ottengono i limiti superiore e inferiore che definiscono la banda di confidenza.

12.2 Regressione Lineare Multipla

La regressione lineare multipla è un'estensione del modello semplice che permette di utilizzare p variabili indipendenti (predittori) per modellare una singola variabile dipendente Y.

12.2.1 Modello e Notazione Matriciale

Il modello per una singola osservazione i assume che la risposta sia una combinazione lineare dei predittori, più un termine di errore Gaussiano:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i, \quad \text{con } e_i \sim \mathcal{N}(0, \sigma^2)$$

Definizione 12.1: Modello Matriciale

Per gestire il modello in modo efficiente, si adotta la notazione matriciale. L'intero set di n equazioni può essere scritto come:

$$Y = X\beta + e$$

dove:

- $Y \in \mathbb{R}^n$ è il vettore delle osservazioni della variabile dipendente.
- $X \in \mathbb{R}^{n \times (p+1)}$ è la **matrice dei predittori**. La sua prima colonna è composta da soli 1 per tenere conto dell'intercetta β_0.
- $\beta \in \mathbb{R}^{p+1}$ è il vettore (ignoto) dei parametri del modello.
- $e \in \mathbb{R}^n$ è il vettore degli errori, con distribuzione $e \sim \mathcal{N}(0, \sigma^2 I_n)$.

Da questo ne consegue che il vettore delle risposte Y segue una distribuzione Normale multivariata:

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

12.2.2 Stima dei Parametri (OLS e MLE)

Sotto l'assunzione di errori Gaussiani, lo stimatore di Massima Verosimiglianza (MLE) coincide con lo stimatore dei Minimi Quadrati Ordinari (Ordinary Least Squares, OLS). L'obiettivo è trovare il vettore di coefficienti $\hat{\beta}$ che minimizza la somma dei quadrati dei residui (SSR).

Proposizione 12.1: Stimatore OLS per β

Lo stimatore dei minimi quadrati per il vettore dei parametri β è dato da:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad \text{(Teorema 10.1)}$$

Questo stimatore è corretto (cioè $E[\hat{\beta}] = \beta$) e la sua distribuzione è:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}) \qquad \text{(Proposizione 10.1)}$$

Definizione 12.2: Stimatore della Varianza dell'Errore σ^2

La stima della varianza dell'errore σ^2 si basa sulla somma dei quadrati dei residui, $SSR = \|Y - X\hat{\beta}\|^2$. Lo stimatore corretto (unbiased) per σ^2 è:

$$S_e^2 = \frac{SSR}{n-p-1}$$

Il denominatore $n-p-1$ rappresenta i gradi di libertà, dati dal numero di osservazioni n meno il numero di parametri stimati $(p+1)$.

Nota 12.3: Legame con il Teorema di Cochran

Il Teorema di Cochran è fondamentale per l'inferenza. Esso garantisce che:

1. La quantità $\frac{SSR}{\sigma^2}$ segue una distribuzione χ^2 con $n-p-1$ gradi di libertà.
2. Lo stimatore dei coefficienti $\hat{\beta}$ è statisticamente indipendente da SSR (e quindi da S_e^2).

Questi due punti sono cruciali perché permettono di costruire la statistica t-Student per il test di ipotesi.

12.2.3 Inferenza sui Singoli Coefficienti (t-test)

Obiettivo: Verificare se una singola variabile di ingresso x_j abbia un potere predittivo statisticamente significativo su Y, al netto delle altre variabili presenti nel modello.

Ipotesi: Per ogni coefficiente β_j (con $j = 1, \dots, p$), il sistema di ipotesi è:

- **Ipotesi Nulla** H_0: $\beta_j = 0$ (la variabile x_j non ha un'influenza lineare su Y).
- **Ipotesi Alternativa** H_1: $\beta_j \neq 0$ (la variabile x_j ha un'influenza lineare significativa su Y).

Procedura di Derivazione: La costruzione della statistica test segue una procedura a più passi, che parte dalla distribuzione dello stimatore.

1. **Distribuzione dello Stimatore B_j:** Partiamo dalla distribuzione del vettore degli stimatori $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$. La distribuzione del singolo stimatore B_j è la sua marginale, anch'essa Normale. La sua varianza è il j-esimo elemento sulla diagonale della matrice di covarianza.

$$B_j \sim \mathcal{N}\left(\beta_j, \sigma^2[(X^TX)^{-1}]_{jj}\right)$$

2. **Funzione Ancillare (Normale):** Standardizzando B_j, otteniamo una variabile $\mathcal{N}(0,1)$ che però dipende ancora dal parametro ignoto σ.

$$\frac{B_j - \beta_j}{\sigma\sqrt{[(X^TX)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

3. **Funzione Ancillare (t-Student):** Sostituiamo σ con la sua stima S_e. Poiché S_e è indipendente da B (per il Teorema di Cochran), la distribuzione cambia da Normale a t-Student con $n-p-1$ gradi di libertà.
$$\frac{B_j - \beta_j}{S_e\sqrt{[(X^TX)^{-1}]_{jj}}} \sim t(n-p-1)$$
4. **Statistica Test Finale:** Valutiamo la funzione ancillare sotto l'ipotesi nulla $H_0 : \beta_j = 0$ per ottenere la statistica test finale.
$$T_j = \frac{B_j}{S_e\sqrt{[(X^TX)^{-1}]_{jj}}}$$

Nota 12.4: Coefficiente Normalizzato

La statistica T_j viene anche chiamata **coefficiente normalizzato**.

- Sotto H_0, ci aspettiamo che T_j assuma valori vicini a 0.
- Sotto H_1, la distribuzione di T_j ha una media diversa da 0, portando a valori più estremi.

Metodi di Decisione: Il test può essere eseguito calcolando la regione di accettazione o, più comunemente, il p-value.

- **Regione di Accettazione (R.A.):** Si definisce un intervallo $[-q, +q]$ dove $q = F^{-1}_{t(n-p-1)}(1-\alpha/2)$ è il quantile critico. Se $|T_j| > q$, si rifiuta H_0.
- **p-value (α_j^*):** Si calcola la probabilità di osservare un valore altrettanto o più estremo di $|T_j|$. Per un test a due code, la formula è:
$$\alpha_j^* = 2 \cdot (1 - F_{t(n-p-1)}(|T_j|))$$
dove F è la funzione di ripartizione della distribuzione t-Student.

Interpretazione dei Risultati: La conclusione del test dipende dal valore del p-value calcolato.

- $\alpha_j^* \gtrsim 30\%$: il p-value è molto grande. Il potere predittivo della variabile x_j è trascurabile e si può considerare di rimuoverla dal modello.
- $\alpha_j^* \lesssim 0.1\%$: il p-value è molto piccolo. La variabile x_j ha un chiaro e forte potere predittivo.
- $0.1\% < \alpha_j^* < 30\%$: la decisione dipende dal contesto, dal task specifico e dalla strategia di selezione del modello adottata.

12.3 Il Problema dei Test Multipli e la Correzione di Bonferroni

Quando eseguiamo un test di ipotesi per ogni predittore in un modello di regressione multipla, stiamo conducendo *test multipli* simultaneamente. Questo introduce un problema statistico significativo: l'inflazione della probabilità di commettere un errore di I specie.

Il Problema: Inflazione dell'Errore di Tipo I Il livello di significatività α (solitamente 0.05) rappresenta la probabilità di rifiutare erroneamente l'ipotesi nulla (commettere un errore di I specie, o falso positivo) in un *singolo* test. Se eseguiamo n test indipendenti, la probabilità di ottenere *almeno un* falso positivo nell'intera famiglia di test aumenta drasticamente. Questa probabilità cumulativa è nota come **Family-Wise Error Rate (FWER)**.

Nota 12.5: Limite Superiore per il FWER

La probabilità di commettere almeno un errore di I tipo nell'intera famiglia di test, ovvero il FWER o α_{globale}, cresce con il numero di test n. Sebbene una stima esatta possa essere complessa, è possibile derivare un limite superiore utilizzando la disuguaglianza di Boole (nota anche come "union bound"). Questa disuguaglianza ci fornisce un'importante garanzia sul controllo dell'errore.

Dimostrazione 12.3: Limite superiore per α_{globale}

Definiamo l'errore di I specie globale come la probabilità di commettere almeno un errore di I specie, assumendo che tutte le ipotesi nulle (H_0) siano vere per ogni test.

$$\begin{aligned}
\alpha_{\text{globale}} &= P(\text{almeno un errore di I specie} \mid \text{tutte le } H_0 \text{ sono vere}) \\
&= P\left(\bigcup_{i=1}^{n} \{\text{rifiuto } H_0 \text{ nel test } i\} \,\middle|\, \text{tutte le } H_0 \text{ sono vere} \right) \\
&\leq \sum_{i=1}^{n} P(\text{errore di I specie nel test } i \mid H_0 \text{ è vera nel test } i) \quad \text{(disuguaglianza di Boole)}
\end{aligned}$$

Se il livello di significatività per ogni singolo test è lo stesso, ovvero $\alpha_i = \bar{\alpha}$ per $i = 1, \ldots, n$, allora la relazione si semplifica come segue:

$$\alpha_{\text{globale}} \leq \sum_{i=1}^{n} \bar{\alpha} = n \cdot \bar{\alpha}$$

Questo dimostra che il FWER è limitato superiormente dal numero di test moltiplicato per il livello di significatività individuale. Ad esempio, con $n = 20$ predittori e $\alpha = 0.05$, la probabilità di ottenere almeno un falso positivo può arrivare fino a $20 \cdot 0.05 = 1$, rendendo quasi certo un risultato errato.

□

La Soluzione: Correzione di Bonferroni Per controllare questo errore cumulativo e mantenere il FWER al di sotto di una soglia desiderata, si possono usare delle procedure di correzione. La più nota e semplice è la correzione di Bonferroni, che deriva direttamente dalla disuguaglianza appena dimostrata.

Definizione 12.3: Correzione di Bonferroni

La procedura consiste nel fissare un livello di significatività globale desiderato (es. $\alpha_{\text{globale}} = 0.05$) e dividere questo valore per il numero di test n che si intende eseguire. Il nuovo livello di significatività α_{corretto} per ogni singolo test sarà:

$$\alpha_{\text{corretto}} = \frac{\alpha_{\text{globale}}}{n}$$

Un risultato per un singolo test verrà considerato statisticamente significativo solo se il suo p-value è inferiore a questa nuova soglia, che è molto più restrittiva. In questo modo si garantisce che $n \cdot \alpha_{\text{corretto}} = n \cdot \frac{\alpha_{\text{globale}}}{n} = \alpha_{\text{globale}}$, mantenendo il FWER sotto controllo.

Nota 12.6: Il Costo della Correzione: Perdita di Potenza

La correzione di Bonferroni è molto conservativa e ha un "prezzo altissimo": **abbassa drasticamente la potenza statistica** dei singoli test. Questo significa che, mentre si è più protetti dai falsi positivi, aumenta la probabilità di commettere errori di II tipo (falsi negativi), non riuscendo a identificare degli effetti che in realtà esistono.

Raccomandazioni Pratiche

- **Ridurre il numero di test:** Ove possibile, è buona norma ridurre il numero di ipotesi da testare *prima* di iniziare l'analisi, ad esempio tramite una pre-selezione delle feature basata su conoscenza del dominio.
- **Pre-specificare le ipotesi:** È metodologicamente cruciale decidere quali test eseguire *prima* di osservare i dati, per evitare pratiche di "p-hacking" o "cherry-picking" che invalidano i risultati statistici.

13 Selezione delle Variabili

Nei modelli di regressione multipla, spesso ci si trova a dover decidere quali predittori includere nel modello finale. L'obiettivo è duplice: da un lato si desidera un modello che spieghi al meglio la variabilità della risposta, dall'altro si cerca un modello **parsimonioso**, ovvero semplice, interpretabile e che eviti l'overfitting. Le procedure di selezione delle variabili, come i metodi *stepwise*, sono algoritmi che automatizzano questo processo.

