TECNICHE MULTIVARIATE

- **descrittive/causali (filosofia della ricerca sociale vs. pratica):** ANOVA, regressione multipla, conjoint analysis
- **associazioni tra variabili:** factor analysis, discriminant analysis*, MDS*
- **classificazione di soggetti/oggetti:** cluster analysis

* La discriminant analysis e l'MDS hanno in verità anche caratteristiche di classificazione; in questo ambito, però, saranno applicate a finalità di posizionamento e quindi di obiettivi di associazione/determinazione di posizioni in uno spazio a due dimensioni

ANALISI DI DATI SPERIMENTALI

Tecnica	Variabile/i dipendente/i (DV)		Variabile/i indipendente/i (IV)	
	Qualitativa/e	Quantitativa/e	Qualitativa/e	Quantitativa/e
t-test	---	1	1 (2 gruppi)	---
One-way ANOVA	---	1	1 (≥ 3 gruppi)	---
ANOVA	---	1	≥ 2	---
MANOVA	---	≥ 2	≥ 1	---
Regressione (o ANCOVA)	---	1	≥ 1	≥ 1
Logit/Probit	1	---	≥ 0	≥ 0

ANOVA

Come nella one-way ANOVA, l'analisi della varianza mira a testare ipotesi sulle differenze tra medie di **una variabile dipendente quantitativa** nei gruppi formati da 2 o più **variabili indipendenti qualitative** e dalle loro **interazioni**

L'ANOVA è applicata su dati sperimentali per analizzare (nel caso comune di 2 variabili indipendenti) tramite il test F:

- **l'effetto diretto** della prima variabile indipendente: le medie della variabile dipendente sono diverse nei gruppi formati dalla prima variabile indipendente
- **l'effetto diretto** della seconda variabile indipendente: le medie della variabile dipendente sono diverse nei gruppi formati dalla seconda variabile indipendente
- **l'effetto di interazione** tra le due variabili indipendenti: le medie della variabile dipendente sono diverse nei gruppi formati dall'interazione delle due variabili indipendenti – l'effetto di una delle variabili indipendenti dipende dal valore assunto dall'altra variabile indipendente

ANOVA: UN ESEMPIO

Variabile dipendente: atteggiamento verso un ad

Variabili indipendenti: Endorser (quality vs. image) e Involvement (low vs. high)

Tests of Between-Subjects Effects

Dependent Variable: Atteggiamento

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	73.400[a]	3	24.467	16.251	.000
Intercept	608.400	1	608.400	404.103	.000
involvement	.400	1	.400	.266	.609
endorser	.100	1	.100	.066	.798
involvement * endorser	72.900	1	72.900	48.421	.000
Error	54.200	36	1.506		
Total	736.000	40			
Corrected Total	127.600	39			

a. R Squared = .575 (Adjusted R Squared = .540)

I *p*-value associati ai test F degli effetti diretti di endorser e involvement sono entrambi > 0,05; quindi tali effetti sono non significativi

Il *p*-value associato al test F dell'effetto d'interazione è < 0,05; quindi tale effetto è significativo

ANOVA: UN ESEMPIO

Infatti, le medie di atteggiamento nei gruppi formati dalle IV sono tendenzialmente uguali (no effetti diretti)

Dall'analisi dei contrasti, invece, emerge che le medie dei gruppi Endorser cambiano al variare del gruppo di involvement

Pairwise Comparisons

Dependent Variable: Atteggiamento

involvement	(I) endorser	(J) endorser	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a] Lower Bound	Upper Bound
low	image	quality	2.600*	.549	.000	1.487	3.713
	quality	image	-2.600*	.549	.000	-3.713	-1.487
high	image	quality	-2.800*	.549	.000	-3.913	-1.687
	quality	image	2.800*	.549	.000	1.687	3.913

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

2. involvement

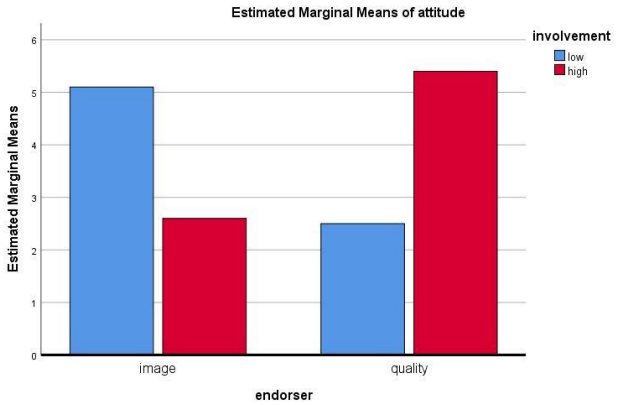
Dependent Variable: Atteggiamento

involvement	Mean	Std. Error	95% Confidence Interval Lower Bound	Upper Bound
low	3.800	.274	3.244	4.356
high	4.000	.274	3.444	4.556

