UNIVERSITÀ DEGLI STUDI DEL PIEMONTE ORIENTALE

DIPARTIMENTO DI STUDI PER L'ECONOMIA E L'IMPRESA

Corso di Laurea Magistrale in Management e Finanza

Curriculum Marketing and Operations Management

Tesi di Laurea Magistrale

RICERCA DELLE DETERMINANTI DEL PREZZO DELLE CAMERE D'HOTEL. UN CONFRONTO TRA MODELLI AD ALTA DIMENSIONALITA'.

RELATORE:

Prof. Aldo GOIA

CORRELATORE:

Prof. Graziano ABRATE

CANDIDATO:

Ilaria ROSSINI

Matricola: 20030255

ANNO ACCADEMICO 2022/2023

INDICE

INTRODUZIONE

Il settore dell'ospitalità riveste un ruolo importante nell'economia di molti Paesi nonostante sia un settore particolare poiché deve far fronte alle continue variazioni della domanda; per questo motivo, le strutture ricettive devono adattare continuamente i prezzi delle proprie camere, le quali però rappresentano allo stesso tempo un vincolo di capacità fissa. Di conseguenza risulta fondamentale per le strutture alberghiere fissare il giusto prezzo per ogni transazione, in modo da ottimizzare i ricavi. In questo contesto si sviluppa una nuova strategia di prezzo: il Revenue Management che, attraverso l'applicazione di prezzi dinamici, consente di sfruttare meglio l'eterogeneità della domanda. Inoltre, i prezzi delle camere degli hotel dipendono da numerosi fattori sia interni che esterni, come ad esempio: il numero di stelle, la presenza di una piscina, la distanza dalla stazione, il momento della prenotazione, ecc.

Con lo sviluppo della tecnologia è cambiato profondamente il modo di prenotare un soggiorno, infatti, oggi è sempre maggiore la percentuale delle prenotazioni online: sia attraverso i siti web delle strutture ricettive sia attraverso le piattaforme delle agenzie di viaggio online che permettono agli utenti di accedere a numerose informazioni e di confrontare direttamente e con trasparenza i prezzi applicati dai diversi hotel ad uno stesso prodotto o servizio, in questo modo gli utenti riescono a compiere una scelta più ponderata. Durante il processo di ricerca, gli utenti possono essere anche influenzati dalle recensioni, sia numeriche che con commenti testuali, lasciate da altri consumatori su queste piattaforme.

L'obiettivo di questo elaborato consiste nell'andare a studiare, attraverso diversi approcci, i prezzi del settore alberghiero per capire da quali fattori, presenti sulla piattaforma Booking.com, sono influenzati maggiormente anche in base al momento della prenotazione. Il dataset analizzato in questo lavoro è un subset del database studiato nell'articolo "The impact of dynamic price variability on revenue maximization" (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019). Nel dataset in esame si analizzano i prezzi ed altri fattori di 255 hotel della città di Londra per tutto il mese di aprile del 2016.

Nel primo capitolo dell'elaborato si analizza la letteratura esistente relativa alle strategie di prezzo applicate nel settore alberghiero, con un approfondimento particolare sul Revenue Management e sul Dynamic Pricing. Successivamente, attraverso l'analisi di studi precedenti, si prosegue con la definizione delle caratteristiche del settore alberghiero e dei principali fattori

che impattano maggiormente sul prezzo delle camere; si studiano inoltre le conseguenze apportate dallo sviluppo tecnologico nel panorama delle prenotazioni.

Il secondo capitolo fornisce una descrizione approfondita del dataset utilizzato, sottolineando le caratteristiche principali di ogni variabile osservata; nella parte finale del capitolo, invece, vengono identificate le variabili correlate ossia quelle variabili che sono una la copia dell'altra e che quindi apportano sulla variabile di risposta le stesse informazioni. Si precisa che le variabili osservate sono otto: camera, numero di stelle, valutazione, numero di preferiti, numero di revisioni, pagina, posizione e prezzo, le quali vengono osservate in otto istanti temporali differenti, ipotizzando cioè di prenotare 60, 45, 30, 20, 10, 4 giorni prima, il giorno precedente e il giorno stesso del soggiorno.

Nel terzo capitolo invece si stimano ed analizzano otto diversi modelli di regressione lineare, uno per ogni istante temporale osservato, con lo scopo di identificare quali sono le variabili esplicative significative e come quest'ultime impattano sulla variabile di risposta, ossia il relativo prezzo. Successivamente attraverso dei grafici a dispersione, si analizza come variano i coefficienti delle variabili significative nell'orizzonte temporale osservato; in questo modo è possibile capire, in base al momento della prenotazione, quali fattori incidono maggiormente sul prezzo delle camere degli hotel analizzati.

Nel quarto e quinto capitolo, si realizza un unico modello avente come variabile di risposta, il prezzo finale, ossia il prezzo del giorno del soggiorno e che comprenda tutte le variabili osservate nell'intero orizzonte temporale studiato. La realizzazione di un unico modello, però, non è immediata poiché, sia tra le componenti autoregressive sia tra le altre variabili in esame, vi è elevata correlazione che, se non controllata, porta a risultati non precisi; è necessario dunque utilizzare delle tecniche che permettano di gestire questa correlazione. In particolare, sono stati utilizzati i seguenti metodi: analisi delle componenti principali nel capitolo quattro, e metodo di compressione nel capitolo cinque.

Nel quarto capitolo, infatti, si comincia con la stima delle componenti principali (PC), le quali sono combinazioni lineari delle variabili originarie ponderate con pesi differenti; si procede poi con la stima del modello di regressione lineare che mette in relazione il prezzo del giorno del soggiorno con le PC ottenute e si confrontano i risultati ottenuti con quelli del capitolo precedente.

Nel quinto capitolo si effettua invece una selezione delle variabili significanti attraverso la tecnica lasso che permette di controllare il problema dell'elevata correlazione tra le variabili; grazie a questo metodo, è possibile procedere con un'analisi più generale che considera il dataset nella sua interezza e non suddiviso per istanti temporali osservati. Terminata l'analisi del modello ottenuto, si procede con un confronto dei risultati con quelli ottenuti nei capitoli precedenti, al fine di identificare quale sia l'approccio che spiega meglio i dati in esame.

Il sesto ed ultimo capitolo presenta un'analisi differente in quanto come predittori non si considerano le variabili originarie, ma si studiano piuttosto le variazioni delle variabili tra due istati successivi, in questo modo si eliminano le eventuali correlazioni seriali legate al fatto che i predittori sono osservati per più instanti nel tempo. Dopo aver stimato un modello di regressione lineare, si procede dunque con la stima di un altro modello comprendente le stesse variabili, ma utilizzando la tecnica lasso per ridurre le ulteriori correlazioni tra le variabili.si confrontano infine i risultati ottenuti dai vari modelli per determinare quale modello funzioni meglio.

CAPITOLO 1: STRATEGIE DI PREZZO DEL SETTORE ALBERGHIERO

Il tempo è uno dei criteri che può guidare una differenziazione di prezzo; ques'ultimo infatti varia nel tempo. Per capire il motivo del cambiamento del prezzo nel tempo, bisogna considerare i tre elementi chiave utili a definire una strategia di prezzo: domanda di mercato, costi di produzione e contesto competitivo; ognuno di questi tre elementi è soggetto a cambiamenti (Simon, Zatta, & Fassnacht, 2013).

Con riferimento alla domanda di mercato si intende la disponibilità a pagare dei consumatori, la quale varia in relazione a diversi fattori: si ha infatti una diversa disponibilità a pagare in base al grado di innovazione dei prodotti oppure in base al momento in cui avviene l'acquisto, per esempio se si vuole magiare fuori a pranzo o a cena, la disponibilità a pagare varia anche in base alle esigenze del consumatore, ad esempio si è disposti a pagare di più per l'utilizzo dell'aria condizionata in un giorno di caldo eccezionale.

Per quanto riguarda i costi, invece, anche quest'ultimi non rimangono costanti, ma subiscono delle variazioni: basti pensare ai cambiamenti tecnologici nella funzione di produzione. Anche i prezzi stessi delle materie prime possono cambiare e, di conseguenza, si avrà un aumento del prezzo del prodotto finale.

Si possono avere anche delle variazioni nel contesto competitivo come: l'ingresso di un nuovo concorrente sul mercato. Poiché tutti e tre gli elementi che definiscono il prezzo variano nel tempo allora anche le decisioni di prezzo variano in funzione del criterio temporale.

Quando si parla di decisioni di prezzo, però, bisogna fare una distinzione tra strategia e tattica. Con strategia si intende un piano di azione di medio-lungo termine al fine di raggiungere un determinato obiettivo, questo significa che le variazioni di prezzo intertemporali possono essere pianificate in anticipo sulla base di ciò che l'impresa può prevedere. Quando si parla di decisioni di prezzo strategiche, si fa riferimento alle traiettorie di prezzo rispetto al ciclo di vita del prodotto e al Peak-load pricing, ossia l'investimento ottimale in capacità produttiva (Abrate, Pricing a creazione di valore. Strumenti e applicazioni manageriali., 2020).

Con tattica, invece, si intende la predisposizione di una singola azione con un obiettivo di breve termine che deve tener conto di una serie di vincoli contingenti; le opzioni tattiche sono normalmente vincolate alle decisioni strategiche. Nel breve termine, si possono adattare le

decisioni di prezzo rispetto a cambiamenti imprevisti delle condizioni di mercato. Quando si parla di decisioni di prezzo tattiche, si fa riferimento al Revenue Management e al Dynamic Princing; questi modelli servono per ottimizzare le vendite, tenendo in considerazione l'evoluzione dinamica delle condizioni di mercato (Abrate, Pricing a creazione di valore. Strumenti e applicazioni manageriali., 2020).