I metodi stepwise più comuni sono:

- **Selezione Backward**: Si parte dal modello completo con tutte le p variabili e si eliminano una alla volta.
- **Selezione Forward**: Si parte da un modello senza variabili e se ne aggiungono una alla volta.
- **Selezione Mista (Stepwise)**: Unisce le due logiche, permettendo sia di aggiungere che di eliminare variabili ad ogni passo.

13.1 Selezione Backward

La selezione backward è una delle tecniche più diffuse. L'algoritmo parte dal modello più complesso e lo semplifica progressivamente.

Procedura

1. **Inizio**: Si stima il modello completo, includendo tutti i p predittori disponibili.
2. **Verifica**: Si calcolano i p-value per i test t su ogni singolo coefficiente β_j (con $H_0 : \beta_j = 0$).
3. **Decisione**:
   - Se tutti i p-value sono inferiori a una soglia di significatività predefinita (α_{out}), la procedura si arresta. Il modello corrente è il modello finale.
   - Altrimenti, si individua il predittore con il **p-value più alto** e lo si rimuove dal modello.
4. **Iterazione**: Si torna al punto 1, stimando un nuovo modello con i predittori rimanenti, e si ripete il ciclo.

Nota 13.1: Sulla scelta della soglia

In pratica, la soglia suggerita per la rimozione (α_{out}) è spesso alta, tipicamente intorno al **30%** (0.30). Questo perché si vuole essere conservativi e non eliminare variabili potenzialmente utili. Una regola pratica per interpretare i p-value (α_j^*) in questo contesto è:

- Se $\alpha_j^* < 0.1\%$: si è ragionevolmente sicuri che $\beta_j \neq 0$ e la variabile non va tolta.
- Se $0.1\% \leq \alpha_j^* < 30\%$: la situazione è incerta e, di solito, si tende a mantenere la variabile.
- Se $\alpha_j^* \geq 30\%$: non vi è alcuna evidenza che $\beta_j \neq 0$ e si può considerare di togliere la variabile.

Nota 13.2: Sulla variabile con p-value massimo

È importante notare che la scelta di rimuovere la variabile con il p-value massimo è una convenzione, ma non è detto che sia la meno utile. Quando H_0 è vera, il p-value si distribuisce come un'Uniforme(0,1) e non privilegia valori vicini a 1. È utile testare i modelli togliendo una alla volta le variabili con p-value massimali e controllare quale tra questi è il migliore.

Criticità: la Multicollinearità La selezione backward risente pesantemente della presenza di **multicollinearità**, ovvero una forte correlazione tra le variabili di ingresso.

Se sono presenti forti correlazioni, si verificano diverse problematiche:

- La matrice $X^T X$ diventa instabile o, nel caso di correlazione perfetta, non invertibile.
- La varianza degli stimatori dei coefficienti $(\hat{\beta}_j)$ diventa molto elevata. Matematicamente, gli elementi sulla diagonale della matrice $[(X^T X)^{-1}]_{jj}$ diventano grandi.
- Di conseguenza, le statistiche test T_j per i singoli coefficienti saranno piccole.
- I **p-value (α_j^*) risulteranno artificialmente alti**, anche per variabili che potrebbero essere importanti.

Questo inganna l'algoritmo backward, che potrebbe eliminare predittori utili semplicemente perché la loro informazione è ridondante a causa della correlazione con altre variabili. I p-value rimarranno alti finché la multicollinearità non viene ridotta togliendo una delle variabili correlate.

Esempio 13.1: Effetto della multicollinearità

Supponiamo di voler prevedere una misura Y usando quattro predittori molto correlati tra loro, come il numero di dipendenti (x_2), il numero di impiegati (x_3), i posti a sedere (x_4) e un indice di fotoritocco (x_1). Le correlazioni sono $x_1 \approx 22x_2$, $x_3 \approx 20x_2$, $x_4 \approx 5x_2$. Un modello di regressione potrebbe produrre coefficienti instabili e difficili da interpretare. Ad esempio, una stima potrebbe essere:

$$Y = 3 + 0.1x_1 + 1.1x_2 + 0.1x_3 + \dots$$

A causa della forte correlazione, i contributi delle singole variabili si confondono. L'effetto di x_2 viene "spalmato" anche sugli altri coefficienti, rendendoli imprecisi e potenzialmente non significativi singolarmente, anche se collettivamente importanti.

Nota 13.3: Robustezza del modello

All'inizio della procedura backward, il numero di variabili è massimo (p). Se il rapporto tra il numero di osservazioni e il numero di predittori, n/p, è piccolo, la stima della regressione è meno robusta e più soggetta a instabilità.

13.2 Selezione Forward

La selezione Forward è un approccio *bottom-up*, opposto a quello Backward. Invece di semplificare un modello complesso, ne costruisce uno partendo da zero.

Procedura

1. **Inizio**: Si parte dal modello nullo, contenente solo l'intercetta.
2. **Aggiunta**: Si provano ad aggiungere, una alla volta, tutte le variabili non ancora incluse nel modello. Si stima una regressione per ciascuna di queste "prove".
3. **Decisione**: Si sceglie la variabile che, una volta aggiunta, migliora maggiormente il modello. Questa scelta viene guidata da indicatori globali:
   - La variabile che produce il modello con l'Errore Standard della Regressione (S_e) più basso.
   - In modo equivalente, quella che produce l'R^2_{corretto} più alto.
   - Al primo passo, ciò equivale a scegliere la variabile con il p-value (α_j^*) più basso.
4. **Iterazione**: La variabile scelta viene aggiunta definitivamente al modello. Il ciclo riparte dal punto 2, provando ad aggiungere le restanti variabili al nuovo modello, finché non si raggiunge una condizione di arresto.

La procedura si ferma quando l'aggiunta di una qualsiasi delle variabili rimanenti non porta a un miglioramento significativo del modello (ad esempio, quando l'S_e del modello smette di diminuire e inizia ad aumentare).

Nota 13.4: Implementazione nei Software

È comune che i software statistici che automatizzano le procedure stepwise chiedano di fissare una soglia per il valore della statistica F invece che per il p-value.

- **"F to enter"**: soglia per la selezione Forward.
- **"F to remove"**: soglia per la selezione Backward.

Entrambe le soglie vengono usate nella selezione mista. I valori di default di solito corrispondono a un livello di significatività α^* intorno al 30%.

È importante notare che i percorsi della selezione Forward e Backward non sono necessariamente simmetrici e possono portare a modelli finali differenti.

13.3 Metodi Globali (Best Subset Selection)

A differenza dei metodi stepwise, che esplorano solo un percorso limitato tra i possibili modelli, i metodi globali adottano un approccio esaustivo.

Definizione 13.1: Best Subset Selection

La selezione "Best Subset" prevede di testare **tutti i possibili sottoinsiemi** di variabili. Per p predittori, questo significa stimare e valutare 2^p modelli di regressione. A causa della crescita esponenziale del numero di modelli, questo approccio è fattibile solo per un numero di predittori non troppo grande (es. $p < 20$).

Per confrontare un numero così elevato di modelli, è necessario utilizzare delle metriche di performance, o "punteggi" (score).

Criteri di Valutazione (Score) I criteri più utilizzati per confrontare i modelli e selezionare il "migliore" sono:

- S_e (Errore Standard della Regressione): si cerca il modello con l'S_e minimo.
- $R^2_{\textbf{corretto}}$ (R-quadro corretto): si cerca il modello con l'R^2_{corretto} massimo (scelta equivalente a minimizzare S_e).
- **AIC** (Akaike Information Criterion): un criterio che bilancia la bontà di adattamento del modello con la sua complessità. La formula è $AIC = 2k - 2\ln(L)$, dove k è il numero di parametri e L è la verosimiglianza (Likelihood). Si cerca il modello con l'AIC minimo.
- **Validazione**: Si valutano le performance del modello (es. S_e, SSR) su un set di dati di validazione, non usato per la stima.

Nota 13.5: Il Pericolo di Scegliere il "Migliore"

Bisogna essere molto cauti nell'interpretare il modello con lo score in assoluto migliore. I valori degli score (come S_e) calcolati sul campione sono **stime campionarie**, e quindi sono essi stessi delle variabili casuali. Selezionare il modello che ha, ad esempio, l'S_e minimo tra 2^p modelli significa scegliere il minimo tra 2^p valori casuali. È molto probabile che il modello "migliore" lo sia semplicemente per effetto del caso, e che le sue performance non siano altrettanto buone su nuovi dati. Non è opportuno cercare il minimo di una funzione usando una sua stima casuale.

Nonostante le differenze negli approcci, molto spesso i metodi Forward, Backward e Best Subset portano alla selezione di modelli molto simili o identici.

13.4 Criteri Basati sui Coefficienti di Determinazione

Per confrontare l'efficacia di diversi modelli di regressione, specialmente durante il processo di selezione delle variabili, si usano spesso degli indici globali. I più noti sono il coefficiente di determinazione R^2 e la sua versione corretta.

Coefficiente di Determinazione (R^2_D) Il coefficiente di determinazione, noto come R^2, misura la proporzione della variabilità totale della variabile dipendente Y che viene spiegata dal modello di regressione.

Definizione 13.2: Coefficiente di Determinazione R^2_D

È definito come:

$$R^2_D = 1 - \frac{SSR}{SSY}$$

dove:

- **SSY** (Total Sum of Squares): è la **devianza totale** di Y, calcolata come $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Rappresenta la variabilità della variabile risposta prima di considerare i predittori.
- **SSR** (Sum of Squared Residuals): è la **devianza residua**, $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, ovvero la parte di variabilità di Y che il modello *non* è riuscito a spiegare.

Il rapporto $\frac{SSR}{SSY}$ è la frazione della devianza di Y che rimane non spiegata nonostante l'uso dei predittori X. Di conseguenza, R^2_D è la frazione di devianza che *viene* spiegata dal modello.

Per un dato set di dati, la devianza totale **SSY è una quantità fissa**: dipende solo dalla variabile Y. Pertanto, per aumentare R^2_D e migliorare il modello, l'unico modo è diminuire la devianza residua **SSR**, ovvero migliorare la regressione trovando stime dei coefficienti che minimizzino l'errore.

Tuttavia, l'R^2_D ha un grosso limite: **aumenta sempre (o al più rimane uguale) all'aumentare del numero di predittori** p nel modello. Questo perché aggiungendo una variabile, la minimizzazione di SSR sul nuovo set di predittori troverà una soluzione con un errore uguale o inferiore a prima. Di conseguenza, massimizzare l'R^2_D porterebbe sempre a scegliere il modello con tutte le variabili, rendendolo un criterio **inadatto per la selezione** di un modello parsimonioso.

Coefficiente di Determinazione Corretto (R^2_A) Per superare il limite di R^2_D, è stato introdotto il coefficiente di determinazione corretto (*adjusted R-squared*), che penalizza l'aggiunta di variabili inutili.

Definizione 13.3: Coefficiente di Determinazione Corretto R^2_A

È definito utilizzando le varianze campionarie (le somme dei quadrati divise per i rispettivi gradi di libertà):

$$R^2_A = 1 - \frac{S^2_e}{S^2_Y}$$

dove $S^2_e = \frac{SSR}{n-p-1}$ è la stima della varianza degli errori (MSE) e $S^2_Y = \frac{SSY}{n-1}$ è la varianza campionaria di Y.

L'R^2_A **non è monotono** rispetto al numero di variabili p. Quando si aggiunge un nuovo predittore al modello:

- Il numeratore di S^2_e, cioè SSR, diminuisce (o rimane uguale).
- Il denominatore di S^2_e, cioè $n-p-1$, diminuisce anch'esso, perché p aumenta di 1.

Il valore di S^2_e (e quindi di R^2_A) dipende dal bilanciamento di questi due effetti. Se la nuova variabile è utile, la riduzione di SSR sarà significativa e compenserà la perdita di un grado di libertà, facendo diminuire S^2_e (e aumentare R^2_A). Se la variabile è inutile, la riduzione di SSR sarà minima e non basterà a compensare la diminuzione del denominatore; di conseguenza S^2_e aumenterà e R^2_A diminuirà. Questa sua proprietà di "penalizzare" la complessità rende l'R^2_A un criterio valido per la selezione delle variabili, dove l'obiettivo è massimizzarlo.