3. endorser

Dependent Variable: Atteggiamento

endorser	Mean	Std. Error	95% Confidence Interval Lower Bound	Upper Bound
image	3.850	.274	3.294	4.406
quality	3.950	.274	3.394	4.506



REGRESSIONE LINEARE

L'analisi di regressione lineare mira a testare ipotesi sugli effetti di variabili indipendenti quantitative o qualitative su una variabile dipendente quantitativa

Modello di regressione lineare multipla con dati temporali

$$y_t = a + b_1 x_{1t} + b_2 x_{2t} + \ldots + b_k x_{kt} + e_t$$

In cui l'indice t indica le osservazioni temporali

Modello di regressione lineare multipla con dati cross-sectional

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + e_i$$

In cui l'indice i indica le osservazioni cross-sectional (es., consumatori)

REGRESSIONE LINEARE

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + e_i$$

Variabile dipendente

Intercetta

Variabili indipendenti

Errore stocastico

Parametri di regressione

IL MODELLO DI REGRESSIONE LINEARE

$$e_i = y_i - \hat{y}_i$$

Errore stocastico per il soggetto "i": la discrepanza tra la variabile dipendente osservata e quella prevista dal modello

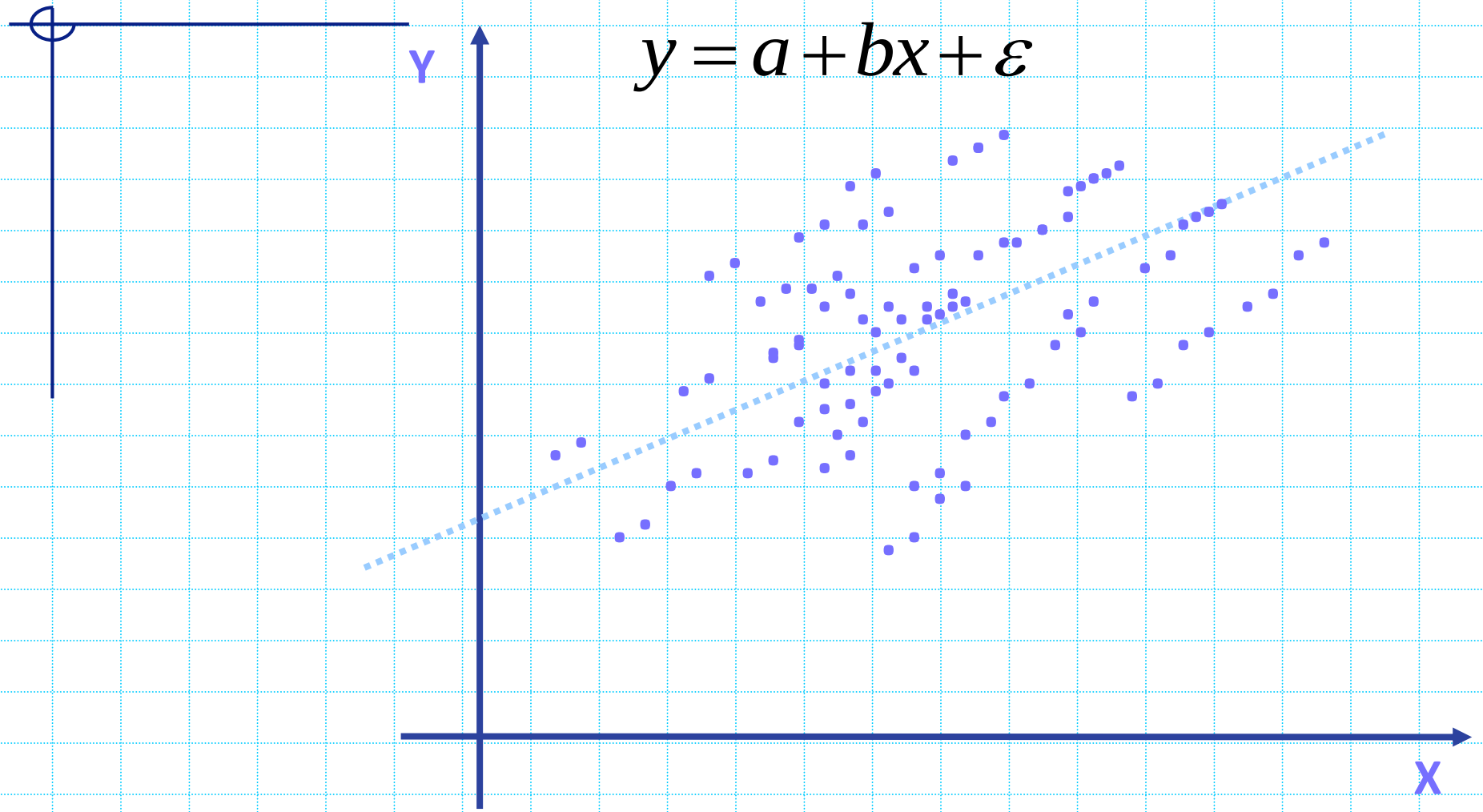
Variabile dipendente: valore osservato per il soggetto "i"