In questo capitolo, si vuole analizzare la letteratura esistente riguardante le strategie di prezzo applicate nel settore alberghiero, in particolare Revenue Management e Dynamic Pricing, per poi analizzare gli studi precedenti sulle caratteristiche e sui fattori tipici del settore alberghiero.

1.1 LA DEFINIZIONE DI REVENUE MANAGEMENT

Il Revenue Management, o gestione dei ricavi, consiste nell'ottimizzazione dei ricavi, dei profitti e del valore del cliente, ovvero è indicato come il processo di massimizzazione delle entrate da ogni transazione commerciale, attraverso la determinazione dinamica dei prezzi e l'allocazione efficiente delle scorte disponibili alla domanda prevista. Si tratta quindi del processo di assegnazione del giusto servizio al cliente specifico, al prezzo giusto e nel momento esatto, in modo da massimizzare i ricavi. È chiaro tuttavia che, se da un lato bisogna guardare alla massimizzazione dei ricavi, dall'altro si deve guardare alla costruzione di una relazione a lungo termine con i clienti, al fine di garantire una continuità nell'attività commerciale (Nair, 2019).

Il Revenue Management viene applicato principalmente in settori peculiari, come il turismo, i viaggi e le strutture ricettive, in quanto devono costantemente confrontarsi con le diverse elasticità della domanda, con i vincoli di capacità spesso fissi e con le caratteristiche di intangibilità e deperibilità del prodotto. Infatti, se il servizio non viene venduto, allora vuol dire che è perso. In questo scenario di incertezza, il ruolo del pricing è quello di massimizzare i profitti dei venditori catturando le valutazioni dei prodotti da parte dei consumatori (Abrate & Viglia, Strategic and tactical price decisions in hotel revenue management, 2016).

Una delle caratteristiche chiave di qualsiasi attività nel settore dell'ospitalità è la sua abilità di massimizzare i ricavi e i profitti durante i periodi di fluttuazione della domanda. Per questo motivo, il Revenue Management è diventato sempre più importante da quando è stato introdotto negli anni '80 nel settore turistico (Binesh, Belarmino, & Raab, 2021).

1.1.1 LA STORIA DEL REVENUE MANAGEMENT

Il Revenue Management nasce nell'industria aerea dove viene implementato per la prima volta dalla compagnia American Airlines. Nel 1978, ci fu la deregolamentazione del settore aereo e questo portò molte compagnie "low-cost" a diventare nuovi competitori, creando così una forte pressione concorrenziale. Per questo motivo, American Airlines iniziò a studiare un metodo per reagire alla concorrenza sui prezzi, partendo da due presupposti (Talluri & Van Ryzin, 2004):

- una guerra di prezzo alle compagnie "low-cost" era stata una strategia impraticabile, a causa dei costi medi troppo elevati;

- sulla capacità in eccesso, il costo marginale era sostanzialmente nullo e, quindi, era possibile competere sui prezzi attuando delle promozioni.

Bisognava però predisporre un limite sulla quantità di posti venduti con la tariffa promozionale su ciascun volo ed era anche necessario introdurre delle restrizioni sui biglietti venduti in promozione, affinché scoraggiassero i clienti che erano già disposti a pagare il prezzo pieno (come: l'obbligo di acquisto trenta giorni prima della partenza, la non rimborsabilità del biglietto, ecc.).

Inizialmente, quindi, si era deciso di impostare una porzione fissa di posti venduti in promozione per ogni volo, ma questo sistema risultò imperfetto, in quanto non tutti i voli avevano le stesse caratteristiche. Per questo venne introdotto un sofisticato sistema di ottimizzazione dinamica, in grado di controllare i posti ancora a disposizione su ciascun volo e fu chiamato Dynamic Inventory Allocation and Maintenance Optimizer system (DINAMO). Grazie a questo sistema, la compagnia American Airlines era in grado di rispondere non appena un rivale pubblicizzava una promozione speciale, confidando sulla capacità del sistema "DINAMO" di controllare e limitare automaticamente il numero di posti venduti in offerta su ciascun volo, cambiando così le dinamiche della concorrenza nel mercato aereo (Talluri & Van Ryzin, 2004).

1.1.2 LE TIPOLOGIE DI REVENUE MANAGEMENT

Il Revenue Management cerca di sfruttare al meglio l'eterogeneità della domanda che riguarda contemporaneamente tre aspetti: il tipo di prodotto, il tipo di cliente e il momento dell'acquisto. Per fare ciò, si identificano tre tipi di Revenue Management:

- Quantity-based Revenue Management;
- Price-based Revenue Management;
- Overbooking.

Nel Quantity-based Revenue Management (modalità di Revenue Management più diffusa), la variabile sotto controllo è la quantità: l'impresa prestabilisce varie categorie di prezzo, ma non definisce a priori la quantità che venderà a quel determinato prezzo; infatti, la quantità effettivamente allocata per ogni categoria di prezzo viene definita in base alla capacità produttiva disponibile controllata in tempo reale. Ad esempio, la capacità di un volo è fissa, ma la compagnia aerea, una volta definito un certo numero di categorie di prezzo, ha la massima

flessibilità nell'allocare i clienti alle diverse categorie di prezzo. La migliore tariffa disponibile cambia dunque nel tempo e, in questo caso, si parla di dynamic pricing o prezzi dinamici (Talluri & Van Ryzin, 2004).

Il modello Quantity-based si basa sulla regola di Littlewood (Talluri & Van Ryzin, 2004), secondo la quale ci sono due gruppi di consumatori: il primo gruppo (Y_1) ha una disponibilità a pagare pari a V_1, mentre il secondo gruppo (Y_2) ha una disponibilità a pagare pari a V_2. Si precisa inoltre che V_1 è maggiore di V_2, questo significa che il primo gruppo ha una disponibilità a pagare maggiore. Si potrebbe decidere di applicare due opzioni di prezzo differenti: se si fissasse un prezzo pari a V_1, acquisterebbe solo il primo gruppo (Y_1), oppure se si abbassasse il prezzo a V_2, invece, acquisterebbero entrambi i gruppi ($Y_1 + Y_2$).

Bisogna tuttavia ricordare che, nei settori in cui si applica il Revenue Management, la capacità produttiva è fissa (basti pensare al numero di posti di un volo aereo o al numero di stanze di un hotel) e inferiore alla somma dei componenti di entrambi i gruppi, ossia la capacità produttiva è inferiore a $Y_1 + Y_2$.

Secondo il modello Littlewood conviene quindi applicare una tariffa promozionale tutte le volte in cui V_2 è maggiore di V_1 per la probabilità che il numero di clienti del primo gruppo (Y_1) sia maggiore della capacità produttiva residua (Kr), ovvero si applica una tariffa promozionale tutte le volte in cui: $V_2 > [V_1 * \text{Prob}(Y_1 > Kr)]$. Bisogna dunque controllare in tempo reale la probabilità che un cliente del primo gruppo acquisti e, successivamente, bisogna confrontarla con la capacità residua che evolve nel tempo e per questo motivo bisogna avere un modello previsionale (Talluri & Van Ryzin, 2004).

Nel Price-based Revenue Management, invece, la variabile sotto controllo è il prezzo. Ad esempio nei negozi al dettaglio, gli spazi sono preallocati e la variabile che si può modificare con più facilità nel breve periodo è il prezzo; la facilità con la quale si può modificare il prezzo può dipendere dal canale di distribuzione e dal tipo di pubblicità, per esempio online o cataloghi (Talluri & Van Ryzin, 2004).

L'overbooking consiste nel vendere le prenotazioni in eccedenza rispetto alla capacità produttiva, considerando le statistiche sui clienti che poi non usufruiscono effettivamente della prenotazione; questa strategia permette di aumentare le vendite e può essere conveniente anche a costo di rischiare di trovarsi nella condizione di non poter dare il servizio al cliente, pagando in tal caso una penalità (Talluri & Van Ryzin, 2004).

1.2 DYNAMIC PRICING

Il dynamic pricing o prezzi dinamici è una strategia di prezzo che consente alle aziende di regolare i prezzi in tempo reale in base alle richieste del mercato, con l'obiettivo di massimizzare i ricavi. La diffusione dei prezzi dinamici è più profonda nelle aziende del settore dei viaggi e dell'ospitalità, dove è difficile cambiare la capacità nel breve periodo e dove i costi variabili sono relativamente bassi (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019).

L'implementazione di questa strategia è stata studiata in diverse aree disciplinari come economia, psicologia, marketing, strategia ed organizzazione, sistemi informativi e problemi di ottimizzazione statistici e matematici. Le strategie di prezzi dinamici sono state infatti approfondite in almeno quattro diversi filoni della letteratura (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019):

- Differenziazione di prezzo intertemporale, che suggerisce di applicare prezzi diversi nel tempo per rivolgersi a consumatori diversi;
- Percezione di ingiustizia dei prezzi, che ha studiato principalmente la relazione tra le variazioni di prezzo e l'equità percepita dei prezzi;
- Controllo e gestione delle scorte, che ha incorporato la presenza di consumatori strategici come una possibile minaccia all'efficacia di tali strategie;
- Cultura organizzativa, che potrebbe spiegare se queste strategie vengono attuate o meno.