Esercizio 13.1: HW: Relazione tra R^2_A e R^2_D

Trovare la relazione $R^2_A = a + bR^2_D$ e mostrare che $R^2_A \leq R^2_D$ e che $R^2_A = 1 \iff R^2_D = 1$.

Dimostrazione 13.1

Partiamo dalla definizione di R_A^2 e sostituiamo le formule di S_e^2 e S_Y^2:

$$R_A^2 = 1 - \frac{S_e^2}{S_Y^2} = 1 - \frac{SSR/(n-p-1)}{SSY/(n-1)} = 1 - \frac{SSR}{SSY} \cdot \frac{n-1}{n-p-1}$$

Sappiamo dalla definizione di R_D^2 che $\frac{SSR}{SSY} = 1 - R_D^2$. Sostituendo otteniamo:

$$R_A^2 = 1 - (1 - R_D^2)\frac{n-1}{n-p-1} = 1 - \frac{n-1}{n-p-1} + R_D^2 \frac{n-1}{n-p-1}$$

Questa è la relazione cercata, con $a = 1 - \frac{n-1}{n-p-1} = \frac{-p}{n-p-1}$ e $b = \frac{n-1}{n-p-1}$.

1. Dimostrazione che $R_A^2 \leq R_D^2$: Dobbiamo dimostrare che $1 - \frac{SSR}{SSY} \cdot \frac{n-1}{n-p-1} \leq 1 - \frac{SSR}{SSY}$. Questo equivale a $-\frac{SSR}{SSY} \cdot \frac{n-1}{n-p-1} \leq -\frac{SSR}{SSY}$, e quindi a $\frac{n-1}{n-p-1} \geq 1$. Poiché n, p sono interi e $p \geq 0$, si ha $n - 1 \geq n - p - 1$. Essendo i gradi di libertà positivi, la disuguaglianza è vera.

2. Dimostrazione che $R_A^2 = 1 \iff R_D^2 = 1$: Se $R_D^2 = 1$, allora $SSR = 0$. Di conseguenza, $S_e^2 = 0$ e $R_A^2 = 1 - 0 = 1$. Viceversa, se $R_A^2 = 1$, allora $S_e^2 = 0$. Poiché $S_e^2 = SSR/(n-p-1)$, questo implica $SSR = 0$. Di conseguenza, $R_D^2 = 1 - 0 = 1$. L'equivalenza è dimostrata.

□

Esercizio 13.2: HW: R^2 e Correlazione Lineare

Verificare che nel caso della regressione lineare semplice ($p = 1$), il coefficiente di determinazione R_D^2 è uguale al quadrato del coefficiente di correlazione lineare di Pearson tra x e Y.

Dimostrazione 13.2

Nella regressione semplice, $R^2_D = \frac{ESS}{SSY}$, dove $ESS = \sum(\hat{Y}_i - \bar{Y})^2$ è la devianza spiegata. Il modello è $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Sostituendo $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$, otteniamo:

$$\hat{Y}_i - \bar{Y} = (\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i) - \bar{Y} = \hat{\beta}_1 (x_i - \bar{x})$$

Quindi, la devianza spiegata è:

$$ESS = \sum(\hat{\beta}_1 (x_i - \bar{x}))^2 = \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}$$

Sostituiamo la stima di $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$:

$$ESS = \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

Infine, calcoliamo R^2_D:

$$R^2_D = \frac{ESS}{SSY} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

Questa espressione è esattamente il quadrato del coefficiente di correlazione lineare $r_{x,y}$:

$$(r_{x,y})^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}\right)^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

L'identità è quindi verificata.

□

14 Analisi della Varianza (ANOVA) per il Confronto tra Modelli

L'Analisi della Varianza (ANOVA) offre un metodo formale per confrontare modelli di regressione annidati, ovvero quando un modello può essere considerato un caso speciale di un altro.

Definizione 14.1: Confronto tra modelli annidati

Siano dati due insiemi di variabili candidati, C e $\tilde{C}$, tali che $C \subset \tilde{C}$. Indichiamo con d il numero di variabili in C e con $\tilde{d}$ il numero di variabili in $\tilde{C}$. Vogliamo testare se le variabili aggiuntive presenti in $\tilde{C}$ (cioè in $\tilde{C} \setminus C$) portano un contributo significativo al modello.

Poiché il modello basato su $\tilde{C}$ contiene più variabili, il suo adattamento ai dati sarà sempre migliore o uguale a quello del modello basato su C. Questo si traduce in:

$$R^2_D(\tilde{C}) \geq R^2_D(C) \iff SSR(\tilde{C}) \leq SSR(C)$$

L'ANOVA ci permette di quantificare se la riduzione dell'errore (la differenza $SSR(C) - SSR(\tilde{C})$) è statisticamente significativa o solo dovuta al caso.

La Statistica F per il Confronto Si definisce la somma dei quadrati addizionale (SS_D) come la riduzione dell'errore ottenuta passando dal modello più piccolo al più grande:

$$SS_D := SSR(C) - SSR(\tilde{C})$$

Questa quantità è associata a $\tilde{d} - d$ gradi di libertà. La statistica test per il confronto tra i due modelli è data dal seguente rapporto:

$$V := \frac{SS_D/(\tilde{d} - d)}{SSR(C)/(n - d - 1)}$$

Teorema 14.1: Distribuzione della statistica V (di Cochran)

Sotto l'ipotesi nulla H_0 che le variabili aggiuntive in $\tilde{C}$ non siano utili (ovvero $H_0 : \beta_j = 0 \quad \forall j \in \tilde{C} \setminus C$), la statistica V segue una legge F di Fisher:

$$V \sim F(\tilde{d} - d, n - d - 1)$$

Sotto l'ipotesi alternativa, il valore di V è tipicamente grande, perciò il test è unilaterale destro. Il p-value si calcola come $\alpha^* = 1 - F_{(\tilde{d}-d, n-d-1)}(V)$.

Esercizio 14.1: HW: Test F Globale di Regressione

Cosa succede se si confronta il modello completo con tutte le p variabili contro il modello nullo (contenente solo l'intercetta)?

Dimostrazione 14.1

Questo è il test F globale, che verifica se complessivamente la regressione è significativa. In questo caso, il modello più piccolo è $C = \emptyset$ (modello nullo) e quello più grande è $\tilde{C} = \{x_1, \ldots, x_p\}$ (modello completo). Abbiamo:

- $d = |C| = 0$
- $\tilde{d} = |\tilde{C}| = p$

Le somme dei quadrati dei residui sono:

- $SSR(C) = SSR(\emptyset)$, che corrisponde alla devianza totale di Y, quindi $SSR(C) = SSY$.
- $SSR(\tilde{C})$ è la devianza residua del modello completo, che chiamiamo semplicemente SSR.

La somma dei quadrati addizionale è $SS_D = SSR(C) - SSR(\tilde{C}) = SSY - SSR$. I gradi di libertà per il numeratore sono $\tilde{d} - d = p - 0 = p$. I gradi di libertà per il denominatore sono $n - d - 1 = n - 0 - 1 = n - 1$.
Sostituendo nella formula della statistica V:

$$V = \frac{SS_D/(\tilde{d} - d)}{SSR(C)/(n - d - 1)} = \frac{(SSY - SSR)/p}{SSY/(n-1)}$$

L'ipotesi nulla è $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$. Un p-value ($\alpha^*$) molto piccolo per questo test indica che almeno uno dei predittori è utile per spiegare la Y.

□

Nota 14.1: Applicazione nei Metodi Stepwise

Questo test F è alla base delle decisioni nei metodi di selezione stepwise. Ad esempio, in una procedura Forward, ad ogni passo si valuta se l'aggiunta di una nuova variabile x_j (quindi $\tilde{C} = C \cup \{x_j\}$) porta a una riduzione significativa dell'errore. Questo test F è equivalente al test t sul coefficiente della variabile aggiunta.

15 Metodi di Regolarizzazione

La regolarizzazione è una tecnica utilizzata per evitare l'overfitting nei modelli di regressione. Spesso, un modello in overfitting è caratterizzato da coefficienti β_j molto grandi, che corrispondono a una funzione approssimante con derivate elevate e un andamento molto "nervoso". L'idea fondamentale della regolarizzazione è di modificare la funzione di costo per penalizzare i modelli che presentano coefficienti di grandi dimensioni.

Mentre la regressione standard (Least Squares Estimation, LSE) si limita a minimizzare la somma dei quadrati dei residui (SSR):

$$\min_{\beta} (\text{SSR}) = \min_{\beta} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 \right)$$

i modelli regolarizzati aggiungono a questa un termine di penalità.

15.1 Regressione Ridge

La regressione Ridge aggiunge una penalità proporzionale alla somma dei quadrati dei coefficienti (norma L_2).

Definizione 15.1: Funzione di costo Ridge

La funzione di costo per la regressione Ridge è:

$$\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \right)$$

Il termine $\alpha \sum \beta_j^2$ è la penalità. L'iperparametro $\alpha \geq 0$ controlla l'intensità di questa regolarizzazione: più α è grande, più i coefficienti sono "spinti" verso lo zero.

La regressione Ridge è efficace nel ridurre la varianza del modello, ma tende a "restringere" (*shrinkage*) i coefficienti verso lo zero senza mai annullarli completamente.

15.2 Regressione Lasso

La regressione Lasso (Least Absolute Shrinkage and Selection Operator) utilizza una penalità basata sulla somma dei valori assoluti dei coefficienti (norma L_1).

Definizione 15.2: Funzione di costo Lasso

La funzione di costo per la regressione Lasso è:

$$\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Una formulazione equivalente consiste nel minimizzare l'SSR sotto il vincolo che la somma dei valori assoluti dei coefficienti sia inferiore a una certa soglia α:

$$\min_{\beta}(\text{SSR}) \quad \text{con il vincolo} \quad \sum_{j=1}^{p} |\beta_j| \leq \alpha$$

La caratteristica più importante del Lasso è che, a differenza della Ridge, è in grado di **impostare alcuni coefficienti esattamente a zero**. Questo significa che la regressione Lasso non solo regolarizza il modello, ma effettua anche una **selezione automatica delle variabili**.

Nota 15.1: Regressione Elastic Net

Esiste anche la regressione "Elastic Net", che combina entrambe le penalizzazioni, L_1 e L_2, nella sua funzione di costo. Questo approccio ibrido può essere particolarmente utile in presenza di alta collinearità tra i predittori.

15.3 Considerazioni Pratiche

- **Standardizzazione delle Variabili**: Per far sì che i termini di penalità abbiano senso, è fondamentale che i coefficienti β_j abbiano una scala confrontabile. Per questo motivo, è prassi comune **standardizzare le variabili** predittrici (portandole ad avere media 0 e deviazione standard 1) prima di applicare la regolarizzazione.
- **Scelta degli Iperparametri**: Gli iperparametri di regolarizzazione (α per la Ridge, λ per la Lasso) non vengono stimati dal modello, ma devono essere scelti con cura. Solitamente si utilizza un set di validazione (o la cross-validation) per testare diverse configurazioni e scegliere quella che produce il modello migliore.
- **Ottimizzazione**: A differenza della regressione lineare standard che ha una soluzione analitica, i modelli regolarizzati vengono generalmente risolti tramite un **ottimizzatore iterativo**.

15.4 Nota Finale: il Fenomeno del Double Descent

Tradizionalmente, la teoria statistica suggerisce che la performance di un modello (misurata dall'errore su dati di test) segua una curva a U al crescere della sua complessità. L'errore diminuisce fino a un punto ottimale, per poi risalire a causa dell'overfitting. Tuttavia, in contesti di machine learning moderni, si è osservato un fenomeno più complesso e controintuitivo, noto come **Double Descent**.