Variabile dipendente: valore previsto dal modello per il soggetto "i"

Il metodo OLS assume che l'errore e_i si distribuisca secondo la distribuzione normale, che abbia media = 0 e varianza costante

IL METODO DEI MINIMI QUADRATI (OLS)

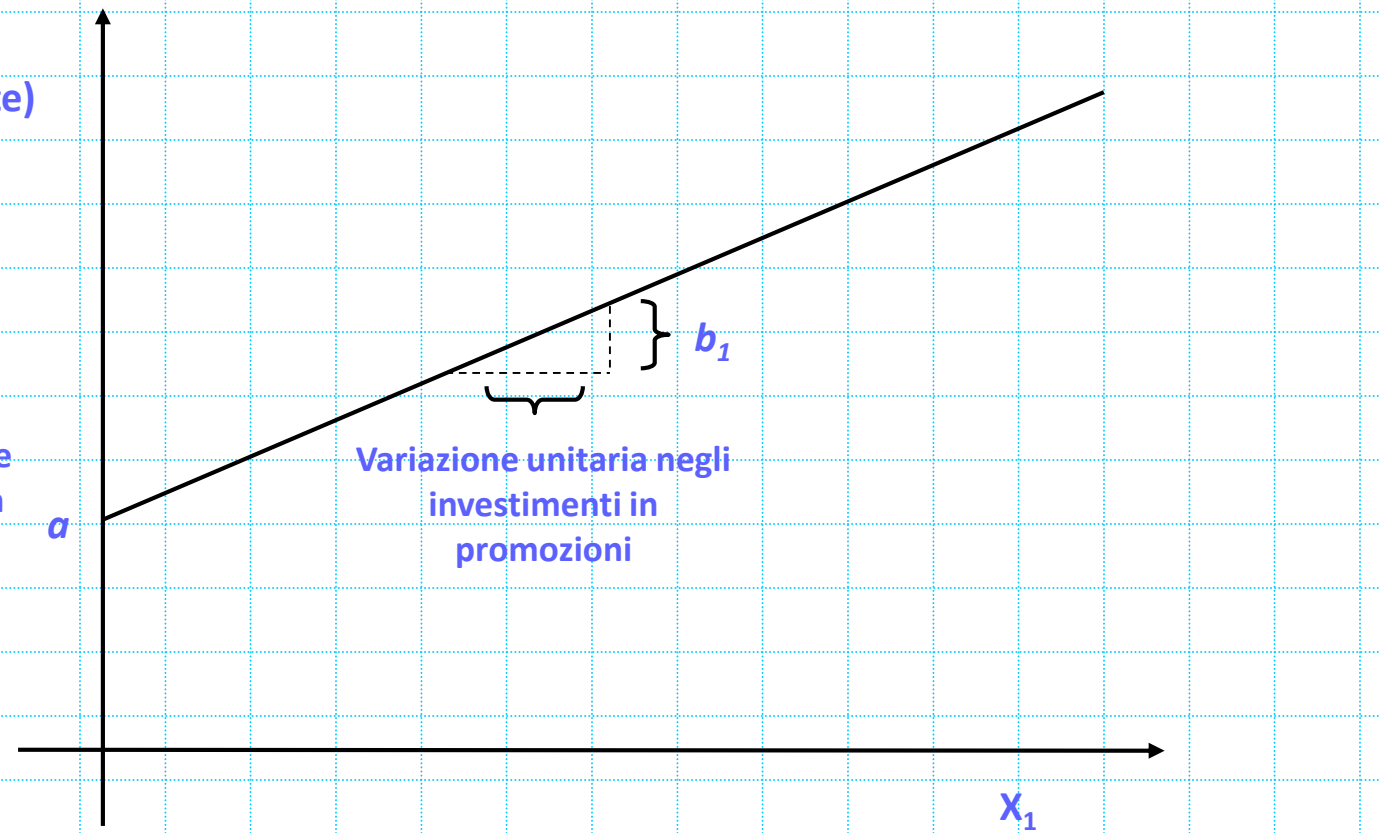
$$y = a + bx + \varepsilon$$



IL METODO DEI MINIMI QUADRATI (OLS)

Graficamente:

$$y_t = a + b_1 x_{1t} + e_t$$



IL MODELLO DI REGRESSIONE LINEARE

Interpretazione

L'**intercetta** rappresenta il valore medio della variabile dipendente quando tutte le variabili indipendenti sono uguali a zero; in alcuni casi, non ha valore interpretativo e può essere ignorata