La discriminazione dei prezzi è la pratica di applicare prezzi diversi per uno stesso articolo o per prodotti simili che hanno gli stessi costi marginali. Quando l'unica differenza tra gli articoli venduti è il momento dell'acquisto, allora la teoria distingue tra differenziazione intertemporale di prezzo e differenziazione dei prezzi basata sul comportamento: la prima si verifica quando i prezzi differiscono in base al momento dell'acquisto, ma in un dato momento sono uguali per tutti i clienti; mentre la seconda si verifica quando i prezzi sono personalizzati per i vari clienti, in base alla loro storia d'acquisto passata.

Quando si utilizza il tempo come criterio di differenziazione di prezzo, che porta alla segmentazione del mercato, bisogna considerare il grado di pazienza del consumatore e la conseguente relazione tra l'impazienza e la disponibilità a pagare (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019); in particolare se i clienti che hanno una minore disponibilità a pagare sono più impazienti, conviene partire da prezzi più bassi per poi aumentare nel tempo, mentre se i clienti con una maggiore disponibilità

a pagare sono impazienti, conviene partire da prezzi più alti e applicare successivamente degli sconti.

Le variazioni di prezzo sono correlate alla percezione di ingiustizia dei prezzi, che è legata al motivo per cui i prezzi variano; infatti, l'applicazione di prezzi dinamici può modificare la percezione del valore del cliente. Si verifica più precisamente una reazione di ingiustizia nel consumatore quando la decisione di prezzo viene percepita come orientata maggiormente al profitto piuttosto che ai costi. I consumatori possono inoltre fare due tipi di confronti: uno esplicito tra più prezzi di uno stesso prodotto, anche grazie allo sviluppo tecnologico, ed uno implicito tra i prezzi effettivi e il prezzo di riferimento, considerato come giusto da parte del consumatore per quella tipologia di bene (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019). È importante quindi gestire in modo efficace la comunicazione al cliente per favorire l'accettabilità di prezzi dinamici; per fare ciò, è necessario accompagnare la variazione di prezzo ad una diversa percezione di valore, agendo sul "prezzo di riferimento", ovvero il prezzo accettabile da parte del consumatore per un dato servizio, cercando di incrementarlo o oscurarlo nella mente del consumatore, in modo da ridurre la percezione di ingiustizia del prezzo effettivo.

Con il termine controllo e gestione delle scorte ci si riferisce a tutti i modelli di ottimizzazione e ai processi che perseguono l'obiettivo di vendere la giusta quantità di ogni articolo in qualsiasi momento; si vuole quindi saturare la capacità produttiva disponibile, partendo dal principio per cui un'unità non può essere non venduta. Nel settore dell'ospitalità, si è in presenza di beni deperibili ossia beni con una limitazione del numero di articoli disponibili; pertanto in questo caso, il profilo di prezzo di questi beni viene progettato prima che la domanda sia nota, ma può essere eventualmente modificato con delle revisioni in un secondo momento; ad esempio se, quando la domanda attuale risulta inferiore a quella prevista, non ci fosse alcun intervento correttivo, che sposti i prezzi verso il basso, molti articoli rimarrebbero invenduti (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019).

I consumatori vengono suddivisi in due categorie: consumatori miopi e consumatori strategici. I primi effettuano un acquisto se il prezzo è inferiore alla loro valutazione (prezzo di prenotazione) senza considerare i prezzi futuri; i mercati, caratterizzati da questa tipologia di consumatori, consentono ai venditori di ignorare gli effetti negativi delle variazioni future dei prezzi sugli acquisti correnti. I secondi invece scelgono non solo se acquistare un prodotto, ma anche quando acquistarlo; questo gruppo di consumatori cerca di massimizzare l'utilità attesa,

aspettando il primo ribasso per procedere con gli acquisti. In presenza di questa seconda tipologia di consumatori, le decisioni di prezzo dinamico dei venditori sono più complesse perché devono considerare i prezzi attuali e quelli futuri (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019).

Il successo derivante dall'applicazione dei prezzi dinamici dipende da due elementi principali: una struttura informatica appropriata e adatta a garantire il corretto flusso di dati e un'opportuna formazione del personale, al fine di superare lo scetticismo iniziale e migliorare la comprensione dei concetti di Revenue Management. Per quanto riguarda il secondo punto, gli aspetti legati alla formazione e all'apprendimento rimangono una potenziale barriera; infatti, a causa della complessa interazione tra le caratteristiche dell'organizzazione interna, la sua cultura, le diverse aree funzionali e le diverse pressioni ambientali, ci sono ancora problemi di apprendimento. L'applicazione dei prezzi dinamici con informazioni incomplete richiede inoltre competenze avanzate (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019). Occorre infine riconoscere il prezzo come un processo strategico e non solo tattico, che coinvolge le diverse aree funzionali e come possibile fonte di vantaggio competitivo, se l'impresa riesce a comunicare che le variazioni di prezzo sono necessarie per coprire i costi derivanti dalla capacità produttiva.

Oggigiorno, dopo l'iniziale scetticismo, le strategie di prezzo dinamico sono state accettate, sebbene i consumatori spesso pensino ancora che queste strategie siano utilizzate solamente per incrementare i profitti delle aziende. In determinate condizioni, l'uso di queste tecniche offre però vantaggi sia ai manager che ai consumatori, infatti se implementati correttamente, i prezzi dinamici permettono ai clienti più pazienti di ottenere delle offerte convenienti e alle aziende di aumentare i propri ricavi, molto di più di quanto non facciano i prezzi fissi. Senza un'appropriata implementazione invece i prezzi tendono ad essere troppo alti quando la domanda è scarsa e troppo bassi quando la domanda supera le aspettative (Abrate & Viglia, Strategic and tactical price decisions in hotel revenue management, 2016).

1.3 IL SETTORE ALBERGHIERO ED I FATTORI CHE NE INFLUENZANO I PREZZI

Secondo l'"Enciclopedia Treccani", l'albergo è un "edificio appositamente costruito o adattato, attrezzato in modo da poter dare a pagamento alloggio ed eventualmente anche vitto a ospiti di passaggio per un soggiorno temporaneo". Il settore alberghiero è quindi una parte fondamentale del settore turistico; infatti, secondo Wang, Sun, & Wen, il turismo è un'industria importante per molte città e l'industria alberghiera di solito contribuisce maggiormente alle entrate del turismo (Wang, Sun, & Wen, Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model, 2019).

I prodotti e i servizi offerti dagli hotel sono eterogenei; infatti, i relativi prezzi sono influenzati da diversi fattori. I prodotti eterogenei comprendono una miriade di caratteristiche intrinseche. La domanda dei consumatori per i beni non è basata sui prodotti stessi, ma piuttosto sulle caratteristiche/attributi contenuti nel prodotto; la combinazione di questi attributi interessa l'utilità dei consumatori e quindi influenza la disponibilità a pagare dei consumatori. Poiché le caratteristiche/attributi dei prodotti eterogenei hanno prezzi impliciti, che non possono essere osservati direttamente, il modello dei prezzi edonici fornisce un metodo per calcolare questi prezzi impliciti, che riflettono la reale disponibilità a pagare dei consumatori e questa offre un modo relativamente oggettivo di analizzare le determinanti dei prezzi degli hotel (Wang, Sun, & Wen, Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model, 2019).

In generale le caratteristiche collegate ai prezzi delle camere degli hotel possono essere suddivise tra fattori interni e fattori esterni. I primi comprendono i servizi forniti dall'hotel e tra questi, i principali sono: far parte di una catena, il numero di stelle, l'età dell'hotel e servizi come la piscina, il parcheggio, la palestra e l'accesso ad Internet. I fattori esterni, invece, sono riferiti alle caratteristiche della posizione dell'hotel, come la distanza dall'aeroporto, la distanza dal centro della località turistica e l'ambiente circostante all'hotel (Wang, Sun, & Wen, Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model, 2019).

Una distinzione più approfondita dei fattori, che impattano sui livelli di prezzo degli hotel, è quella che individua tre principali tipi di variabili (Abrate & Viglia, Strategic and tactical price decisions in hotel revenue management, 2016):

- Variabili tangibili: caratteristiche fisiche oggettive dei prodotti venduti;
- Variabili reputazionali: classificazione da parte di terzi;
- Variabili contestuali: caratteristiche della posizione e dell'ambiente competitivo.

Nell'area dell'ospitalità, un hotel con differenti tipi di camere può assegnarle in modo tattico a seconda della prenotazione o del cliente target. Se si considera il contributo specifico degli attributi tangibili: il numero di camere, le dimensioni delle camere e la presenza di una spa o di un centro benessere, si ha un effetto sul prezzo. Ci sono poi altri fattori che si considerano solo in base alla destinazione, come la presenza di strutture congressuali per le località più orientate al business o la presenza di una piscina per le località più orientate al tempo libero. Nonostante il confronto incrociato di vecchi studi indichi un impatto divergente dei servizi sui livelli dei prezzi, gli attributi tangibili tendono a rimanere una solida base di riferimento nel determinare il prezzo addizionale da applicare ai servizi turistici e di ospitalità (Abrate & Viglia, Strategic and tactical price decisions in hotel revenue management, 2016).

Per quanto riguarda le variabili reputazionali, invece, uno strumento utilizzato dai consumatori per valutare un servizio turistico o di ospitalità è rappresentato dalle valutazioni numeriche. Queste misure sembrano essere utili per i consumatori nel momento in cui si effettua una prenotazione, in cui si preferisce il ricorso a informazioni facili da valutare come, ad esempio il numero di stelle o le valutazioni online, piuttosto che informazioni più dettagliate. Le valutazioni tendono a influenzare molto la scelta di un prodotto; infatti, è stato dimostrato che le recensioni dei viaggiatori contano di più delle informazioni di chi offre i servizi turistici (Abrate & Viglia, Strategic and tactical price decisions in hotel revenue management, 2016).