Nota 15.2: Il Double Descent

In contesti realistici, che utilizzano ottimizzatori iterativi e beneficiano di una forma di *regolarizzazione implicita*, il comportamento dell'errore di test può mostrare una seconda discesa. Come mostrato nel grafico, una volta superata la **soglia di interpolazione** (dove il numero di parametri p eguaglia il numero di dati n), l'errore di test, dopo aver raggiunto un picco, inizia a diminuire di nuovo. Questo avviene nel regime **over-parameterizzato** ($p > n$), tipico di molti modelli attuali come le reti neurali profonde.

Questo fenomeno suggerisce che, quando il rapporto n/p è minore di 1 e tende a 0 (ovvero il modello diventa estremamente più complesso dei dati), l'overfitting tende a diminuire, contrariamente a quanto previsto dalla teoria classica.

16 Estensione dei Modelli Lineari

Quando la relazione tra i predittori e la variabile di risposta non è puramente lineare, o quando l'effetto di un predittore dipende dal valore di un altro, è possibile estendere il framework della regressione lineare introducendo termini non lineari.

16.1 Regressione Polinomiale

Si ricorre alla regressione polinomiale quando il modello lineare non è soddisfacente. Un segnale tipico di questa necessità è osservare degli andamenti non lineari (es. a parabola) nei grafici dei residui rispetto a una delle variabili predittrici.

Metodologia L'idea è di aggiungere al modello dei nuovi predittori "fittizi" (*dummy*) che sono semplicemente monomi di grado 2 o superiore delle variabili originali. Per esempio, partendo da una singola variabile x, possiamo costruire:

- **Modello lineare**: $Y = \beta_0 + \beta_1 x + \epsilon$
- **Modello quadratico**: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- **Modello generale di grado d**: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d + \epsilon$

Nota 16.1: Linearità nei Parametri

È fondamentale notare che, sebbene il modello sia non lineare *nelle variabili*, è ancora un modello **lineare nei parametri** β. Stiamo semplicemente conducendo una regressione lineare multipla in cui i predittori sono $x_1 = x, x_2 = x^2, \ldots, x_d = x^d$. Pertanto, può essere stimato con il metodo dei minimi quadrati (LSE) esattamente come un modello lineare standard.

Esempio 16.1: Correzione della Non-Linearità

Se il grafico dei residui R_i contro un predittore x_2 mostra un'evidente forma a parabola, possiamo sospettare una relazione quadratica. Aggiungendo un nuovo predittore $x_{\text{new}} := x_2^2$ al modello, la non-linearità viene spesso corretta. Se il termine aggiunto risulta statisticamente significativo, l'SSR del nuovo modello sarà inferiore, indicando un miglior adattamento ai dati.

16.2 Termini di Interazione

Il modello lineare standard assume che l'effetto di ogni predittore sulla risposta sia indipendente dagli altri (modello additivo). I termini di interazione vengono introdotti per modellare situazioni in cui l'effetto di una variabile cambia in base al valore di un'altra.

Definizione 16.1: Interazione

Un termine di interazione è un nuovo predittore creato moltiplicando due o più variabili originali (es. $x_1 x_2$). Il modello che lo include è detto *semilineare*.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

L'inclusione del termine di interazione cambia radicalmente l'interpretazione del modello. Mentre nel modello additivo l'effetto di un aumento unitario di x_1 su Y è sempre β_1, nel modello con interazione questo effetto diventa $\beta_1 + \beta_3 x_2$. In altre parole, **il modo in cui Y dipende da x_1, dipende a sua volta dal valore di x_2 (e viceversa)**. Geometricamente, il modello non rappresenta più un piano, ma una superficie curva.

16.3 Principi Guida e Selezione delle Variabili

Quando si lavora con termini polinomiali e di interazione, è cruciale seguire delle regole per costruire modelli coerenti e per guidare la selezione delle variabili.

Nota 16.2: Regola Gerarchica

Il principio gerarchico stabilisce che, se un modello include un termine di ordine superiore, deve includere anche tutti i termini di ordine inferiore che lo compongono.

- Se il modello contiene x^k, deve contenere anche $x, x^2, \ldots, x^{k-1}$.
- Se il modello contiene l'interazione $x_1^a x_2^b$, deve contenere anche tutti i monomi $x_1^c x_2^d$ che lo dividono.

Questo assicura che il modello sia interpretabile e invariante rispetto a semplici traslazioni dell'origine degli assi.

Selezione delle Variabili in Modelli Non Lineari Le procedure stepwise devono essere adattate per gestire i termini non lineari, tenendo sempre conto del principio gerarchico.

- **Approccio Forward**: È generalmente preferibile perché i termini polinomiali (es. x e x^2) sono spesso correlati. La procedura è la seguente:
  1. Si parte dal modello nullo.
  2. Ad ogni passo, i candidati per l'inclusione non sono solo le variabili originali, ma anche tutti i termini di ordine superiore (potenze, interazioni) che si possono creare con le variabili già presenti nel modello, nel rispetto della regola gerarchica.
- **Approccio Backward**:
  1. Si parte da un modello lineare con tutte le variabili.
  2. Si analizzano i residui per identificare eventuali non-linearità e guidare l'aggiunta di termini polinomiali o di interazione.
  3. Si inizia a rimuovere le variabili una alla volta, partendo da quelle con il p-value più alto, ma **con il vincolo di non violare mai il principio gerarchico**.

17 Regressione Pesata

La regressione pesata è una tecnica che si utilizza quando i dati non sono **omoschedastici**, ovvero quando la varianza degli errori non è costante per tutte le osservazioni (violazione dell'assunto di omoschedasticità, detta **eteroschedasticità**).

Quando si usa? L'eteroschedasticità si manifesta in diversi contesti:

- Dati derivanti da una distribuzione di **Poisson**, dove la varianza è approssimativamente uguale alla media ($\sigma_i^2 \approx \mu_i$).
- Dati la cui incertezza è proporzionale al valore misurato (es. errore del 10%), dove la deviazione standard è proporzionale alla media ($\sigma_i \propto \mu_i$), e quindi la varianza è proporzionale al quadrato della media ($\sigma_i^2 \propto \mu_i^2$).
- Dati Binomiali, dove la varianza di una proporzione dipende dalla proporzione stessa.
- Quando i grafici dei residui mostrano una forma a "imbuto", indicando che la dispersione dei residui cambia al variare dei valori di un predittore o dei valori stimati.

Metodo di Risoluzione: Minimi Quadrati Pesati (WLS) Quando l'assunzione di omoschedasticità cade, ogni osservazione Y_i segue una distribuzione Normale con una propria varianza: $Y_i \sim N(\mu_i, \sigma_i^2)$. Per trovare gli stimatori dei parametri β, si applica il principio di Massima Verosimiglianza (MLE), come visto in precedenza (Esempio 8.5).

La log-verosimiglianza per questo modello è:

$$l(\beta) = C - \sum_{i=1}^{n} \frac{(Y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2}{2\sigma_i^2}$$

Massimizzare questa funzione rispetto ai β equivale a minimizzare la parte che dipende da loro, ovvero la somma dei quadrati degli errori. Tuttavia, ogni termine della somma è ora "pesato" dall'inverso della sua varianza.

Massimizzare $l(\beta)$ è quindi equivalente a risolvere il problema di minimizzazione:

$$\min_{\beta} \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} (Y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 \right)$$

Questo è esattamente il criterio dei **Minimi Quadrati Pesati (WLS)**. I pesi (w_i) emergono naturalmente dalla derivazione della massima verosimiglianza e sono il reciproco delle varianze:

$$w_i = \frac{1}{\sigma_i^2}$$

L'idea intuitiva di dare più "peso" alle osservazioni più precise (con varianza più piccola) è quindi una conseguenza diretta del principio di massima verosimiglianza.

Nota 17.1: Metodo Equivalente: OLS su Dati Trasformati

La regressione pesata può essere implementata in modo molto semplice come una regressione lineare standard (OLS) applicata a dati trasformati. Se si conoscono i fattori di proporzionalità degli errori (cioè $\sigma_i \propto r_i$), si possono definire delle nuove variabili:

$$\tilde{Y}_i := \frac{Y_i}{r_i} \qquad \text{e} \qquad \tilde{x}_{ij} := \frac{x_{ij}}{r_i}$$

Questo include anche il termine per l'intercetta, che da $x_{i0} = 1$ diventa $\tilde{x}_{i0} = 1/r_i$.
Minimizzare la somma dei quadrati dei residui per queste variabili trasformate, $\sum(\tilde{Y}_i - \sum \beta_j \tilde{x}_{ij})^2$, è matematicamente equivalente a risolvere il problema dei minimi quadrati pesati originale.

18 Gestione delle Variabili

La corretta gestione e codifica delle variabili è un passo fondamentale nella costruzione di qualsiasi modello, inclusa la regressione lineare. Le variabili possono essere di diverse tipologie, e ciascuna richiede un trattamento specifico.

Le variabili di ingresso si possono classificare principalmente in tre categorie:

- **Numeriche**: Variabili quantitative su cui è possibile eseguire operazioni aritmetiche (es. età, temperatura, reddito).
- **Dicotomiche**: Variabili che possono assumere solo due valori (es. maschio/femmina, sì/no, acceso/spento).
- **Categoriali**: Variabili che rappresentano un gruppo o una categoria, a loro volta suddivisibili in nominali e ordinali.

18.1 Variabili Dicotomiche

Le variabili dicotomiche, che rappresentano due sole categorie, sono un caso speciale ma molto comune di variabili categoriali.

Come Variabili di Ingresso Quando usate come predittori, vengono tipicamente codificate con i numeri 0 e 1. In un modello di regressione:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \epsilon$$

il coefficiente β_j associato a una variabile dicotomica x_j ha un'interpretazione molto diretta: rappresenta la **differenza media nella risposta** Y tra la categoria codificata come 1 e la categoria codificata come 0. Ad esempio, se x_j è 1 per "femmina" e 0 per "maschio", β_j misura di quanto, in media, la risposta Y per le femmine differisce da quella per i maschi.

Nota 18.1: Variabili Dicotomiche come Risposta

Utilizzare una variabile dicotomica come variabile di risposta Y non è appropriato per la regressione lineare standard. Per questo tipo di problema esistono modelli specifici, tra cui:

- **Analisi Discriminante**
- **Regressione Logistica**

18.2 Variabili Categoriali

Le variabili categoriali richiedono una codifica specifica per poter essere incluse in un modello di regressione.

Variabili Ordinali vs. Nominali È importante distinguere tra:

- **Variabili Nominali**: Le categorie non hanno un ordinamento intrinseco. Esempi sono le stagioni (primavera, estate, autunno, inverno), le regioni d'Italia o il tipo di carburante.

- **Variabili Ordinali**: Le categorie possiedono un ordine naturale. Esempi sono il titolo di studio (medie, maturità, laurea), una classifica (I, II, III) o la gravità di un sintomo.

Le variabili ordinali possono essere codificate numericamente (es. 1, 2, 3...) rispettando l'ordine, ma bisogna essere consapevoli che questa scelta impone una distanza uniforme tra i livelli, il che potrebbe non essere sempre corretto.

Codifica One-Hot per Variabili Nominali Per inserire una variabile categoriale nominale in un modello di regressione, la tecnica standard è la **codifica one-hot**, nota anche come creazione di *variabili dummy*.

Definizione 18.1: Codifica One-Hot con Categoria di Riferimento

Una variabile categoriale con k livelli viene trasformata in $k-1$ nuove variabili dicotomiche (0/1). Un livello viene escluso e scelto come **categoria di riferimento** (o *default*) per evitare una perfetta multicollinearità.

Esempio 18.1: Codifica delle stagioni

Consideriamo la variabile "stagione" con 4 livelli: primavera, estate, autunno, inverno. Scegliamo "inverno" come categoria di riferimento. Creiamo 3 nuove variabili dummy:

- $x_{\text{primavera}}$: vale 1 se la stagione è primavera, 0 altrimenti.
- x_{estate}: vale 1 se la stagione è estate, 0 altrimenti.
- x_{autunno}: vale 1 se la stagione è autunno, 0 altrimenti.

Il modello di regressione diventerà: $Y = \beta_0 + \beta_1 x_{\text{primavera}} + \beta_2 x_{\text{estate}} + \beta_3 x_{\text{autunno}} + \cdots + \epsilon$.