I **parametri di regressione** indicano il segno e l'intensità dell'effetto medio di una variabile indipendente sulla variabile dipendente, assumendo che le altre variabili indipendenti rimangano costanti: i parametri indicano di quanto varia la variabile dipendente al variare di una unità della variabile indipendente

Quando si trattano dati oggettivi si possono interpretare i parametri di regressione **non standardizzati** (b), che preservano l'informazione relativa alle unità di misura

Quando si trattano dati percettivi è preferibile interpretare i parametri di regressione **standardizzati** (β); i parametri standardizzati, diversamente da quelli non standardizzati, permettono il confronto tra parametri, perché eliminano l'informazione relativa alle unità di misura

IL MODELLO DI REGRESSIONE LINEARE

La stima del modello avviene tramite il metodo dei minimi quadrati (OLS)

Per ogni variabile indipendente si ottiene:

- il parametro non standardizzato
- l'errore standard del parametro
- il t-test (parametro/errore standard); H_0: b = 0
- il *p*-value

Un parametro di regressione può essere:

- significativo (cioè diverso da zero, il risultato è generalizzabile), se il *p*-value è minore di 0,05
- non significativo (cioè non diverso da zero, il risultato non è generalizzabile), se il *p*-value è maggiore di 0,05

IL MODELLO DI REGRESSIONE LINEARE

La bontà del modello è espressa da

- il test **F**, che è un test generale sulla significatività del modello; l'ipotesi nulla è che tutti i parametri di regressione siano pari a zero; ad esso è associato un *p*-value (se $p < 0,05$ si rigetta H_0 e si conclude che almeno un effetto è significativo; se $p > 0,05$ si accetta H_0 e si conclude che il modello nella sua interezza non è significativo)

- R^2, che esprime la parte di varianza della variabile dipendente che è spiegata dal modello (la parte non spiegata riguarda variabili esterne al modello ed è indicata dall'errore *e*); oscilla tra 0 (modello pessimo) e 1 (modello perfetto, cioè un non-modello…); quando si trattano dati oggettivi o storici ci si aspettano valori alti (> 0,60), con dati cross-sectional (percettivi) sono accettabili soglie più basse (> 0,20 o anche > 0,10)

VARIABILI DUMMY

Il modello di regressione lineare prevede che la Y e le X siano variabili numeriche. È possibile però includere variabili indipendenti qualitative nella forma di **variabili dummy**

Le variabili indipendenti sono variabili binarie (0/1); il parametro di regressione indica di quanto varia la variabile dipendente quando la variabile dummy è uguale a uno e non a zero – esempio: genere (femmina = 1; maschio = 0), marca Barilla (sì = 1; no = 0)

ASSUNZIONI

Affinché il metodo degli OLS riporti stime affidabili ed efficienti, è necessario che vengano rispettate alcune assunzioni

- **Assenza di multicollinearità**: è necessario che le variabili indipendenti non siamo fortemente correlate tra di loro (altrimenti i parametri e gli errori standard dei parametri sarebbero distorti) – è necessario verificare le correlazioni tra variabili ed eliminare eventuali variabili ridondanti oppure accorparle
- **Indipendenza** tra l'errore e le variabili indipendenti
- **Assenza di endogeneità (vs. esogeneità)**: la variabile dipendente non determina le variabili indipendenti (simultaneità), non ci sono variabili rilevanti omesse dal modello, assenza di errore di misurazione delle variabili
- **Omoschedasticità**: la varianza degli errori è costante (test di White ed errori robusti all'eteroschedasticità)
- **Assenza di autocorrelazione** (rilevante solo per dati storici): è necessario che gli errori stocastici di periodi successivi non siano correlati tra loro – si usa il test di Durbin-Watson o si stima il coefficiente di autocorrelazione (rho)

PRINCIPALI VIOLAZIONI

Violazione	Conseguenze	Diagnostici	Possibili soluzioni
Multicollinearità perfetta (combinazioni lineari perfette tra le IV)	Il modello non può essere stimato	Correlazioni tra le IV	• Eliminare IV ridondanti • Se possibile, aggregare le IV correlate (factor analysis) • Aumentare la N • Non c'è bisogno di soluzione se gli obiettivi sono puramente predittivi
Multicollinearità imperfetta (correlazioni sostanziali tra le IV)	Parametri e s.e. distorti		
Endogeneità (reciprocità, errore di misurazione, IV rilevanti omesse)	Parametri distorti R^2 ridotti	Teoria Hausman test	• Teoria robusta • Disegni sperimentali • Cambiare metodo di stima (TSLS, GMM) • Modelli di equazioni strutturali (stima simultanea e controllo dell'errore di misurazione)
Eteroschedasticità (varianza dell'errore non costante)	s.e. , test e conclusioni inferenziali non corrette	White test	• Ri-specificare il modello (cluster?) • Cambiare metodo di stima (GLS) • s.e. robusti all'eteroschedasticità
Autocorrelazione (errori serialmente correlati)	s.e. , test e conclusioni inferenziali non corrette	Durbin-Watson Coefficiente di autocorrelazione	• Ri-specificare il modello (variabili temporali) • Cambiare metodo di stima (GLS) • s.e. robusti all'autocorrelazione