Considerando le variabili contestuali, la localizzazione di un servizio turistico e di ospitalità appare di primaria importanza in termini di attrattività e densità dell'area o in termini di concorrenza. Si suggerisce una relazione negativa tra il livello dei prezzi e il numero di concorrenti; questa relazione è stata accentuata soprattutto con la diffusione di Internet che ha semplificato il modo di raccogliere informazioni sul comportamento dei concorrenti. Le variabili contestuali possono essere sfruttate appieno adottando delle tecniche di prezzi dinamici; è stato dimostrato che, in un contesto dinamico, gli hotel diminuiscono i prezzi in tempo reale quando il numero di concorrenti, con almeno una camera disponibile, diminuisce

e che l'effetto della concorrenza è più intenso durante i fine settimana: questo è dovuto alla presenza di viaggiatori turistici che sono più flessibili nella scelta del luogo dove soggiornare (Abrate & Viglia, Strategic and tactical price decisions in hotel revenue management, 2016).

Una variabile contestuale strettamente legata alla concorrenza è il tempo di prenotazione: è stata mostrata l'importanza di regolare i prezzi in base al tempo di prenotazione, al fine di segmentare i consumatori in base alla loro disponibilità a pagare e al loro status, ad esempio tra clienti turistici o clienti business. In generale, se il consumatore che attribuisce un valore elevato al prodotto acquista all'ultimo momento e il consumatore che attribuisce un valore relativamente più basso al prodotto è disposto a pagare in anticipo, la strategia migliore consiste nell'aumentare il prezzo in prossimità del check-in. Il momento della prenotazione può essere considerato una variabile fondamentale per le strategie di prezzi dinamici nel settore dei trasporti, come quello aereo, mentre risulta meno efficiente nel settore alberghiero poiché in quest'ultimo i consumatori possono beneficiare dell'opzione di cancellazione gratuita (Abrate & Viglia, Strategic and tactical price decisions in hotel revenue management, 2016).

Il settore alberghiero è strettamente connesso al concetto di stagionalità, tanto che i livelli di prezzo delle camere degli hotel variano in funzione del periodo del soggiorno. Generalmente si fa una distinzione tra alta e bassa stagione: per alta stagione si intende il periodo in cui si registra una maggiore affluenza turistica che di solito coincide con le festività e vacanze oppure con eventi particolari; per bassa stagione invece si intende il periodo in cui l'affollamento turistico è minore.

Inoltre, come già anticipato all'inizio del capitolo, gli alberghi devono far fronte a funzioni di domanda caratterizzate da diverse elasticità e questo li porta a regolare costantemente i relativi livelli di prezzo: durante l'alta stagione, i prezzi aumentano in seguito ad un aumento della domanda dei consumatori, mentre durante la bassa stagione, per far fronte ad una diminuzione della domanda, c'è una maggiore probabilità che gli hotel possano offrire sconti o promozioni e quindi i prezzi sono più bassi.

Si può dunque affermare che la stagionalità influenza i livelli di prezzo del settore alberghiero. Alcuni studi precedenti hanno infatti analizzato la relazione tra la stagionalità e le strategie di prezzo per i diversi hotel e hanno ottenuto dei risultati interessanti; gli studiosi hanno ad esempio notato che gli hotel caratterizzati da una classificazione a stelle più alta o quelli che appartengono a una catena di solito offrono meno sconti durante la bassa stagione, indicando così variazioni di prezzo stagionali minori, anche in destinazioni di sole e mare. Gli hotel con

una classificazione a stelle più bassa, invece, offrono sconti più frequenti durante la bassa stagione (Wang, Sun, & Wen, Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model, 2019).

Dalla letteratura consultata emerge tuttavia che il fenomeno della stagionalità non viene considerato nella maggior parte degli studi precedenti: in quanto si analizzano dati relativi ad un determinato periodo. Anche in questa ricerca non si osserva la stagionalità poiché il dataset analizzato presenta i prezzi delle camere degli hotel di Londra per il mese di aprile del 2016: è stato, quindi, osservato un unico periodo preciso.

Altri due fattori che stanno diventando sempre più importanti nella ricerca turistica e dell'ospitalità sono le valutazioni e le recensioni degli utenti online: queste possono influenzare in modo significativo l'atteggiamento dei clienti nei confronti di un hotel e le loro decisioni di acquisto: poiché, come segnale di qualità, riflettono la reputazione di un hotel e di conseguenza una valutazione online più alta da parte degli utenti genererebbe un sovrapprezzo maggiore (Wang, Sun, & Wen, Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model, 2019).

È stato inoltre dimostrato che le valutazioni degli utenti sono strettamente collegate al prezzo delle camere degli hotel e hanno un impatto positivo su questi ultimi in tutti i periodi; è infatti emerso che un hotel con una valutazione online più alta aumenti ancora di più il prezzo delle camere durante le stagioni di punta. Infine, è apparso che l'impatto delle valutazioni degli utenti online sul prezzo delle camere degli hotel è eterogeneo e che sembrerebbe maggiore per gli alberghi di fascia media piuttosto che per quelli di lusso (Wang, Sun, & Wen, Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model, 2019).

Oggigiorno, i fattori legati alle nuove piattaforme tecnologiche sono estremamente importanti e per questo vengono analizzati in modo più approfondito nella prossima sezione, legata allo sviluppo tecnologico nel settore alberghiero.

1.4 LO SVILUPPO TECNOLOGICO NEL SETTORE ALBERGHIERO

Negli ultimi anni sono emerse nuove tecnologie che hanno cambiato il panorama delle prenotazioni alberghiere; i siti web si sono infatti evoluti: da contenuti statici si sono trasformati in dinamici generati dagli utenti. Virtualmente ogni utente dunque è un produttore di contenuti e può fornire un feedback su qualsiasi prodotto o servizio, influenzando così gli altri utenti attraverso il passaparola elettronico (electric word-of-mouth, eWOM). Nel settore dell'ospitalità sono risultati particolarmente rilevanti i siti di feedback, come TripAdvisor, dove gli utenti possono dare un punteggio ai vari hotel e/o analizzare le opinioni lasciate dagli altri turisti (Moro, Rita, & Oliveira, 2018).

Con il rapido sviluppo dei social media e delle piattaforme tecnologiche, risultano prevalenti le prenotazioni online (Wang, Sun, & Wen, Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model, 2019): sia attraverso i siti web degli hotel che attraverso le agenzie di viaggio online globali, con quest'ultima che prevale come fonte dominante di prenotazione online (Moro, Rita, & Oliveira, 2018). Durante il processo di acquisto del consumatore, infatti, le agenzie di viaggio online sono diventate il canale di distribuzione più popolare nel settore alberghiero, con circa il 70% delle camere vendute. Le agenzie di viaggio online hanno ottenuto molto successo, in quanto offrono agli hotel una dimensione di mercato significativamente più grande (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019).

La prevalenza delle prenotazioni online deriva inoltre dal fatto che i social media e le aziende di viaggio online danno la possibilità ai consumatori di valutare con trasparenza le opzioni di acquisto: oggi, nel percorso di prenotazione di una camera di un hotel, un turista può accedere ad una piattaforma di recensioni online per valutare le opinioni degli altri utenti prima di effettuare un acquisto, probabilmente integrando le informazioni con il sito web dell'hotel per avere un'idea più precisa. L'utente può accedere ad un'azienda di viaggio online per confrontare i prezzi applicati dalle diverse alternative individuate in precedenza, prima di prenotare definitivamente una camera (Moro, Rita, & Oliveira, 2018).

Queste agenzie operano secondo due possibili modelli: il modello mercantile e il modello di agenzia. Nel modello mercantile, le agenzie di viaggio online acquistano le camere dall'hotel a un prezzo all'ingrosso e le rivendono ai propri clienti ad un prezzo maggiore, in modo da generare profitti; di conseguenza, l'agenzia si fa carico del rischio di inventario delle camere invendute. Nel modello di agenzia, invece, le agenzie di viaggio online si comportano come

agenti nelle transazioni e trasmettendo quindi le prenotazioni che ricevono sulla propria piattaforma all'hotel e ricevono, in cambio, una commissione concordata per ogni transazione effettuata. In questo caso è l'hotel ad assumersi il rischio di inventario per camere non vendute, ma, proprio per questo, ha un margine di profitto maggiore per ogni camera venduta (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019).

La letteratura del marketing riconosce infatti che i consumatori hanno la capacità di influenzarsi a vicenda: su Internet quest'influenza è presente dappertutto e si esercita attraverso raccomandazioni, valutazioni numeriche e recensioni verbali. L'influenza sociale, ossia l'influenza degli atri sul proprio comportamento, assume due forme, denominate rispettivamente influenza normativa ed influenza informativa: la prima si riferisce a quella esercitata dai gruppi di riferimento primari ed ha origine da comportamenti che promuovono la conformità con le aspettative degli altri individui, con lo scopo finale di ottenere ricompense o di eludere le sanzioni. L'influenza informativa, invece, implica l'accettazione di informazioni o consigli da parte di persone che non sono conosciute dal soggetto, ma che forniscono prove affidabili della realtà (Gavilan, Avello, & Martinez-Navarro, 2018).