L'interpretazione dei coefficienti è cruciale: ogni coefficiente β_j misura l'effetto medio sulla risposta Y di quella categoria **rispetto alla categoria di riferimento**.

- Per l'inverno (riferimento), tutte le dummy sono 0 e il valore atteso di Y è legato a β_0.
- Per la primavera, il valore atteso di Y è legato a $\beta_0 + \beta_1$. Quindi, β_1 rappresenta la differenza media in Y tra la primavera e l'inverno.

Nota 18.2: Variabili Categoriali come Risposta

Utilizzare una variabile dicotomica o categoriale come variabile di risposta Y in una regressione lineare standard è problematico. Per questi casi, esistono modelli più appropriati come la **regressione logistica**, gli alberi di classificazione, le random forest o le reti neurali.

18.3 Interpretazione e Test sui Coefficienti delle Variabili Dummy

Una volta inserite le variabili dummy nel modello, è fondamentale saper interpretare correttamente i loro coefficienti e capire come testare ipotesi specifiche.

Esempio 18.2: Interpretazione dei coefficienti delle stagioni

Riprendendo il modello delle stagioni con "inverno" come riferimento:

$$Y = \beta_0 + \beta_1 x_{\text{primavera}} + \beta_2 x_{\text{estate}} + \beta_3 x_{\text{autunno}} + \epsilon$$

Supponiamo che la stima del modello fornisca i seguenti coefficienti: $\beta_0 = 16.4$, $\beta_1 = 5.4$, $\beta_2 = 6.7$, $\beta_3 = 1.8$. Il valore medio della risposta Y per ciascuna stagione è:

- **Inverno (riferimento)**: $E[Y|\text{inverno}] = \beta_0 = 16.4$
- **Primavera**: $E[Y|\text{primavera}] = \beta_0 + \beta_1 = 16.4 + 5.4 = 21.8$
- **Estate**: $E[Y|\text{estate}] = \beta_0 + \beta_2 = 16.4 + 6.7 = 23.1$
- **Autunno**: $E[Y|\text{autunno}] = \beta_0 + \beta_3 = 16.4 + 1.8 = 18.2$

Il test t standard su un coefficiente, ad esempio $H_0 : \beta_3 = 0$, verifica se c'è una differenza significativa tra l'autunno e l'inverno. Se non si rifiuta l'ipotesi nulla, si conclude che le due categorie sono statisticamente indistinguibili.

Confronto tra Categorie non di Riferimento Per confrontare due categorie non di riferimento (es. primavera vs. estate), non è sufficiente guardare i singoli p-value. Bisogna testare l'ipotesi $H_0 : \beta_1 = \beta_2$.

Esercizio 18.1: HW: Testare la differenza tra primavera ed estate

Come si può testare l'ipotesi $H_0 : \beta_1 = \beta_2$? Ci sono due approcci.

Dimostrazione 18.1

1. **Test Manuale (Test Lineare Generale)**: L'ipotesi è equivalente a $H_0 : \beta_1 - \beta_2 = 0$. Si può costruire una statistica t per questa combinazione lineare di coefficienti:
$$T = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{SE}(\hat{\beta}_1 - \hat{\beta}_2)}$$
dove l'errore standard al denominatore si calcola dalla matrice di varianza-covarianza dei coefficienti:
$$\text{SE}(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)}$$
Questo test permette di ottenere un p-value per l'ipotesi di uguaglianza.

2. **Cambiare la Categoria di Riferimento**: Un metodo più semplice e pratico consiste nel modificare il modello, scegliendo una delle categorie da confrontare (es. "estate") come nuovo livello di riferimento. A questo punto si riesegue la regressione. Il nuovo coefficiente associato alla variabile "primavera" misurerà direttamente la differenza media di Y rispetto all'"estate", e il suo test t standard fornirà il p-value desiderato.

□

18.4 Criticità e Note Pratiche

Interazione con le Procedure Stepwise

- **Stepwise Backward**: Se la procedura elimina una variabile dummy (es. x_{autunno}), significa che il test non ha trovato una differenza significativa tra quella categoria e la categoria di riferimento. L'effetto è che le due categorie vengono fuse o "collassate".
- **Stepwise Forward**: La procedura sceglie autonomamente quale categoria usare implicitamente come riferimento, e non è detto che la scelta sia ottimale. Potrebbe non accorgersi che due categorie (nessuna delle quali è il riferimento) sono molto simili e andrebbero accorpate.

Nota 18.3: Attenzione

- **Aumento di p**: L'uso della codifica one-hot può aumentare drasticamente il numero di predittori p, specialmente per variabili con molte categorie. Questo peggiora il rapporto n/p e può rendere il modello instabile. Per risolvere il problema, si possono accorpare manualmente alcune categorie in macro-categorie più generali e significative.
- **Non-Linearità**: Per le variabili dummy (0/1), l'aggiunta di potenze (es. x^2) è inutile, poiché $x^2 = x$. Similmente, l'interazione tra dummy della stessa variabile categoriale (es. $x_{\text{primavera}} \cdot x_{\text{estate}}$) è sempre zero ed è quindi inutile. Sono invece utili le interazioni tra dummy di variabili categoriali diverse.

18.5 Gestione delle Variabili Numeriche

Anche le variabili numeriche possono essere suddivise in base alla loro natura, e questo influenza la decisione di trasformarle o meno prima di inserirle in un modello.

Variabili di tipo "Differenza" vs. "Rapporto"

- **Tipo "Differenza"**: Sono variabili per cui le differenze additive sono significative e interpretabili (es. Quoziente Intellettivo, statura, temperatura). Generalmente, queste variabili vengono inserite nel modello così come sono, in un'ottica additiva del tipo $Y = \beta_0 + \beta_1 x_1 + \cdots + \epsilon$.
- **Tipo "Rapporto"**: Sono variabili tipicamente positive che possono coprire diversi ordini di grandezza, per le quali i rapporti (e quindi le variazioni percentuali) sono più significativi delle differenze assolute (es. reddito, popolazione di una città, intensità sonora). Per queste variabili è spesso opportuno applicare una trasformazione logaritmica.

Nota 18.4: Trasformazioni Lineari

Una trasformazione lineare della variabile di risposta ($\tilde{Y} = mY + q$) non cambia la sostanza della regressione. I coefficienti vengono riscalati ($\tilde{\beta}_j = m\beta_j$), ma i risultati dei test di ipotesi (p-value, α^*) e gli indici di bontà del modello (R^2_D, R^2_A) rimangono invariati.

Le trasformazioni non-lineari, invece, possono essere usate per migliorare le proprietà distributive di una variabile. Ad esempio, se un predittore x ha una distribuzione asimmetrica a destra (*right-skewed*), applicare una trasformazione come la radice quadrata ($\sqrt{x}$) o il logaritmo ($\ln(x)$) può renderne la distribuzione più simmetrica, aiutando a soddisfare le assunzioni del modello di regressione lineare.

18.6 Modelli Logaritmici e Interpretazione

Una delle trasformazioni più potenti e comuni è l'uso del logaritmo sulla variabile di risposta, specialmente quando questa è di tipo "rapporto".

Il Modello Log-Level Quando si modella il logaritmo di Y (modello log-level), la relazione diventa:

$$\ln(Y) = \beta_0 + \beta_1 x_1 + \cdots + \epsilon$$

Questo modello è ancora lineare nei parametri, ma implica una relazione **moltiplicativa** sulla scala originale di Y:

$$Y = e^{\beta_0} \cdot e^{\beta_1 x_1} \cdot \ldots \cdot e^{\epsilon}$$

Questo approccio è preferibile quando si ritiene che i predittori abbiano un effetto percentuale, e non assoluto, sulla risposta.

Esempio 18.3: Interpretazione del coefficiente in un modello Log-Level

Supponiamo di modellare lo stipendio (Y) in funzione di una variabile dicotomica x_1 (1 se una persona ha frequentato un certo corso, 0 altrimenti). Il modello è $\ln(Y) = \beta_0 + \beta_1 x_1$.

- Per chi non ha frequentato il corso ($x_1 = 0$): $\ln(Y_0) = \beta_0 \implies Y_0 = e^{\beta_0}$

- Per chi ha frequentato il corso ($x_1 = 1$): $\ln(Y_1) = \beta_0 + \beta_1 \implies Y_1 = e^{\beta_0} e^{\beta_1} = Y_0 \cdot e^{\beta_1}$

Il coefficiente β_1 non rappresenta più una differenza additiva. Il fattore e^{β_1} è un **fattore moltiplicativo**. Se la stima del modello fornisce $\hat{\beta}_1$ tale che $e^{\hat{\beta}_1} = 1.40$, significa che frequentare il corso è associato a un aumento dello stipendio del 40% ($Y_1 = 1.40 \cdot Y_0$). La variazione percentuale si calcola come $(e^{\hat{\beta}_1} - 1) \times 100\%$.

19 Regressione Logistica

La regressione logistica è un modello statistico utilizzato quando la variabile di risposta Y è categoriale. A differenza della regressione lineare, non modella direttamente il valore della risposta, ma la **probabilità** che la risposta appartenga a una determinata categoria.

19.1 Regressione Logistica Binomiale

Il caso più comune è quello in cui la variabile di risposta è **dicotomica**, ovvero può assumere solo due valori (es. 0/1, successo/fallimento, malato/sano).

Il Modello L'obiettivo è modellare la probabilità che la risposta sia 1, dato un set di predittori X. Poiché una probabilità deve essere compresa tra 0 e 1, si utilizza la **funzione logistica (o sigmoide)** per mappare la combinazione lineare dei predittori (chiamata **logit**) nell'intervallo (0, 1).

Definizione 19.1: Funzione Logistica (Sigmoide)

Il modello lega i predittori alla probabilità di successo $p(X) = P(Y = 1|X)$ attraverso la seguente relazione:

$$p(X) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

dove $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ è il logit.

Stima tramite Massima Verosimiglianza (MLE) Per stimare i coefficienti β_j, si assume che ogni osservazione y_i sia il risultato di una **prova Bernoulliana**. La probabilità di successo di questa prova, p_i, è data dal modello logistico. La funzione di massa di probabilità per una singola osservazione è quindi:

$$P(Y = y_i|X_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

La verosimiglianza (*likelihood*) per l'intero dataset di N osservazioni indipendenti è il prodotto delle singole probabilità. Per semplicità, si massimizza la sua versione logaritmica (la **log-verosimiglianza**):

$$\begin{aligned} l(\beta) &= \log\left(\prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1-y_i}\right) \\ &= \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \end{aligned}$$

Questa espressione è esattamente l'opposto della funzione di costo di **cross-entropy**. Pertanto, massimizzare la verosimiglianza è equivalente a minimizzare la cross-entropy.

Nota 19.1: Inferenza sui Coefficienti

A differenza della regressione lineare, i coefficienti β_j non seguono una distribuzione t di Student. Tuttavia, si basano su approssimazioni alla distribuzione normale per grandi campioni. L'inferenza statistica e la selezione delle variabili sono quindi possibili e si basano comunemente su:

- **Wald Test**: Un test simile al t-test per verificare l'ipotesi nulla che un singolo coefficiente sia uguale a zero.
- **Likelihood-Ratio Test (LRT)**: Un test più robusto per confrontare modelli annidati e valutare la significatività di un gruppo di variabili.

Una volta addestrato, il modello si usa per prevedere la probabilità di successo per nuove osservazioni.

19.2 Regressione Logistica Multinomiale

Questa è la generalizzazione del modello a casi in cui la variabile di risposta ha più di due categorie (es. $k \in \{1, \dots, m\}$).

Il Modello Il modello predice un vettore di probabilità $\pi = (\pi_1, \pi_2, \dots, \pi_m)$, dove π_k è la probabilità che l'osservazione appartenga alla categoria k, con il vincolo che $\sum \pi_k = 1$. Per fare ciò, si usa la **funzione softmax**, che è la generalizzazione della sigmoide.