UN ESEMPIO

$$y_t = a + b_1 x_{1t} + b_2 x_{2t} + b_3 x_{3t} + b_4 x_{4t} + e_t$$

in cui

y_t = vendite della marca Y al tempo t

x_1 = prezzo medio della marca Y al tempo t

x_2 = store promotions della marca Y al tempo t

x_3 = advertising della marca Y al tempo t

x_4 = dummy holiday tempo t

e_t = errore stocastico al tempo t

Dati: 95 settimane

UN ESEMPIO

Risultati (OLS): $R^2 = 0.62$; $F = 36.04$, $p = 0{,}000$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4564.584	327.792		13.925	.000
	prezzo	-1725.136	164.849	-.707	-10.465	.000
	store_promo	2.239	1.030	.150	2.174	.032
	advert	1.412	.626	.151	2.256	.026
	holiday	-5.051	44.087	-.008	-.115	.909

a. Dependent Variable: sales

UN ESEMPIO

Interpretazione

Il test F è significativo ($p < 0.000$), quindi, globalmente, il modello è significativo; l'R^2 è soddisfacente

L'intercetta è significativa, ma in questo caso non è interessante (il prezzo non può essere uguale a zero)

Il parametro per il prezzo medio è significativo ($p < 0.000$); il valore –1725.14 indica che all'aumentare di un'unità di prezzo medio le vendite diminuiscono di 1725.14 unità, posti store_promo, advertising e holiday costanti

Il parametro per store_promo è significativo ($p < 0.05$); il valore 2.24 indica che all'aumentare di un'unità di store_promo le vendite aumentano di 2.24 unità, posti prezzo medio, advertising e holiday costanti

Il parametro per advertising è significativo ($p < 0.05$); il valore 1.41 indica che all'aumentare di un'unità di advertising le vendite aumentano di 1.41 unità, posti prezzo medio, store_promo e holiday costanti

L'effetto di holiday non è significativo ($p > 0.10$)

LA CONJOINT ANALYSIS

La **conjoint analysis** è una forma particolare di analisi di regressione in cui

- la **variabile dipendente** può essere quantitativa (e.g., valutazione da 1 a 9 di vari prodotti, stima con OLS) o qualitativa (e.g., scelta tra set di 3 prodotti alternativi, stima modello Logit, non discusso in questo corso)
- le **variabili indipendenti** possono essere quantitative o qualitative; le variabili indipendenti vengono **manipolate** nell'ambito di un **disegno ortogonale**, che genera combinazioni alternative dei livelli delle variabili indipendenti

LA CONJOINT ANALYSIS

Rispetto alle tecniche univariate per misurare il valore per il cliente, basate su attributi (e.g., indice di Fishbein), la conjoint analysis risulta più **realistica**, perché ai rispondenti viene richiesta una valutazione delle alternative in senso **globale**

Attraverso la stima del modello sottostante, è possibile inferire i **valori di importanza** di ogni variabile indipendente (o fattore) e il **valore di utilità** (parametro di regressione) di ogni livello delle variabili indipendenti