Oggi i consumatori per prendere decisioni d'acquisto devono confrontarsi con un'enorme quantità di informazioni, nuovi motori di ricerca, diversi dispositivi e nuove strategie di approccio alle informazioni. In questo nuovo contesto, le valutazioni online sono diventate una delle fonti più affidabili quando si prendono decisioni d'acquisto su e-commerce. I consumatori di solito hanno fiducia in questa tipologia di valutazioni e le considerano affidabili, è stato inoltre dimostrato che i consumatori sono disposti a pagare almeno il 20% in più per i servizi che hanno ricevuto una valutazione "Eccellente" o "cinque stelle", rispetto a quelli che hanno ricevuto una valutazione "Buona" o "quattro stelle" (Gavilan, Avello, & Martinez-Navarro, 2018).

La scelta del consumatore è un processo a più stadi in cui i consumatori costruiscono set mentali con opzioni di scelta sempre più piccoli. Secondo il modello del "Consideration set", nei primi stadi del processo decisionale del consumatore, il suo compito è quello di restringere l'insieme delle numerose opzioni a disposizione a quelle più rilevanti. Nel contesto attuale, le valutazioni e il numero di recensioni rappresentano un fattore rilevante per ottenere importanza per una possibile opzione di scelta. Quando prende una decisione d'acquisto, il consumatore si trova in una modalità orientata all'obiettivo, che favorisce un approccio di elaborazione delle informazioni semplice. Le valutazioni sono facili da elaborare e possono essere facilmente

utilizzate per gestire una grande quantità di informazioni ed aiutano a stabilire criteri di selezione, ad esempio solo le opzioni superiori a 4 in una scala a 5. Le valutazioni diventano uno spunto informativo facilmente accessibile che influisce sulla scelta di un prodotto (Gavilan, Avello, & Martinez-Navarro, 2018).

In realtà, ricerche precedenti hanno dimostrato che l'efficacia delle valutazioni e delle recensioni online, come fonte di informazioni per i consumatori, è relativamente limitata. In primo luogo, le recensioni online possono rappresentare solo le preferenze dei consumatori. In secondo luogo, i recensori non sono un campione casuale della popolazione di utenti: i clienti estremamente soddisfatti o insoddisfatti hanno maggiori probabilità di iniziare i passaparola (Gavilan, Avello, & Martinez-Navarro, 2018).

È stato inoltre osservato che le recensioni positive e le recensioni negative influenzano i consumatori in modo diverso. Infatti, gli utenti dedicano più tempo a esaminare e commentare le recensioni negative o miste, ossia quelle che includono sia contenuti positivi che negativi, rispetto a quelle positive; quest'ultime sembrano avere anche un impatto minore sui consumatori. In effetti, l'impatto marginale (negativo) delle recensioni a una stella è maggiore dell'impatto (positivo) delle recensioni a cinque stelle. Questa asimmetria è legata al fatto che tra i recensori ci siano imprenditori, proprietari di aziende o altri soggetti di parte che possono facilmente manipolare i forum online, aggiungendo recensioni anonime positive per elogiare i propri prodotti. Al contrario, le recensioni negative sono considerate provenienti solo da fonti affidabili. Ciò ignora la plausibilità della falsificazione delle recensioni negative per danneggiare un concorrente. Pertanto, l'attendibilità delle valutazioni negative tende a essere più alta di quella delle valutazioni positive (Gavilan, Avello, & Martinez-Navarro, 2018).

Tuttavia, nessuna di queste argomentazioni sembra alterare la fiducia dei clienti nelle valutazioni e recensioni altrui. Al contrario, ogni anno la fiducia dei clienti nelle recensioni online aumenta e diventa addirittura importante quanto le raccomandazioni personali, quando si tratta di prendere decisioni di acquisto (Gavilan, Avello, & Martinez-Navarro, 2018).

Come già detto in precedenza, attraverso i siti di recensioni e i social network, i turisti sono in grado di riferire le loro esperienze sia in termini di punteggi sia con commenti testuali, influenzando i potenziali utenti. Queste informazioni, però, possono essere utilizzate efficacemente anche dai manager per supportare le loro strategie di prezzo (Moro, Rita, & Oliveira, 2018).

Da un punto di vista manageriale, invece, gli albergatori devono stabilire delle strategie di prezzo adeguate a un contesto caratterizzato da trasparenza nei prezzi dato che gli utenti possono simulare le prenotazioni a distanza di un click e confrontare facilmente i prezzi offerti per servizi simili (Gavilan, Avello, & Martinez-Navarro, 2018).

Comprendere le diverse dimensioni che influenzano il comportamento degli utenti è una risorsa fondamentale per supportare le decisioni manageriali nell'attuale mondo dei Big Data; per questo motivo, gli albergatori devono affrontare tutte le variabili disponibili, comprese quelle che non possono controllare, al fine di incorporare una conoscenza approfondita nelle loro strategie di e-marketing per crescere in un mondo sempre più interconnesso (Moro, Rita, & Oliveira, 2018).

CAPITOLO 2: ANALISI DEL DATASET

Il dataset analizzato è un subset di quello usato nell'articolo "The impact of dynamic price variability on revenue maximization" (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019): si tratta di rilevazioni riguardanti i prezzi e altri fattori, che li influenzano, presenti sulla piattaforma Booking.com di 255 hotel della città di Londra per tutto il mese di aprile del 2016. Ogni ricerca fornisce infatti informazioni sul miglior prezzo disponibile per ogni hotel a tre, quattro e cinque stelle nel centro della città, in base ad alcune opzioni di filtraggio attivate nel motore di ricerca. Inoltre, per uniformare la raccolta dati, iniziata due mesi prima, le diverse prenotazioni riguardano la permanenza di una sola persona per una sola notte; in questo modo, risulta più semplice e veloce effettuare un confronto diretto tra i vari hotel, ma anche tra le variabili di uno stesso hotel durante i due mesi studiati.

Per quanto riguarda il processo di rilevazione dei dati, nell'articolo di riferimento (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019), si afferma che il dataset deriva dall'abbinamento di due fonti d'informazione: la raccolta delle informazioni pubbliche disponibili sul web da parte di un'agenzia di viaggio online (online travel agency, OTA) e l'accesso a una banca dati offerta da STR, un'organizzazione specializzata nell'archiviazione di dati alberghieri per il mondo accademico ed imprenditoriale. L'agenzia di viaggio online selezionata per la raccolta dati è la piattaforma Booking.com che opera secondo il modello di agenzia, ossia si comporta come agente nelle transizioni: trasmette quindi le prenotazioni che riceve sulla propria piattaforma all'hotel e, in cambio, riceve una commissione concordata per ogni transazione effettuata. Lascia di conseguenza le decisioni sull'inventario e sui prezzi ai singoli hotel, generando in questo modo un ambiente caratterizzato da strategie di prezzo dinamiche, individuali ed eterogenee.

Nel dataset, le variabili analizzate, oltre al prezzo, sono: tipologia di camera, numero di stelle, numero di preferiti, ossia il numero di utenti che hanno indicato uno specifico hotel come preferito, numero di recensioni, pagina, ossia la pagina di ricerca di Booking.com in cui è apparso uno specifico hotel, posizione, ossia la posizione nell'elenco dei risultati occupata da un determinato hotel durante la ricerca e valutazione. Tutte queste variabili sono state osservate, non solo per quattro settimane consecutive nel mese di aprile (ossia per ventotto giorni), ma anche in otto distanze di tempo diverse a partire da due mesi prima del soggiorno: ipotizzando,

cioè, la prenotazione con diverso anticipo. In particolare, sono stati considerati le seguenti istanti temporali:

- 60 giorni prima del soggiorno;
- 45 giorni prima del soggiorno;
- 30 giorni prima del soggiorno;
- 20 giorni prima del soggiorno;
- 10 giorni prima del soggiorno;
- 4 giorni prima del soggiorno;
- 1 giorno prima del soggiorno;
- Il giorno stesso del soggiorno.

Il dataset analizzato, però, non coincide con quello iniziale poiché quest'ultimo è composto da 28 osservazioni per ognuno dei 255 hotel considerati, per un totale quindi di 7140 osservazioni. Per comprendere al meglio la struttura del dataset iniziale si osservi la **Figura 2-1**, dove sono riportate le ventotto osservazioni per il primo hotel per il giorno del soggiorno.

id_hotel	data_checkin	data_checkout	city	camera0	prezzo0	stelle0	valutazior	n_rev0	n_preferit	posizione	pagina0
1	02/04/2016	03/04/2016	London								
1	03/04/2016	04/04/2016	London	Twin/Double Room	219	4	9	374	438	255	17
1	04/04/2016	05/04/2016	London								
1	05/04/2016	06/04/2016	London	Twin/Double Room	345	4			440	218	15
1	06/04/2016	07/04/2016	London	Twin/Double Room	342	4			440	216	15
1	07/04/2016	08/04/2016	London	Twin/Double Room	245	4	9	378	441	234	16
1	08/04/2016	09/04/2016	London								
1	09/04/2016	10/04/2016	London								
1	10/04/2016	11/04/2016	London	Twin/Double Room	216	4	9	380	441	254	17
1	11/04/2016	12/04/2016	London	Twin/Double Room	364	4	9	380	441	247	17
1	12/04/2016	13/04/2016	London								
1	13/04/2016	14/04/2016	London								
1	14/04/2016	15/04/2016	London	Twin/Double Room	280	4				208	14
1	15/04/2016	16/04/2016	London	Twin/Double Room	220	4	9	383		245	20
1	16/04/2016	17/04/2016	London	Twin/Double Room	220	4			445	234	16
1	17/04/2016	18/04/2016	London								
1	18/04/2016	19/04/2016	London								
1	19/04/2016	20/04/2016	London								
1	20/04/2016	21/04/2016	London	Twin/Double Room	281	4	9	384		172	12
1	21/04/2016	22/04/2016	London	Twin/Double Room	311	4				227	16
1	22/04/2016	23/04/2016	London	Twin/Double Room	282	4	9	384		240	16
1	23/04/2016	24/04/2016	London	Twin/Double Room	323	4	9	384		99	7
1	24/04/2016	25/04/2016	London	Twin/Double Room	285	4				250	17
1	25/04/2016	26/04/2016	London	Twin/Double Room	254	4				233	16
1	26/04/2016	27/04/2016	London								
1	27/04/2016	28/04/2016	London								
1	28/04/2016	29/04/2016	London	Twin/Double Room	298	4				242	17
1	29/04/2016	30/04/2016	London	Twin/Double Room	225	4	9	389		254	17

Figura 2-1: Composizione dataset originale.