Definizione 19.2: Funzione Softmax

Per ogni categoria k si calcola un logit z_k con un set di parametri dedicato w_k:

$$z_k = w_{k0} + w_{k1}x_1 + \cdots + w_{kp}x_p$$

La probabilità per la categoria k è data dalla funzione softmax, che normalizza i logit:

$$\pi_k = \text{softmax}(z_k) = \frac{e^{z_k}}{\sum_{j=1}^{m} e^{z_j}}$$

Stima del Modello La stima dei parametri w avviene, anche in questo caso, massimizzando la verosimiglianza, la cui derivazione è discussa in 8.4. Ciò corrisponde a minimizzare la **loss di cross-entropy categoriale**, che per una singola osservazione è:

$$H(y_i, \pi_i) = -\sum_{k=1}^{m} y_{ik} \log(\pi_{ik})$$

dove y_i è il vettore one-hot della classe vera (es. '[0, 1, 0]') e π_i è il vettore delle probabilità predette. La loss totale da minimizzare è la media di questa quantità su tutto il dataset.

Nota 19.2: Nota Pratica sull'Implementazione della Loss

Sebbene la cross-entropy sia concettualmente definita tra la distribuzione di probabilità vera (y_i) e quella predetta (π_i), nella pratica è sconsigliato calcolare manualmente la funzione softmax per ottenere le probabilità π_i e poi passarle alla funzione di loss.
Il calcolo esplicito dell'esponenziale nella funzione softmax (e^{z_k}) può portare a instabilità numerica (errori di *overflow* o *underflow*) quando i valori dei logit z_k sono molto grandi o molto piccoli.

Per questo motivo, le librerie software di machine learning (come TensorFlow, PyTorch, Scikit-learn) offrono implementazioni della cross-entropy loss che prendono in input direttamente i **logit** grezzi. Queste funzioni utilizzano internamente degli accorgimenti matematici (come il trucco "Log-Sum-Exp") per calcolare la loss in modo numericamente stabile.
Pertanto, la regola pratica è: **passare sempre i logit, non le probabilità, alla funzione di loss fornita dalla libreria**.

20 Analisi della Varianza (ANOVA)

L'Analisi della Varianza (ANOVA) è una tecnica statistica che può essere vista come una variante della regressione lineare, utilizzata quando le variabili di ingresso sono **categoriali**. Il modello analizza la relazione tra una o più variabili categoriali indipendenti e una variabile dipendente numerica, assumendo che l'errore sia additivo, Gaussiano e omoschedastico.

I principali tipi di ANOVA sono:

- **ANOVA a una via**: utilizzata quando si ha una sola variabile di ingresso categoriale.
- **ANOVA a due vie**: utilizzata con due variabili di ingresso categoriali. Si distingue tra disegni con o senza repliche (più di un'osservazione per ogni combinazione di categorie).

20.1 ANOVA a Una Via (One-Way ANOVA)

Questo è il modello ANOVA più semplice. Si usa per confrontare le medie di tre o più gruppi (categorie) definiti da un singolo fattore.

Il Modello Statistico L'idea centrale è che ogni categoria della variabile di ingresso possa avere una propria media per la variabile di risposta. Si assume che la varianza all'interno di ogni gruppo sia la stessa per tutti (omoschedasticità).

Definizione 20.1: Modello ANOVA a una via

Sia Y_{ij} la j-esima osservazione nel i-esimo gruppo, dove $i = 1, \dots, m$ (numero di gruppi) e $j = 1, \dots, n_i$ (numerosità del gruppo i). Il modello statistico è:

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

I dati sono assunti indipendenti. I parametri incogniti del modello sono le m medie dei gruppi $\mu_1, \mu_2, \dots, \mu_m$ e la varianza comune σ^2.

Stima dei Parametri del Modello La stima dei parametri incogniti del modello avviene a partire dai dati campionari.

- **Stima delle Medie (μ_i)**: La stima naturale per la media di ciascun gruppo è la sua media campionaria:
$$\hat{\mu}_i = \bar{Y}_{i*} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$
- **Stima della Varianza Comune (σ^2)**: La stima della varianza comune σ^2 è un processo a due passi.
    1. **Stima per singolo gruppo**: Per prima cosa, si calcola la varianza campionaria per ciascun gruppo i, definita come S_i^2:
$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i*})^2$$

Ognuno di questi S_i^2 è uno stimatore corretto di σ^2 (cioè $E[S_i^2] = \sigma^2$), ma utilizza solo una parte dei dati. Si ottengono così m stimatori diversi per lo stesso parametro.

2. **Combinazione degli stimatori (Stimatore Pooled)**: Per ottenere una stima singola e più robusta, si combinano gli m stimatori in una media pesata. Lo stimatore risultante è detto **stimatore pooled della varianza** o **Within** (S_W^2), noto anche come **Mean Square Within** (MSW):

$$S_W^2 = \frac{\sum_{i=1}^{m}(n_i - 1)S_i^2}{N - m}$$

Definizione 20.2: Somma dei Quadrati Entro i Gruppi (SSW)

Il numeratore dello stimatore S_W^2 è la Somma dei Quadrati Entro i Gruppi (Sum of Squares Within), o **devianza within**:

$$SS_W = \sum_{i=1}^{m}(n_i - 1)S_i^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i*})^2$$

Questa quantità rappresenta la variabilità totale all'interno dei gruppi e corrisponde alla Somma dei Quadrati dei Residui (SSR) del modello ANOVA. Lo stimatore della varianza può essere scritto come:

$$S_W^2 = \frac{SS_W}{N - m}$$

Inferenza sui Singoli Parametri Disponendo degli stimatori e delle loro distribuzioni, è possibile costruire test e intervalli di confidenza sui singoli parametri, come ad esempio per una singola media μ_i tramite la statistica $T = \frac{\bar{Y}_{i*} - \mu_i}{S_W/\sqrt{n_i}} \sim t(N - m)$.

20.2 Il Test F nell'ANOVA a Una Via

Lo scopo principale dell'ANOVA a una via è testare se le medie dei vari gruppi sono tutte uguali o se almeno una è diversa. Questo permette di determinare se la variabile categoriale ha un effetto statisticamente significativo sulla variabile di risposta.

Ipotesi del Test Il test fondamentale dell'ANOVA confronta due ipotesi:

- **Ipotesi Nulla (H_0)**: Tutte le medie dei gruppi sono uguali.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$$

  Sotto H_0, la variabile categoriale non ha alcun effetto sulla risposta Y.

- **Ipotesi Alternativa (H_1)**: Non tutte le medie sono uguali (almeno una è diversa). Sotto H_1, la variabile categoriale ha un effetto su Y.

Decomposizione della Varianza La logica del test F si basa sulla scomposizione della variabilità totale dei dati in due parti: la variabilità *tra* i gruppi e la variabilità *all'interno* dei gruppi.

Definizione 20.3: Somma dei Quadrati Totale (SST)

La Somma dei Quadrati Totale misura la variabilità totale di tutte le osservazioni attorno alla media generale ($\bar{Y}_{**}$).

$$SS_T = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{**})^2$$

Definizione 20.4: Somma dei Quadrati Tra i Gruppi (SSB)

La Somma dei Quadrati Tra i Gruppi, o **devianza between**, misura la variabilità delle medie di ciascun gruppo attorno alla media generale. Ha $m-1$ gradi di libertà.

$$SS_B = \sum_{i=1}^{m} n_i(\bar{Y}_{i*} - \bar{Y}_{**})^2$$

Queste quantità sono legate alla devianza within (SS_W) dalla relazione fondamentale:

$$SS_T = SS_B + SS_W$$

La Statistica del Test F Il test si basa sul confronto tra la varianza stimata "tra" i gruppi e quella "entro" i gruppi. Queste stime sono chiamate **Medie dei Quadrati** (Mean Squares).

Definizione 20.5: Medie dei Quadrati (MS)

Le Medie dei Quadrati si ottengono dividendo le somme dei quadrati per i rispettivi gradi di libertà.

- **Media dei Quadrati Tra i Gruppi**: $MS_B = S_B^2 = \frac{SS_B}{m-1}$
- **Media dei Quadrati Entro i Gruppi**: $MS_W = S_W^2 = \frac{SS_W}{N-m}$

L'idea è che MS_W è sempre una stima corretta della varianza σ^2, mentre MS_B lo è solo se H_0 è vera. Se H_0 è falsa, MS_B tenderà ad essere più grande.

Teorema 20.1: Test F per l'ANOVA a una via

La statistica del test è il rapporto tra le due medie dei quadrati:

$$F_{ANOVA} = \frac{MS_B}{MS_W} = \frac{S_B^2}{S_W^2}$$

Sotto l'ipotesi nulla H_0, questa statistica segue una distribuzione F di Fisher con $m-1$ e $N-m$ gradi di libertà:

$$F_{ANOVA} \sim F(m-1, N-m)$$

Poiché sotto H_1 la statistica F tende ad assumere valori grandi, il test è **unilaterale destro**.

20.3 Verifica delle Assunzioni e Note Pratiche

Nota 20.1: Attenzione alle Cose Pratiche

- **Analisi dei Residui**: È fondamentale verificare le assunzioni del modello analizzando i residui ($R_{ij} = Y_{ij} - \bar{Y}_{i*}$). Graficamente, si deve controllare che siano approssimativamente Gaussiani, omoschedastici (varianza costante tra i gruppi), indipendenti dal predittore e che non presentino outlier evidenti.
- **Trasformazioni**: Se le assunzioni non sono soddisfatte, a volte una trasformazione non lineare della variabile di risposta Y (es. logaritmo) può risolvere il problema.
- **Accettazione di H_0**: Se il test F non porta a rifiutare l'ipotesi nulla, significa che non c'è evidenza statistica che la variabile categoriale influenzi la risposta. In questo caso, si potrebbe considerare di ignorare la suddivisione in gruppi e analizzare i dati come un unico campione Gaussiano i.i.d. $Y_i \sim N(\mu, \sigma^2)$.

20.4 ANOVA a Due Vie (Two-Way ANOVA)

L'ANOVA a due vie si utilizza quando si vuole analizzare l'effetto di **due variabili categoriali** (o fattori) su una variabile di risposta numerica. Questo modello permette di studiare non solo l'effetto individuale di ciascun fattore (effetto principale), ma anche se l'effetto di un fattore dipende dal livello dell'altro (effetto di interazione).

Notazione e Modelli Sia Y_{ijk} la k-esima osservazione (o replica) per il livello i del primo fattore (righe) e il livello j del secondo fattore (colonne). La media della cella (i, j) viene scomposta per separare i diversi effetti.

Definizione 20.6: Modelli Additivo e con Interazione

- **Modello con Interazione** (usato con repliche, $l \geq 2$): scompone la media in quattro parti:
$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$
dove μ è la media globale, α_i è l'effetto principale del fattore di riga, β_j è l'effetto principale del fattore di colonna e γ_{ij} è il **termine di interazione**.
- **Modello Additivo** (usato senza repliche, $l = 1$): assume che non ci sia interazione:
$$\mu_{ij} = \mu + \alpha_i + \beta_j$$
In questo modello, l'effetto di un fattore è costante a tutti i livelli dell'altro fattore. Per garantire l'unicità del modello, si impongono vincoli di somma zero sugli effetti ($\sum \alpha_i = 0, \sum \beta_j = 0$).

Analisi per il Modello Additivo (Senza Repliche) Quando si ha una sola osservazione per cella ($l = 1$), si assume che il modello sia additivo. La stima dei parametri e degli effetti si basa sulle medie di riga ($\bar{Y}_{i*}$), di colonna ($\bar{Y}_{*j}$) e sulla media generale ($\bar{Y}_{**}$).