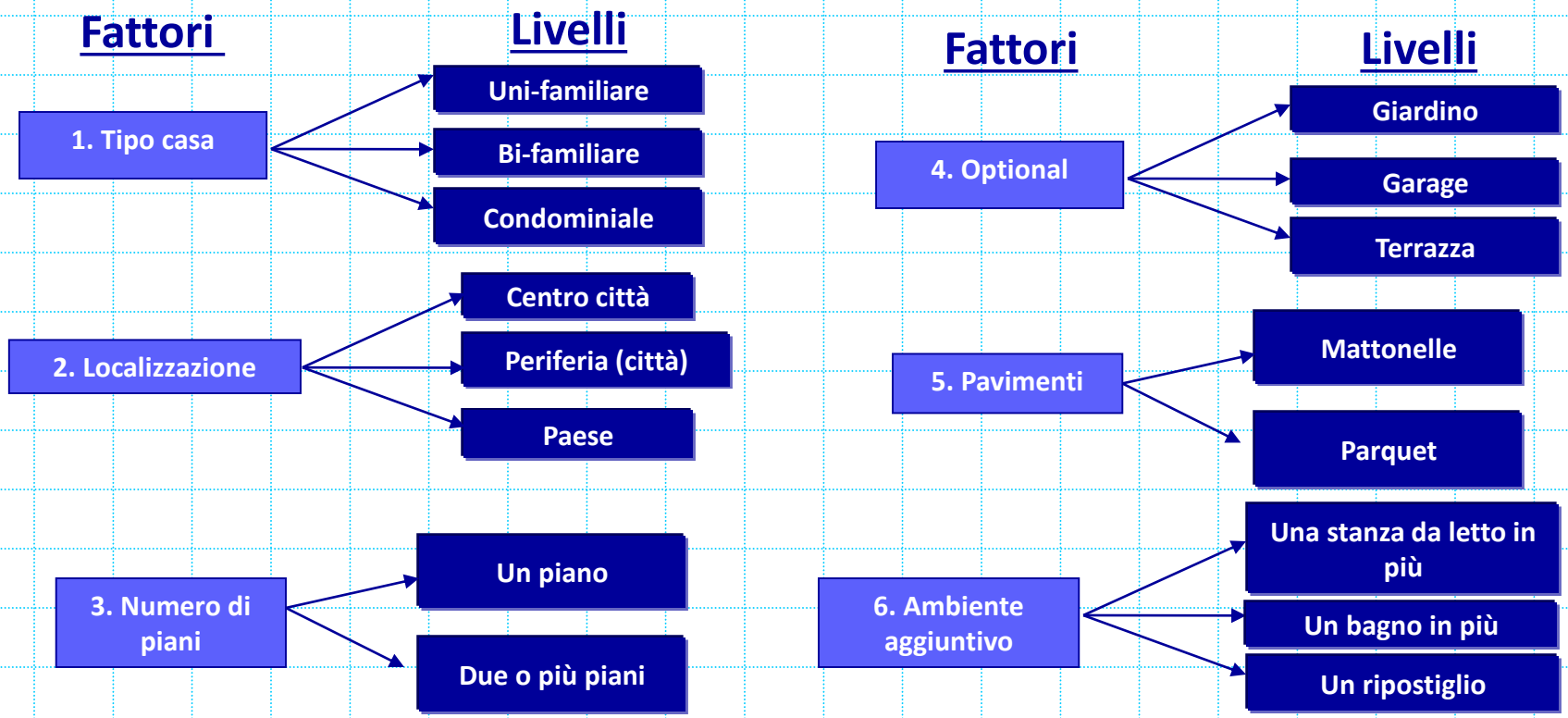
Date queste caratteristiche, la conjoint analysis è particolarmente utile per misurare il valore per il cliente, testare nuovi prodotti, definire il prodotto ideale, simulare quote di mercato, definire il prezzo.

LA CONJOINT ANALYSIS

Nel manipolare i fattori, è preferibile

- che i fattori abbiano (tendenzialmente) lo stesso numero di livelli
- che i livelli siano effettivamente in competizione tra loro (evitare livelli palesemente migliori o peggiori di altri): trade-off tra i livelli
- controllare che le combinazioni ottenute (rappresentate in forma di card) siano sostanzialmente diverse tra di loro (livelli diversi su almeno due fattori)

UN ESEMPIO


Fattori
Livelli
1. Tipo casa
Uni-familiare
Bi-familiare
Condominiale
2. Localizzazione
Centro città
Periferia (città)
Paese
3. Numero di piani
Un piano
Due o più piani
Fattori
Livelli
4. Optional
Giardino
Garage
Terrazza
5. Pavimenti
Mattonelle
Parquet
6. Ambiente aggiuntivo
Una stanza da letto in più
Un bagno in più
Un ripostiglio

UN ESEMPIO

Disegno ortogonale

card	Tipo casa	Localizzazione	Livelli	Optional	Pavimenti	Ambiente aggiuntivo
1	Bifamiliare	Paese	Un livello	Terrazzo	Mattonelle	Un bagno in più
2	Unifamiliare	Periferia (città)	Un livello	Giardino	Mattonelle	Un ripostiglio
3	Unifamiliare	Periferia (città)	Un livello	Garage	Parquet	Un bagno in più
4	Unifamiliare	Centro (città)	Un livello	Terrazzo	Parquet	Una camera da letto in più
5	Bifamiliare	Centro (città)	Due o più livelli	Garage	Parquet	Un ripostiglio
6	Unifamiliare	Paese	Due o più livelli	Garage	Mattonelle	Un bagno in più
7	Condominiale	Periferia (città)	Due o più livelli	Terrazzo	Parquet	Un bagno in più
8	Bifamiliare	Periferia (città)	Due o più livelli	Terrazzo	Mattonelle	Una camera da letto in più
9	Condominiale	Centro (città)	Un livello	Garage	Mattonelle	Una camera da letto in più
10	Condominiale	Paese	Un livello	Terrazzo	Parquet	Un ripostiglio
11	Unifamiliare	Periferia (città)	Due o più livelli	Garage	Mattonelle	Un ripostiglio
12	Unifamiliare	Centro (città)	Due o più livelli	Terrazzo	Parquet	Un bagno in più
13	Condominiale	Centro (città)	Due o più livelli	Giardino	Mattonelle	Un bagno in più
14	Bifamiliare	Centro (città)	Un livello	Giardino	Parquet	Un bagno in più
15	Unifamiliare	Paese	Due o più livelli	Giardino	Parquet	Una camera da letto in più
16	Unifamiliare	Centro (città)	Un livello	Terrazzo	Mattonelle	Un bagno in più