Poiché si vuole studiare come cambia l'effetto delle variabili esplicative nell'orizzonte temporale considerato, da sessanta giorni prima al giorno stesso del soggiorno, si è deciso quindi di calcolare il valore medio tra le ventotto osservazioni di uno stesso hotel (**Figura 2-2**); in questo modo, le variabili utilizzate nei modelli rappresentano i valori medi di ciascuna grandezza per ogni hotel. Di conseguenza, si è arrivati ad avere un dataset costituito da 255 osservazioni: una per ogni hotel.

id_hotel	data_chec	data_chec	city	camera0	prezzo0	stelle0	valutazion	n_rev0	n_preferiti	posizioneC	pagina0
1	02/04/201	03/04/201	London								
1	03/04/201	04/04/201	London	Twin/Double Room	219	4	9	374	438	255	17
1	04/04/201	05/04/201	London								
1	05/04/201	06/04/201	London	Twin/Double Room	345	4			440	218	15
1	06/04/201	07/04/201	London	Twin/Double Room	342	4			440	216	15
1	07/04/201	08/04/201	London	Twin/Double Room	245	4	9	378	441	234	16
1	08/04/201	09/04/201	London								
1	09/04/201	10/04/201	London								
1	10/04/201	11/04/201	London	Twin/Double Room	216	4	9	380	441	254	17
1	11/04/201	12/04/201	London	Twin/Double Room	364	4	9	380	441	247	17
1	12/04/201	13/04/201	London								
1	13/04/201	14/04/201	London								
1	14/04/201	15/04/201	London	Twin/Double Room	280	4				208	14
1	15/04/201	16/04/201	London	Twin/Double Room	220	4	9	383		245	20
1	16/04/201	17/04/201	London	Twin/Double Room	220	4			445	234	16
1	17/04/201	18/04/201	London								
1	18/04/201	19/04/201	London								
1	19/04/201	20/04/201	London								
1	20/04/201	21/04/201	London	Twin/Double Room	281	4	9	384		172	12
1	21/04/201	22/04/201	London	Twin/Double Room	311	4				227	16
1	22/04/201	23/04/201	London	Twin/Double Room	282	4	9	384		240	16
1	23/04/201	24/04/201	London	Twin/Double Room	323	4	9	384		99	7
1	24/04/201	25/04/201	London	Twin/Double Room	285	4				250	17
1	25/04/201	26/04/201	London	Twin/Double Room	254	4				233	16
1	26/04/201	27/04/201	London								
1	27/04/201	28/04/201	London								
1	28/04/201	29/04/201	London	Twin/Double Room	298	4				242	17
1	29/04/201	30/04/201	London	Twin/Double Room	225	4	9	389		254	17
1			London	Twin/Double Room	=MEDIA(F2:F29)		9	382	441	226	16

Figura 2-2: Composizione dataset finale.

Nel dataset iniziale, però, ci sono ventinove hotel che non presentano informazioni per tutti e otto gli istanti di tempo considerati; questo significa che, al momento della raccolta dati, questi hotel non presentano camere disponibili sulla piattaforma Booking.com per tutti gli otto momenti temporali analizzati. Si è quindi deciso di eliminare i ventinove hotel dalla matrice dati; in quanto altrimenti risulterebbe incompleta (**Figura 2-3**). Il dataset finale è dunque composto da 226 osservazioni, anziché 255.

id_hotel	city	camera0	prezzo0	stelle0	valutazion	n_rev0	n_preferiti	posizione0	pagina0
1	London	Twin/Doul	277	4	9	382	441	226	16
2	London	Twin/Doul	366	5	8	575	1686	156	11
3	London	Twin/Doul	144	4	7	3707	7441	62	5
4	London	Twin/Doul	212	4	9	479	1496	173	12
5	London	Twin/Doul	244	4	8	3188	7450	102	7
6	London	Twin/Doul	109	3	8	1010	1786	319	22
7	London	Twin/Doul	185	4	9	852	2844	325	22
8	London	Twin/Doul	259	5	9	224	598	251	17
9	London	Twin/Doul	160	4	8	5675	11403	77	6
10	London	Twin/Doul	266	4	9	256	390	290	20
11	London	Twin/Doul	266	4	9	1930	1695	306	21
12	London	Twin/Doul	155	3	8	2035	2523	316	22
13	London	Twin/Doul	270	4	9	302	1069	107	8
14	London	Twin/Doul	187	3	8	784	1277	267	18
15	London	Twin/Doul	178	4	9	1890	6555	118	8
16	London	Twin/Doul	293	4	9	1039	2975	179	12
17	London	Twin/Doul	205	4	8	672	1461	192	13
19	London	Twin/Doul	240	4	9	8646	17637	26	2
20	London	Twin/Doul	170	3	9	1201	1852	341	23
21	London	Twin/Doul	579	5	9	361	1360	90	6
22	London	Twin/Doul	219	4	9	2438	3397	269	18
23	London	Twin/Doul	560	5	9	998	1794	166	12
24	London	Twin/Doul	265	4	9	1076	1437	111	8
25	London	Twin/Doul	214	4	9	645	1611	114	8
26	London	Twin/Doul	463	5	9	133	617	286	20
27	London	Twin/Doul	303	4	8	2841	4709	26	2

Figura 2-3: Dataset finale.

Si passa ora ad analizzare le singole variabili osservate durante l'indagine, sia con grafici che con commenti descrittivi, utilizzando i seguenti programmi: RStudio ed Excel. Nella parte finale del capitolo, invece, si procede con la verifica dell'eventuale presenza di correlazione tra le variabili in esame.

2.1 LE VARIABILI

Il dataset oggetto di studio, come già anticipato, è composto da 226 osservazioni, una per ogni hotel, e da otto variabili: “camera”, “stelle”, “n_preferiti”, “n_rev”, “pagina”, “posizione”, “prezzo” e “valutazione”. Le variabili sono state osservate in otto istanti temporali differenti; per questo motivo, presentano i seguenti suffissi: 0, 1, 4, 10, 20, 30, 45 e 60, i quali indicano i giorni prima del soggiorno. Tra le variabili in esame è possibile individuare due gruppi distinti: il primo gruppo comprende le variabili che per ogni hotel assumono un valore che rimane costante per l’intero orizzonte temporale considerato; tra queste vi sono:

- Camera;
- Stelle;
- Valutazione.

Il secondo gruppo, invece, è costituito da quelle variabili che cambiano velocemente nel tempo, ovvero assumono, per ogni hotel, valori differenti per i vari istanti di tempo osservati; tra queste vi sono:

- N_preferiti;
- N_rev;
- Pagina e Posizione;
- Prezzo.

2.1.1 CAMERA

La variabile “camera” indica la tipologia di camera che è apparsa sulla piattaforma Booking.com come opzione principale, per ciascun hotel, al momento della raccolta dati. Le tipologie di camere osservate sono le seguenti:

- Twin/Double Room;
- King Junior Suite;
- Standard Suite.

In particolare, per 224 hotel su 226 è apparsa come opzione principale la tipologia Twin/Double Room, la quale rappresenta quindi la quasi totalità delle camere analizzate. Invece, sia per la tipologia King Junior Suite che per la tipologia Standard Suite si è osservato un solo caso, ovvero sono entrambe apparse come opzione principale per un solo albergo.

Per questo motivo, la variabile “camera” non viene considerata nei modelli studiati poiché le opzioni King Junior Suite e Standard Suite sono ininfluenti, in quanto rappresentano insieme solo l’1% delle osservazioni.

2.1.2 STELLE

La variabile “stelle” indica il numero di stelle degli hotel osservati; normalmente ci si aspetta un prezzo maggiore per gli hotel che hanno un numero di stelle più alto. Si tratta di una variabile qualitativa a tre livelli:

- Hotel con tre stelle;
- Hotel con quattro stelle;
- Hotel con cinque stelle.

Tra gli hotel considerati, la maggior parte sono hotel con quattro stelle (122 alberghi su 226, il 54% del dataset), il restante 46% si divide tra hotel con tre stelle e hotel con cinque stelle; in particolare, quelli con tre stelle sono 23 su 226, mentre quelli con cinque stelle sono 81 su 226. Questo significa che gli hotel con tre stelle rappresentano solo il 10% dei casi osservati. È possibile visualizzare tale distribuzione col grafico della **Figura 2-4**.