Definizione 20.7: Stimatori degli Effetti Principali

Gli effetti principali sono stimati come lo scostamento delle medie parziali dalla media generale:
$$\hat{\alpha}_i = \bar{Y}_{i*} - \bar{Y}_{**} \qquad \text{e} \qquad \hat{\beta}_j = \bar{Y}_{*j} - \bar{Y}_{**}$$

Definizione 20.8: Valori Previsti e Residui

Il valore previsto (o stimato) dal modello per la cella (i, j) è $\hat{Y}_{ij} = \bar{Y}_{i*} + \bar{Y}_{*j} - \bar{Y}_{**}$. I residui sono la differenza tra i valori osservati e quelli previsti:
$$R_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i*} - \bar{Y}_{*j} + \bar{Y}_{**}$$

Definizione 20.9: Stima della Varianza d'Errore

La Somma dei Quadrati dell'Errore (SS_e) è la somma dei residui al quadrato. Lo stimatore corretto della varianza σ^2, indipendente dalle ipotesi, è la Media dei Quadrati dell'Errore (MS_e):

$$\hat{\sigma}^2 = S_e^2 = MS_e = \frac{SS_e}{(m-1)(n-1)} \quad \text{dove} \quad SS_e = \sum_{i,j} R_{ij}^2$$

Questa stima, scalata, segue una distribuzione Chi-Quadrato: $\frac{SS_e}{\sigma^2} \sim \chi^2_{(m-1)(n-1)}$.

L'analisi completa dei dati e i test di ipotesi vengono quindi riassunti nelle seguenti tabelle.

Tabella 2: Tabella ANOVA a due fattori (caso senza repliche) - Scomposizione della Varianza

Fonte di variabilità	**Somma di quadrati (SS)**	**Gradi di libertà (df)**
Riga (Fattore 1)	$SS_r = n\sum(\bar{Y}_{i*} - \bar{Y}_{**})^2$	$m-1$
Colonna (Fattore 2)	$SS_c = m\sum(\bar{Y}_{*j} - \bar{Y}_{**})^2$	$n-1$
Errore	$SS_e = \sum(Y_{ij} - \bar{Y}_{i*} - \bar{Y}_{*j} + \bar{Y}_{**})^2$	$(m-1)(n-1)$

Tabella 3: Tabella ANOVA a due fattori (caso senza repliche) - Test di Ipotesi

Ipotesi Nulla	**Statistica del Test**	**Un test con significatività α deve...**	**p-dei-dati se $D_{ts} = v$**
H_0 : Tutte le $\alpha_i = 0$	$D_{ts} = \frac{SS_r/(m-1)}{SS_e/((m-1)(n-1))}$	rifiutare H_0 se $D_{ts} > F_{\alpha,m-1,(m-1)(n-1)}$	$P(F_{m-1,(m-1)(n-1)} \geq v)$
H_0 : Tutte le $\beta_j = 0$	$D_{ts} = \frac{SS_c/(n-1)}{SS_e/((m-1)(n-1))}$	rifiutare H_0 se $D_{ts} > F_{\alpha,n-1,(m-1)(n-1)}$	$P(F_{n-1,(m-1)(n-1)} \geq v)$

Teorema 20.2: Test F per il modello additivo

Per testare gli effetti principali si usano due distinte statistiche F, che confrontano la varianza spiegata da ciascun fattore con la varianza residua $MS_e = S_e^2$.

- **Test per le Righe**: $F_r = \frac{MS_r}{S_e^2} \sim F(m-1, (m-1)(n-1))$
- **Test per le Colonne**: $F_c = \frac{MS_c}{S_e^2} \sim F(n-1, (m-1)(n-1))$

Analisi per il Modello con Repliche La presenza di repliche ($l \geq 2$) è fondamentale perché permette di stimare separatamente la media di ogni cella μ_{ij} e, di conseguenza, di isolare l'effetto di interazione dalla variabilità casuale.

- **Stimatori delle Medie e degli Effetti**: Lo stimatore naturale per la media della cella (i, j) è la media campionaria delle osservazioni in quella cella:

$$\hat{\mu}_{ij} = \bar{Y}_{ij*} = \frac{1}{l}\sum_{k=1}^{l} Y_{ijk}$$

Da questi si ricavano gli stimatori per gli effetti: $\hat{\alpha}_i = \bar{Y}_{i**} - \bar{Y}_{***}$, $\hat{\beta}_j = \bar{Y}_{*j*} - \bar{Y}_{***}$ e $\hat{\gamma}_{ij} = \bar{Y}_{ij*} - \bar{Y}_{i**} - \bar{Y}_{*j*} + \bar{Y}_{***}$.

- **Residui e Stima della Varianza d'Errore**: I residui sono la variabilità interna a ciascuna cella, attorno alla propria media:
$$R_{ijk} = Y_{ijk} - \bar{Y}_{ij*}$$
La Somma dei Quadrati dell'Errore (SS_e) è la somma dei residui al quadrato. Lo stimatore corretto e sempre valido della varianza σ^2 è la Media dei Quadrati dell'Errore (MS_e):
$$\hat{\sigma}^2 = S_e^2 = MS_e = \frac{SS_e}{mn(l-1)} \quad \text{dove} \quad SS_e = \sum_{i,j,k} R_{ijk}^2$$

L'analisi completa, che scompone la varianza totale, è riassunta nelle tabelle seguenti.

Tabella 4: Tabella ANOVA a due fattori (caso con repliche) - Scomposizione della Varianza

Fonte di variabilità	**Somma di quadrati (SS)**	**Gradi di libertà (df)**
Riga (Fattore 1)	$SS_r = nl\sum(\bar{Y}_{i**} - \bar{Y}_{***})^2$	$m-1$
Colonna (Fattore 2)	$SS_c = ml\sum(\bar{Y}_{*j*} - \bar{Y}_{***})^2$	$n-1$
Interazione	$SS_{int} = l\sum(\bar{Y}_{ij*} - \bar{Y}_{i**} - \bar{Y}_{*j*} + \bar{Y}_{***})^2$	$(m-1)(n-1)$
Errore	$SS_e = \sum(Y_{ijk} - \bar{Y}_{ij*})^2$	$mn(l-1)$

Tabella 5: Tabella ANOVA a due fattori (caso con repliche) - Test di Ipotesi

Ipotesi Nulla	**Statistica del Test**	**Un test con significatività α deve...**	**p-dei-dati se $F = v$**
H_0^{int}: Le γ_{ij} sono tutte nulle	$F_{int} = \frac{SS_{int}/((m-1)(n-1))}{SS_e/(mn(l-1))}$	rifiutare H_0 se $F_{int} > F_{\alpha,(m-1)(n-1),mn(l-1)}$	$P(F_{(m-1)(n-1),mn(l-1)} \geq v)$
H_0^r: Le α_i sono tutte nulle	$F_r = \frac{SS_r/(m-1)}{SS_e/(mn(l-1))}$	rifiutare H_0 se $F_r > F_{\alpha,m-1,mn(l-1)}$	$P(F_{m-1,mn(l-1)} \geq v)$
H_0^c: Le β_j sono tutte nulle	$F_c = \frac{SS_c/(n-1)}{SS_e/(mn(l-1))}$	rifiutare H_0 se $F_c > F_{\alpha,n-1,mn(l-1)}$	$P(F_{n-1,mn(l-1)} \geq v)$

Teorema 20.3: Test F per il modello con interazione

La procedura di test è gerarchica. Il test per l'interazione ($H_0 : \gamma_{ij} = 0$) è il primo e più importante da valutare.

$$F_{int} = \frac{MS_{int}}{S_e^2} \sim F((m-1)(n-1), mn(l-1))$$

Se l'interazione non è significativa, si procede a testare gli effetti principali.

- **Righe**: $F_r = \frac{MS_r}{S_e^2} \sim F(m-1, mn(l-1))$
- **Colonne**: $F_c = \frac{MS_c}{S_e^2} \sim F(n-1, mn(l-1))$

Nota 20.2: Semplificazione del Modello

Se l'esito del test per uno degli effetti principali (es. righe) porta ad accettare l'ipotesi nulla, si può considerare di eliminare quella variabile dal modello, riducendo di fatto l'analisi a un'ANOVA a una via sul fattore rimanente.

21 Il Test del Chi-Quadrato

Il test del Chi-Quadrato (χ^2) è una delle famiglie più importanti di test statistici. Viene utilizzato principalmente come **test di adattamento** (*goodness-of-fit*) per verificare se una distribuzione di dati osservata si adegua a una distribuzione teorica attesa.

Le principali applicazioni includono:

- **Test elementare**: per distribuzioni discrete con un numero limitato di categorie (es. testare l'equità di un dado).
- **Generalizzazione**: per distribuzioni con molti valori o continue.
- **Test di adattamento a famiglie di distribuzioni**: per verificare se i dati seguono una certa famiglia di distribuzioni (es. Gaussiana) senza specificarne i parametri.
- **Tabelle di contingenza**: per verificare l'indipendenza tra due variabili categoriali.

21.1 Test del Chi-Quadrato Elementare (Goodness-of-Fit)

Questo test è il caso base e si applica quando si vuole confrontare la distribuzione di un campione con una specifica distribuzione discreta.

Scopo e Ipotesi del Test Si parte da un campione $X_1, \ldots, X_n$ proveniente da una distribuzione incognita ϕ. L'obiettivo è verificare se i dati si conformano a una distribuzione "candidata" ϕ_0 completamente specificata.

- **Ipotesi Nulla (H_0)**: la distribuzione del campione è quella candidata.
$$H_0 : \phi = \phi_0$$
- **Ipotesi Alternativa (H_1)**: la distribuzione del campione è diversa da quella candidata.
$$H_1 : \phi \neq \phi_0$$

Frequenze Osservate e Attese La procedura si basa sul confronto tra le frequenze osservate nel campione e quelle che ci aspetteremmo se l'ipotesi nulla fosse vera.

- **Frequenze Osservate (O_j)**: si conta quante volte ciascuna delle k categorie possibili appare nel campione.
$$O_j = \text{numero di volte in cui è uscito il valore } j$$
- **Frequenze Attese (A_j)**: sono le frequenze che ci aspetteremmo in media se i dati seguissero la distribuzione ϕ_0. Si calcolano come:
$$A_j = n \cdot p_{0,j}$$
dove n è la numerosità del campione e $p_{0,j}$ è la probabilità della categoria j secondo ϕ_0.

La Statistica del Test Per misurare la discrepanza tra le frequenze osservate e quelle attese, si usa la statistica Chi-Quadrato di Pearson.

Definizione 21.1: Statistica Chi-Quadrato di Pearson

La statistica del test, indicata con W o χ^2, è calcolata come:

$$W = \sum_{j=1}^{k} \frac{(O_j - A_j)^2}{A_j}$$

Questa statistica è "piccola" quando i dati osservati sono vicini a quelli attesi (supportando H_0) e "grande" quando c'è una forte discrepanza (supportando H_1).

Teorema 21.1: Distribuzione della Statistica del Test

Sotto l'ipotesi nulla H_0, la statistica del test W segue asintoticamente (per n grande) una **distribuzione Chi-Quadrato con $k-1$ gradi di libertà**:

$$W \overset{H_0}{\sim} \chi^2_{k-1}$$

Poiché valori grandi della statistica forniscono evidenza contro H_0, il test è **unilaterale destro**.

Nota 21.1: Condizioni di Validità e Potenza

- **Regola Pratica**: l'approssimazione alla distribuzione χ^2 è considerata valida se tutte le frequenze attese sono $A_j \geq 1$ e almeno l'80% di esse sono $A_j \geq 5$.
- **Potenza**: il test del Chi-Quadrato è noto per essere poco potente. Richiede scostamenti notevoli o campioni molto grandi per rigettare H_0 quando le differenze tra ϕ e ϕ_0 sono modeste.
- **Statistica G**: negli ultimi anni si tende a preferire la statistica G (o del rapporto di verosimiglianza), che segue la stessa distribuzione χ^2_{k-1} ma con un'approssimazione spesso migliore:

$$G = 2\sum_{j=1}^{k} O_j \ln\left(\frac{O_j}{A_j}\right)$$

21.2 Estensioni del Test del Chi-Quadrato

Il test elementare può essere generalizzato per affrontare due scenari più complessi: l'adattamento a distribuzioni continue e l'adattamento a famiglie di distribuzioni con parametri non specificati.

Test di Adattamento per Leggi Continue Quando la distribuzione candidata ϕ_0 è continua (es. Esponenziale, Normale, etc.), non è possibile contare le occorrenze di singoli valori. La procedura viene quindi adattata "discretizzando" il supporto della distribuzione.