UN ESEMPIO

Utilities

		Utility Estimate	Std. Error
tipocasa	Unifamiliare	.848	.068
	Bifamiliare	.133	.078
	Condominiale	-.981	.078
dove	Centro (città)	.021	.071
	Periferia (città)	.050	.076
	Paese	-.071	.077
livelli	Un livello	.010	.051
	Due o più livelli	-.010	.051
optional	Terrazzo	-.236	.071
	Garage	-.050	.076
	Giardino	.287	.077
paviment	Mattonelle	-.020	.051
	Parquet	.020	.051
stanza	Un bagno in più	.092	.068
	Una camera da letto in più	.116	.079
	Un ripostiglio	-.208	.083
(Constant)		5.934	.058

Importance Values

tipocasa	32.789
dove	21.768
livelli	7.692
optional	14.469
paviment	8.187
stanza	15.094

Averaged Importance Score

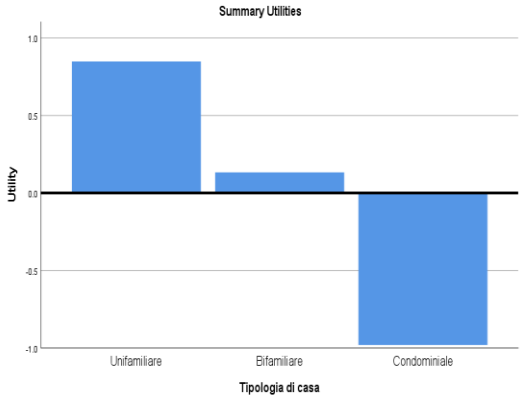
Correlations[a]

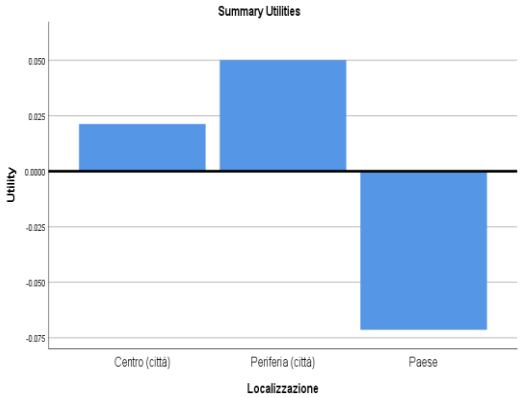
	Value	Sig.
Pearson's R	.990	.000
Kendall's tau	.845	.000

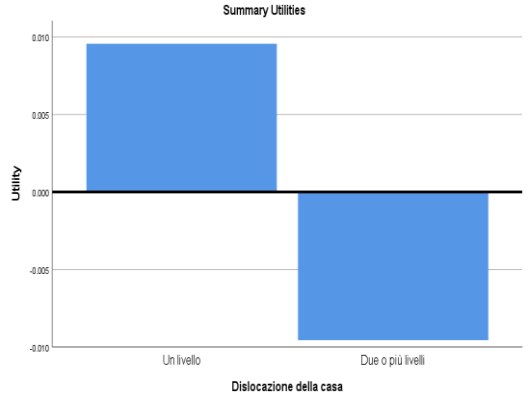
a. Correlations between observed and estimated preferences

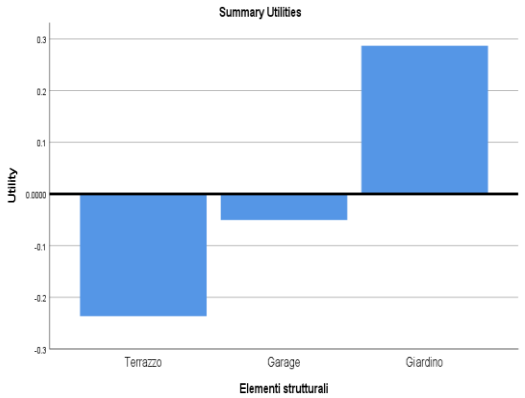
UN ESEMPIO

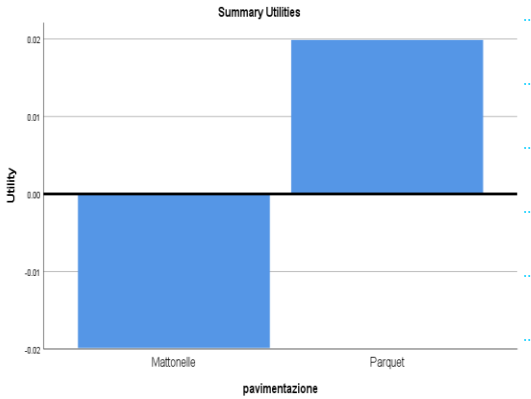
Livelli dei fattori: utilità (parametri di regressione)