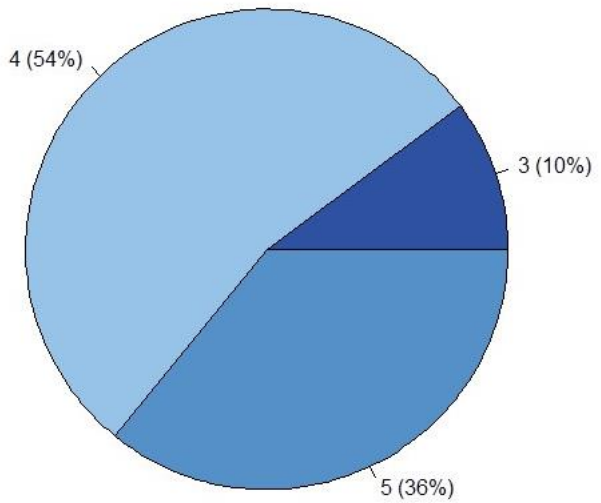

Figura 2-4: Suddivisione degli hotel per il numero di stelle.

2.1.3 VALUTAZIONE

La variabile "valutazione" indica il punteggio medio delle recensioni, per ogni hotel, lasciate dagli utenti sulla piattaforma Booking.com, attraverso l'impostazione di una scala likter che, in questo caso, comprende valori che vanno da sette a nove. Questa variabile rappresenta un elemento di qualità percepita: infatti, se un hotel dovesse avere una valutazione elevata, allora vorrebbe dire che è molto apprezzato dagli utenti; di conseguenza, ci si aspetta che la variabile in esame impatti positivamente sul prezzo.

Dalla **Figura 2-5**, la quale rappresenta la distribuzione delle valutazioni, si osserva che il 90% del campione è rappresentato da hotel con valutazione otto e valutazione nove: in particolare, su un totale di 226 hotel, 114 hanno una valutazione media pari a nove e 91 hanno una valutazione media pari a otto. Il restante 10% del campione, invece, è suddiviso tra hotel con valutazione sette e valutazione dieci: più precisamente, su un totale di 226 hotel, 16 hanno una valutazione media pari a sette e, infine, solo 5 hanno una valutazione media massima, ossia pari a dieci; ciò significa che quest'ultimo 2.2% del campione è stato valutato in modo perfetto dagli utenti della piattaforma Booking.com.

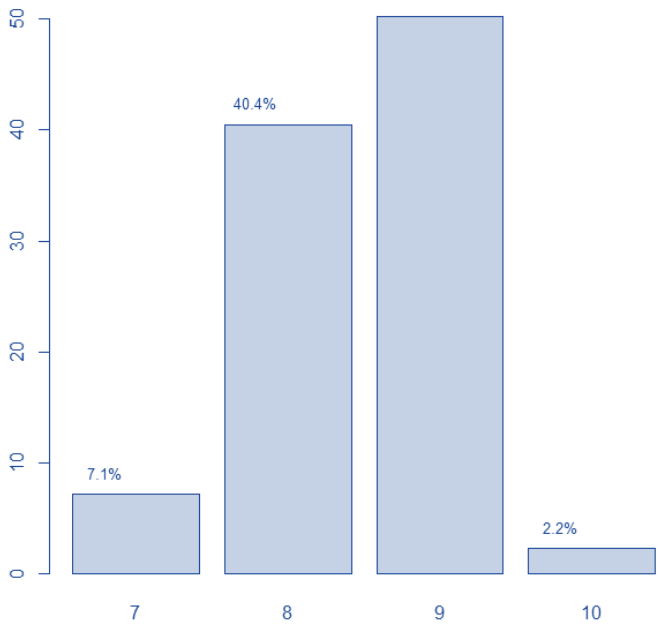

Figura 2-5: Suddivisione degli hotel per valutazione.

Per avere un'analisi più approfondita sono stati calcolati gli indicatori presenti nella **Tabella 2-1**, i quali possono essere anche osservati attraverso il boxplot rappresentato nella **Figura 2-6**.

VALORE MINIMO	**7**
PRIMO QUARTILE	8
MEDIANA	9
MEDIA	8
TERZO QUARTILE	9
VALORE MASSIMO	10

Tabella 2-1: Indicatori della variabile valutazione.

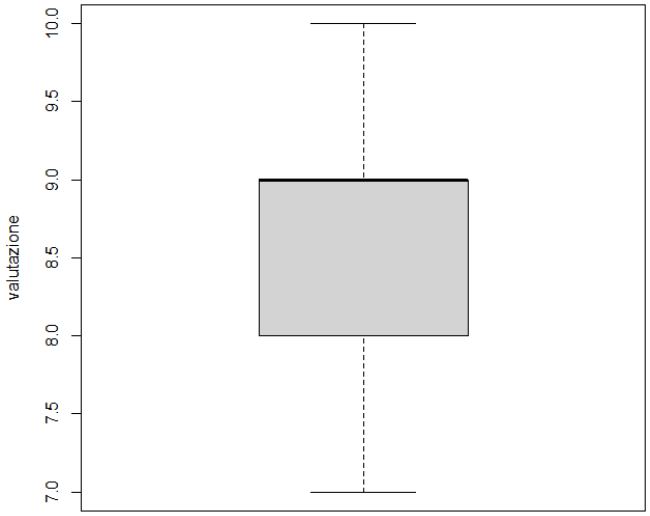

Figura 2-6: Boxplot della variabile valutazione.

Dai risultati ottenuti si può affermare che la maggior parte degli hotel ha ottenuto una valutazione pari o superiore alla media; in quanto, il primo quartile è pari a otto e, quindi, il 75% delle osservazioni ha una valutazione media pari o superiore a quest'ultimo. Questo significa che, per il campione considerato, gli utenti della piattaforma Booking.com hanno in media valutato positivamente la propria esperienza, confermando quindi l'elevata qualità delle strutture ricettive osservate.

	HOTEL 3 STELLE	HOTEL 4 STELLE	HOTEL 5 STELLE
VALORE MINIMO	7	7	8
PRIMO QUARTILE	7	8	9
MEDIANA	8	8	9
MEDIA	8	8	9
TERZO QUARTILE	8	9	9
VALORE MASSIMO	9	10	10

Tabella 2-2: Indicatori della variabile valutazione in base al numero di stelle.

Nella **Tabella 2-2**, dove si mette in relazione la variabile "valutazione" con il numero di stelle, si nota che in generale ad un numero di stelle maggiore corrisponde una valutazione maggiore; in particolare, si può osservare che gli hotel con tre stelle hanno complessivamente una valutazione media compresa tra sette e otto, mentre per gli hotel con quattro stelle si registrano valutazioni medie comprese principalmente tra otto e nove e, infine, gli hotel con cinque stelle presentano una valutazione media pari a nove; inoltre, si osserva che la valutazione pari a dieci, la quale rappresenta la perfezione, si registra solo per alcuni hotel con quattro e cinque stelle. Tali risultati possono essere osservati anche graficamente nella **Figura 2-7**.

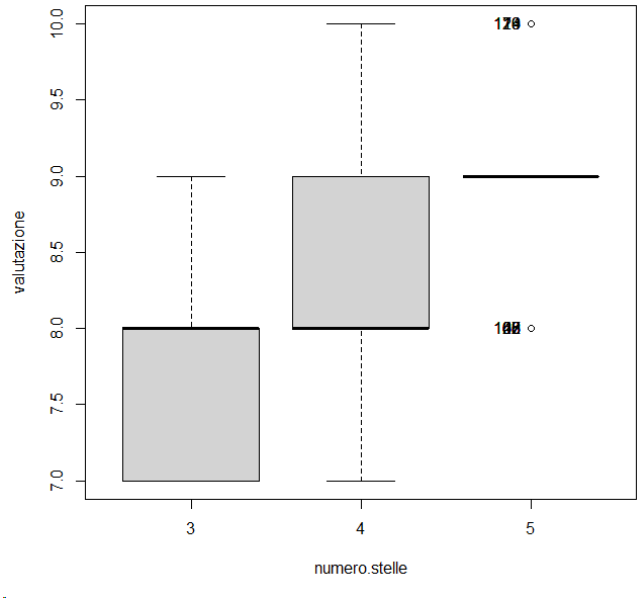

Figura 2-7: Boxplot della variabile valutazione in funzione del numero di stelle.

2.1.4 N_PREFERITI

La variabile "n_preferiti" indica il numero di utenti che ha indicato uno specifico hotel tra i preferiti: se un hotel avesse un elevato numero di preferiti, vorrebbe dire che quell'hotel è molto apprezzato dagli utenti. Si precisa che i valori analizzati sono calcolati come la media del numero di preferiti di uno stesso hotel osservato per ventotto giorni consecutivi.

Nella **Figura 2-8** sono riportati, a scopo esemplificativo, gli andamenti di sei hotel, selezionati all'interno del campione. Si può osservare che, per l'orizzonte temporale considerato, il numero medio di preferiti dei singoli hotel considerati ha un andamento crescente. Si possono notare tuttavia delle differenze tra i singoli hotel: per alcuni, il numero medio di preferiti cresce in

modo esponenziale, per altri invece aumenta, anche se in modo più contenuto, e per altri ancora sembra quasi costante. Vi sono anche degli alberghi che, pur avendo in generale un andamento crescente, in alcuni istanti temporali subiscono una leggera diminuzione del numero medio di preferiti.