1. **Binning**: Si suddivide il dominio della variabile in k intervalli (o *bin*) disgiunti, $B_1, B_2, \ldots, B_k$.

2. **Frequenze Osservate**: Si contano quante osservazioni del campione cadono in ciascun bin per ottenere le frequenze osservate O_j.

3. **Frequenze Attese**: Si calcola la probabilità $p_{0,j}$ che un'osservazione cada nel bin j secondo la distribuzione nulla ϕ_0. Per una distribuzione continua con densità $f_0(x)$, questo si ottiene integrando:
$$p_{0,j} = P(X \in B_j) = \int_{B_j} f_0(x)dx$$
Le frequenze attese sono quindi $A_j = n \cdot p_{0,j}$.

Una volta ottenute le frequenze O_j e A_j, si procede con la statistica di Pearson W e il confronto con la distribuzione χ^2_{k-1} esattamente come nel caso elementare.

Nota 21.2: Strategia di Binning

A parità di k, il test è più potente e l'approssimazione alla χ^2 è migliore se le frequenze attese A_j sono più grandi possibile. Per questo motivo, la strategia ottimale non è creare bin di larghezza uguale, ma **bin equiprobabili**, ovvero intervalli scelti in modo tale che la probabilità $p_{0,j}$ (e quindi la frequenza attesa A_j) sia la stessa per ogni bin. La potenza del test è comunque legata al numero e alla posizione dei bin scelti.

Test di Adattamento a Famiglie di Distribuzioni Spesso non si vuole testare l'adattamento a una distribuzione *completamente* specificata, ma a una *famiglia* di distribuzioni, lasciando i parametri liberi. Un esempio classico è il test di Gaussianità.

- **Ipotesi**: $H_0 : X \sim N(\mu, \sigma^2)$ con μ e σ^2 non specificati.

La procedura è la seguente:

1. **Stima dei Parametri**: Si stimano i parametri incogniti della distribuzione a partire dal campione (es. $\hat{\mu} = \bar{x}$ e $\hat{\sigma}^2 = s^2$).

2. **Test Standard**: si esegue il test di adattamento del Chi-Quadrato (con la procedura di binning vista sopra) usando come distribuzione nulla ϕ_0 quella della famiglia specificata, ma con i parametri appena stimati (es. $N(\bar{x}, s^2)$).

3. **Correzione dei Gradi di Libertà**: la distribuzione della statistica del test W è ancora una Chi-Quadrato, ma i suoi gradi di libertà vengono ridotti.

Teorema 21.2: Gradi di Libertà con Parametri Stimati

Se si stimano s parametri dal campione per definire la distribuzione nulla ϕ_0, la statistica del test W si distribuisce come una Chi-Quadrato con $k - 1 - s$ gradi di libertà.

$$W \overset{H_0}{\sim} \chi^2_{k-1-s}$$

Nel caso del test di Gaussianità, si stimano $s = 2$ parametri (μ e σ^2), quindi i gradi di libertà sono $k - 3$.

Nota 21.3: Uso Pratico dei Test di Adattamento

L'uso di test formali per verificare a priori o a posteriori le assunzioni di un modello (es. la Gaussianità dei residui) non è sempre una buona idea. Con campioni molto grandi, questi test diventano "troppo potenti" e possono rigettare l'ipotesi nulla per deviazioni minime e praticamente irrilevanti. L'ipotesi che spesso ci interessa non è se i dati siano *esattamente* Normali, ma se siano *sufficientemente vicini* alla Normalità perché le nostre procedure funzionino. Per questo, è spesso preferibile affiancare o sostituire il test formale con un **controllo qualitativo e grafico** (es. un Q-Q plot).

21.3 Test del Chi-Quadrato per Tabelle di Contingenza

Questa è una delle applicazioni più comuni del test del Chi-Quadrato e serve a verificare se esista un'associazione tra due variabili categoriali. A differenza della regressione o dell'ANOVA, questa analisi è simmetrica: non si definisce una variabile di ingresso e una di uscita, ma si studia la relazione reciproca $X \leftrightarrow Y$.

Scopo e Ipotesi L'obiettivo del test è determinare se due variabili categoriali sono statisticamente indipendenti o se sono associate.

- **Ipotesi Nulla (H_0)**: Le due variabili sono indipendenti. In termini di probabilità, la probabilità congiunta è il prodotto delle probabilità marginali: $P(X = i, Y = j) = P(X = i) \cdot P(Y = j)$.
- **Ipotesi Alternativa (H_1)**: Le due variabili non sono indipendenti (esiste un'associazione).

Si tratta di un test **non parametrico**, in quanto non fa alcuna ipotesi sulla famiglia di distribuzione sottostante, il che lo rende meno potente di test come l'ANOVA se le assunzioni di quest'ultima sono soddisfatte.

Frequenze Osservate e Attese Il test si basa sul confronto tra le frequenze congiunte osservate nel campione e quelle che ci aspetteremmo in un'ipotetica situazione di indipendenza.

- **Frequenze Osservate (O_{ij})**: sono i conteggi effettivi per ogni cella della tabella di contingenza, ovvero il numero di osservazioni che presentano simultaneamente la categoria i della prima variabile e la categoria j della seconda.
- **Frequenze Attese (A_{ij})**: sono i conteggi che ci aspetteremmo in ogni cella se H_0 (indipendenza) fosse vera. Poiché le probabilità marginali sono incognite, vengono stimate dai dati. La formula per le frequenze attese è:

$$A_{ij} = \frac{(\text{Totale della riga } i) \cdot (\text{Totale della colonna } j)}{\text{Totale generale}}$$

Definizione 21.2: Statistica del Test per l'Indipendenza

La statistica del test è la consueta statistica Chi-Quadrato di Pearson, calcolata su tutte le

$m \times n$ celle della tabella:

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}$$

Teorema 21.3: Distribuzione e Gradi di Libertà

Sotto l'ipotesi nulla di indipendenza, la statistica del test W segue asintoticamente una distribuzione Chi-Quadrato con $(m-1)(n-1)$ gradi di libertà, dove m è il numero di righe e n è il numero di colonne.

$$W \overset{H_0}{\sim} \chi^2_{(m-1)(n-1)}$$

Dimostrazione 21.1: Calcolo dei Gradi di Libertà

La regola generale per i gradi di libertà è:

$$gdl = (\text{numero categorie}) - 1 - (\text{numero parametri stimati})$$

- Numero totale di categorie: $m \cdot n$.
- Per stimare le frequenze attese, abbiamo dovuto stimare le probabilità marginali. Per le m righe, i parametri liberi sono $m-1$ (poiché la somma deve fare 1). Per le n colonne, i parametri liberi sono $n-1$.
- Numero totale di parametri stimati: $(m-1) + (n-1)$.

Quindi: $gdl = (m \cdot n) - 1 - [(m-1) + (n-1)] = mn - 1 - m + 1 - n + 1 = mn - m - n + 1 = (m-1)(n-1)$.

□

Nota 21.4: Note Pratiche

- **Variabili Continue**: Se una o entrambe le variabili non sono discrete con poche categorie, devono essere prima "discretizzate" tramite binning.
- **Condizioni di Validità**: Anche in questo caso, l'approssimazione alla χ^2 è valida se le frequenze attese non sono troppo piccole (la regola pratica è che l'80% delle celle abbia $A_{ij} \geq 5$ e tutte abbiano $A_{ij} \geq 1$).

22 Versioni Esatte dei Test del Chi-Quadro

L'approssimazione della statistica W con una distribuzione χ^2 è un risultato asintotico, valido per grandi campioni. Quando le frequenze attese sono basse e la "rule of thumb" non è rispettata, questa approssimazione non è più affidabile. In questi casi, si ricorre a versioni esatte del test, che calcolano la distribuzione della statistica del test sotto H_0 senza fare affidamento sull'approssimazione.

Questo si può ottenere in due modi:

- **Metodo esatto (esaustione)**: tramite calcolo combinatorio, si enumerano tutti i possibili risultati e si calcola la probabilità esatta di ciascuno. È fattibile solo per problemi di piccole dimensioni.
- **Simulazione Monte Carlo (MCS)**: si genera un gran numero di campioni casuali dalla distribuzione nulla ϕ_0, si calcola la statistica del test per ciascuno e si costruisce una distribuzione empirica, che approssima quella vera.

22.1 Test Esatto di Goodness-of-Fit

Supponiamo di voler eseguire un test di adattamento, ma le frequenze attese sono troppo piccole.

Esempio 22.1: Test esatto per "titolo di studio"

Consideriamo un campione di $n = 20$ studenti e vogliamo testare se la distribuzione del loro titolo di studio si conforma a una ϕ_0 data. Se le frequenze attese calcolate $A_j = n \cdot p_{0,j}$ risultano essere, ad esempio, $[2, 10, 6, 2]$, la regola pratica per l'uso della χ^2 non è rispettata. Non possiamo quindi fidarci del p-value ottenuto da una distribuzione χ^2_3. Per calcolare il p-value corretto, dobbiamo trovare la vera distribuzione della statistica W sotto H_0. Con un metodo di simulazione Monte Carlo, ad esempio, possiamo generare 100.000 campioni da ϕ_0, calcolare W per ciascuno, e usare la distribuzione empirica di questi valori per calcolare il p-value.

22.2 Il Test Esatto di Fisher per Tabelle 2x2

Il test esatto di Fisher è la soluzione combinatoria esatta per il test di indipendenza in una tabella di contingenza 2x2. È particolarmente utile quando le numerosità campionarie sono piccole.

L'Idea Fondamentale Invece di usare la statistica W, il test si concentra su una delle quattro celle della tabella, tipicamente quella in alto a sinistra (O_{11}). L'idea chiave è che, una volta **fissati i totali marginali** della tabella, il valore di una singola cella determina i valori di tutte le altre.

Tabella 6: Tabella di contingenza 2x2 con notazione dei marginali.

	Colonna 1	Colonna 2	Totale Riga
Riga 1	O_{11}	O_{12}	c
Riga 2	O_{21}	O_{22}	d
Totale Colonna	a	b	n

Il problema è analogo a un'estrazione da un'urna. Immaginiamo un'urna con n palline, di cui a nere (colonna 1) e b bianche (colonna 2). Se estraiamo c palline (la prima riga), il numero di palline

nere estratte (k) corrisponde al valore di O_{11}. Questa logica porta direttamente alla distribuzione di probabilità esatta.

Teorema 22.1: Distribuzione per il Test di Fisher

Assumendo l'ipotesi nulla H_0 di indipendenza tra le due variabili e condizionando ai totali marginali, la probabilità di osservare un valore k nella cella $(1,1)$ di una tabella 2x2 segue la **distribuzione Ipergeometrica**:

$$P(O_{11} = k | H_0, \text{marginali}) = \frac{\binom{a}{k}\binom{b}{c-k}}{\binom{n}{c}}$$

Esempio 22.2: Calcolo della distribuzione esatta

Consideriamo una tabella 2x2 con totali marginali: $c = 8$ (riga 1), $a = 21$ (colonna 1), $b = 21$ (colonna 2), e $n = 42$. Sotto H_0, la probabilità per ogni possibile valore di O_{11} (da 0 a 8) è data dalla distribuzione Ipergeometrica. La tabella seguente e il grafico a barre mostrano questa distribuzione.

k	$\mathbf{P}(O_{11} = k)$
0	0.00172
1	0.02069
2	0.09655
3	0.22930
4	0.30348
5	0.22930
6	0.09655
7	0.02069
8	0.00172

Calcolo del p-value Il test è tipicamente bilaterale. Il p-value si calcola sommando le probabilità di tutti i risultati altrettanto o più "estremi" (rari) di quello osservato. Se la distribuzione è simmetrica, come nell'esempio sopra, e osserviamo $O_{11} = 1$, il p-value sarà la somma delle probabilità nelle due code: $P(O_{11} \leq 1) + P(O_{11} \geq 7) \approx 0.02069 + 0.00172 + 0.02069 + 0.00172 = 0.0448$.