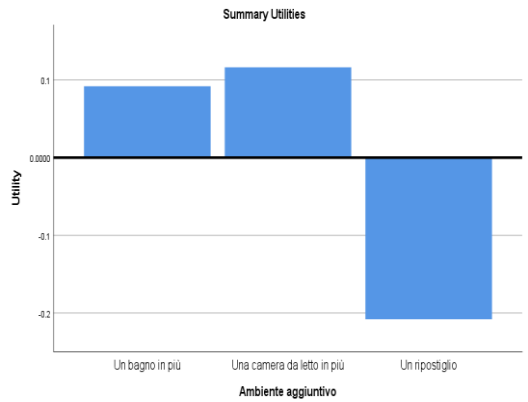






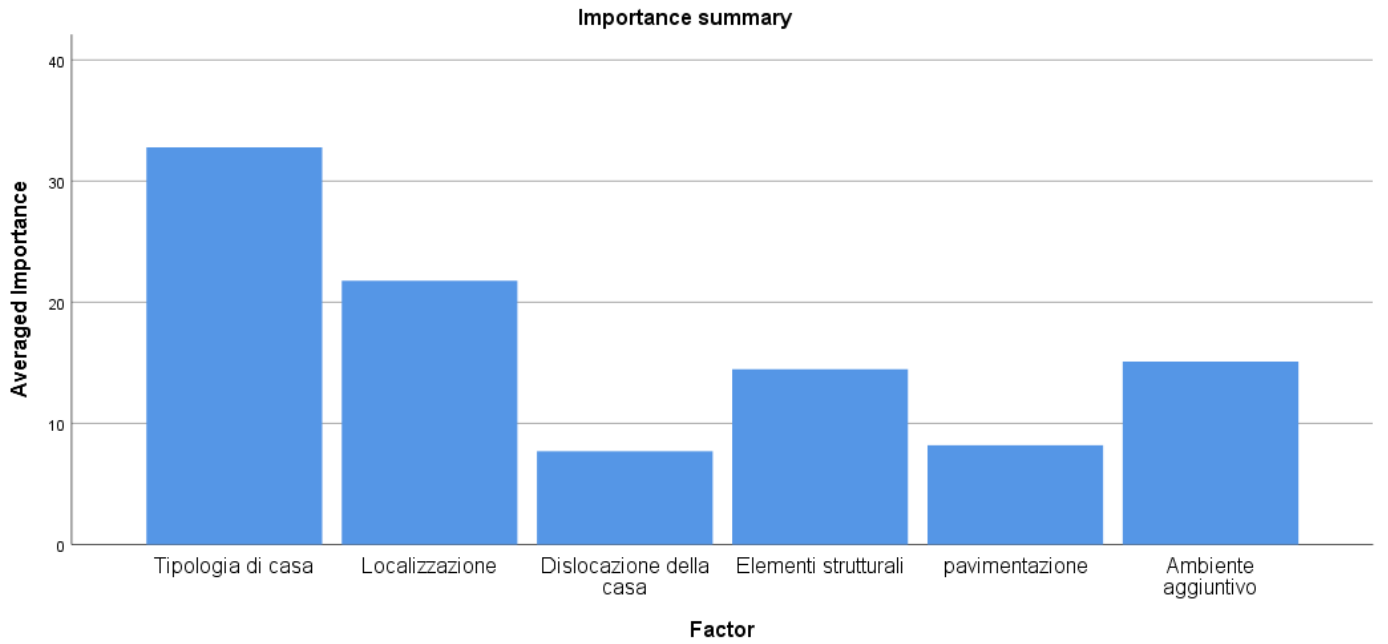






UN ESEMPIO

Importanza media percentuale dei fattori (calcolata sulla base della differenza tra il parametro più alto e quello più basso per ogni fattore)



ULTERIORI APPLICAZIONI

Sulla base della manipolazione e dell'output della conjoint analysis, è possibile

- inserire il prezzo come fattore (ipotesi: relazione negativa con valutazioni card) nell'ambito di analisi di pricing

$$\text{Valore Monetario} = \frac{\text{Pmax - Pmin}}{\text{Umax - Umin}}$$

$$\text{Utilità Unitaria}$$

- effettuare delle simulazioni inserendo nel disegno ortogonale delle card non testate, ma "interessanti" (e.g., card ideale, economica, realistica, ecc.); in output: valutazione media simulata
- stimare con i metodi first-choice (ranking delle card), BTL (quota utilità/sommatoria utilità) e Logit (come BTL, ma con trasformazione esponenziale), le probabilità di scelta delle card (o quote di mercato simulate)