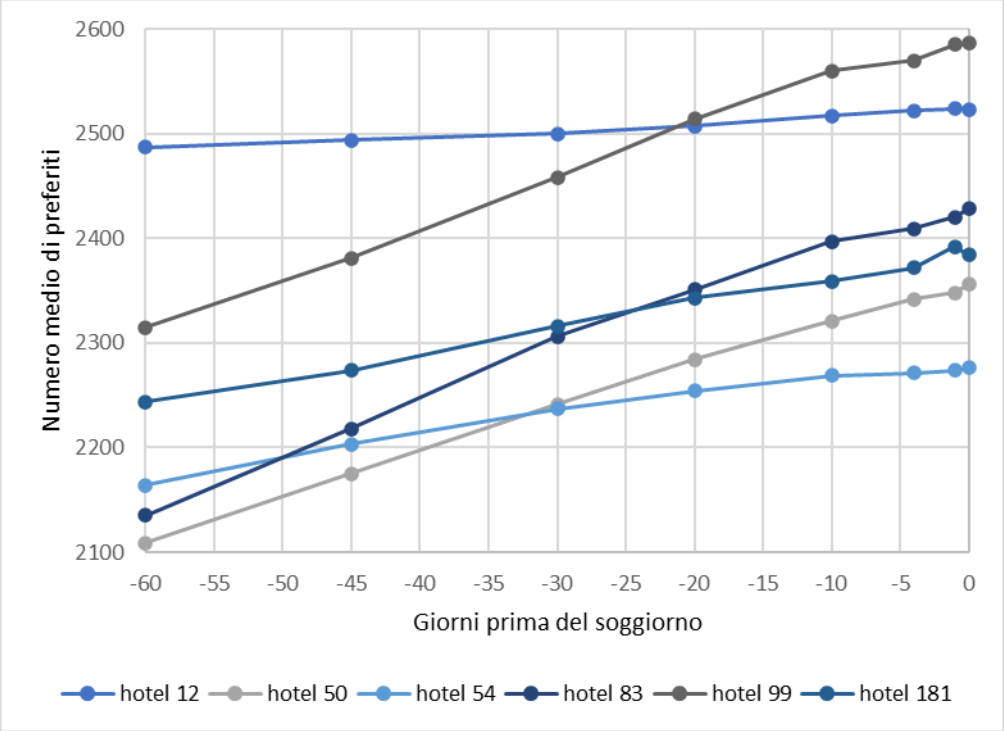

Figura 2-8: Andamenti della variabile numero di preferiti.

Per avere una visione globale della variabile in esame, si è deciso di calcolare anche il valore medio del numero di preferiti, per ogni hotel, tra gli otto istanti temporali considerati, in modo da calcolare gli indicatori presenti sia nella **Tabella 2-3** che nel boxplot rappresentato nella **Figura 2-9**. Dai risultati ottenuti si può notare che c'è un range molto ampio tra il numero medio di preferiti degli hotel considerati: infatti, il valore più piccolo che appare nella distribuzione è pari a 91, ciò vuol dire che nel dataset esiste almeno un hotel che ha un numero medio di preferiti molto basso; invece, il valore massimo è pari a 17189, il quale rappresenta l'hotel con il maggior numero medio di preferiti.

La mediana, che rappresenta il valore centrale della distribuzione, è pari a 1740: questo significa che il 50% delle osservazioni ha un numero medio di preferiti minore a quest'ultimo valore, mentre il restante 50% ha un numero medio di preferiti superiore. In particolare, osservando il

primo quartile, si può affermare che il 25% delle osservazioni ha un numero medio di preferiti inferiore a 941 e quindi un quarto delle osservazioni ha ottenuto un numero di preferiti piuttosto basso. Il terzo quartile, invece, è pari a 2971 dunque il 75% degli hotel considerati ha un numero medio di preferiti inferiore a quest'ultimo, che a sua volta è molto inferiore rispetto al valore massimo registrato, pari a 17189. Di conseguenza, per la maggior parte degli hotel considerati, il numero medio di utenti che ha indicato uno specifico hotel come preferito è compreso tra 91 e 2971; i valori che si discostano in modo significativo da questo range sono chiamati outliers o valori anomali e sono raffigurati nel grafico a scatola e baffi della **Figura 2-9**: la presenza di questi outliers spiega l'enorme differenza che esiste tra il valore massimo registrato e la maggior parte del numero medio di preferiti.

VALORE MINIMO	91
PRIMO QUARTILE	941
MEDIANA	1740
MEDIA	2457
TERZO QUARTILE	2971
VALORE MASSIMO	17189

Tabella 2-3: Indicatori della variabile numero di preferiti.

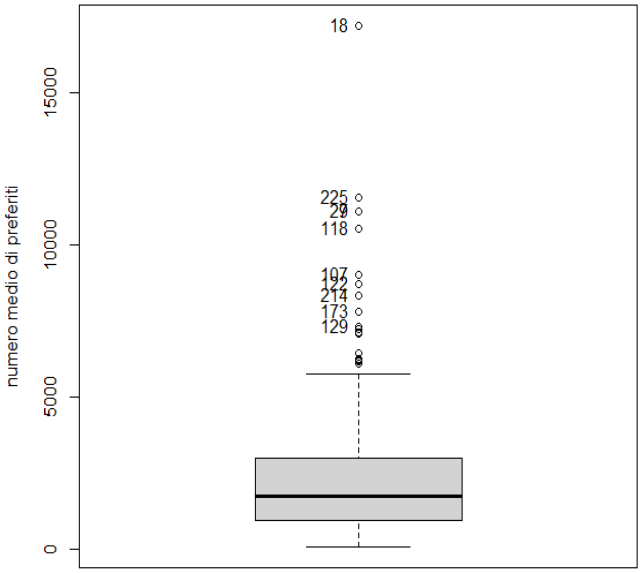

Figura 2-9: Boxplot della variabile numero di preferiti.

Poiché dalla **Figura 2-8** si è osservato che la variabile in esame assume andamenti differenti, per i vari hotel, oltre a calcolare il numero medio di preferiti per hotel, si è deciso di calcolare anche il coefficiente di variazione intertemporale (cv), dato dal rapporto tra deviazione standard e numero medio di preferiti; in questo modo si riesce a capire quanto i dati si discostano dalla loro media in termini relativi. La distribuzione del coefficiente di variazione intertemporale del numero medio di preferiti è stata rappresentata attraverso un grafico di densità presente nella **Figura 2-10**, dove si può osservare che si tratta di una distribuzione bimodale poiché è caratterizzata da due picchi, i quali di conseguenza potrebbero rappresentare due sottogruppi o cluster tra le osservazioni che si differenziano per il valore del coefficiente di variazione, ossia per la variabilità dei dati attorno alla media; tuttavia, non si registra un'elevata differenza, in

quanto i cv dei due gruppi sono entrambi compresi tra lo 0 e lo 0.05. È interessante osservare che entrambi i picchi hanno una densità di probabilità massima pari a 40: questo potrebbe indicare un equilibrio tra i due gruppi. La distribuzione di densità risulta inoltre essere asimmetrica: poiché, nonostante la maggior parte delle osservazioni sia caratterizzata da un coefficiente di variazione intertemporale non superiore allo 0.05, si può notare una lunga coda verso destra, la quale conferma la presenza di outliers e quindi una maggiore variabilità dei dati.

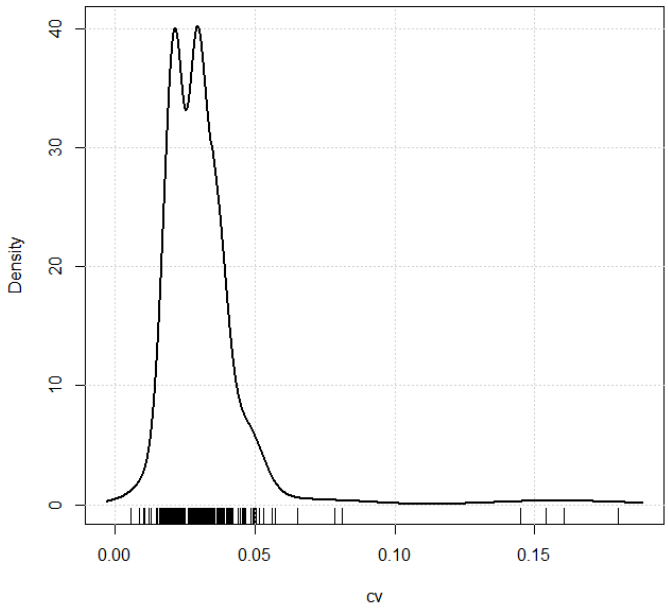

Figura 2-10: Distribuzione del coefficiente di variazione intertemporale del numero di preferiti.

2.1.5 N_REV

La variabile "n_rev" rappresenta il numero di recensioni che sono state lasciate sulla piattaforma Booking.com dagli utenti che hanno soggiornato in ciascuno degli hotel presenti nel dataset; con questa variabile si indicano sia le recensioni positive che quelle negative, di conseguenza, un aumento del numero di recensioni non rappresenta sempre un elemento di qualità per l'hotel, come invece accade per il numero di preferiti. Si precisa, inoltre, che i valori analizzati sono medie tra i numeri di recensioni di uno stesso hotel osservati per quattro settimane consecutive.

Nella **Figura 2-11** sono riportati a scopo esemplificativo gli andamenti di sei hotel selezionati all'interno del campione. Si può notare che in generale la variabile in esame assume un andamento crescente, anche se si osservano delle differenze tra gli andamenti degli hotel

selezionati: per alcuni hotel, il numero medio di recensioni cresce in maniera esponenziale, per altri, invece, aumenta in modo più contenuto e per altri ancora, ha un andamento crescente, anche se, per alcuni degli istanti temporali osservati, diminuisce leggermente.

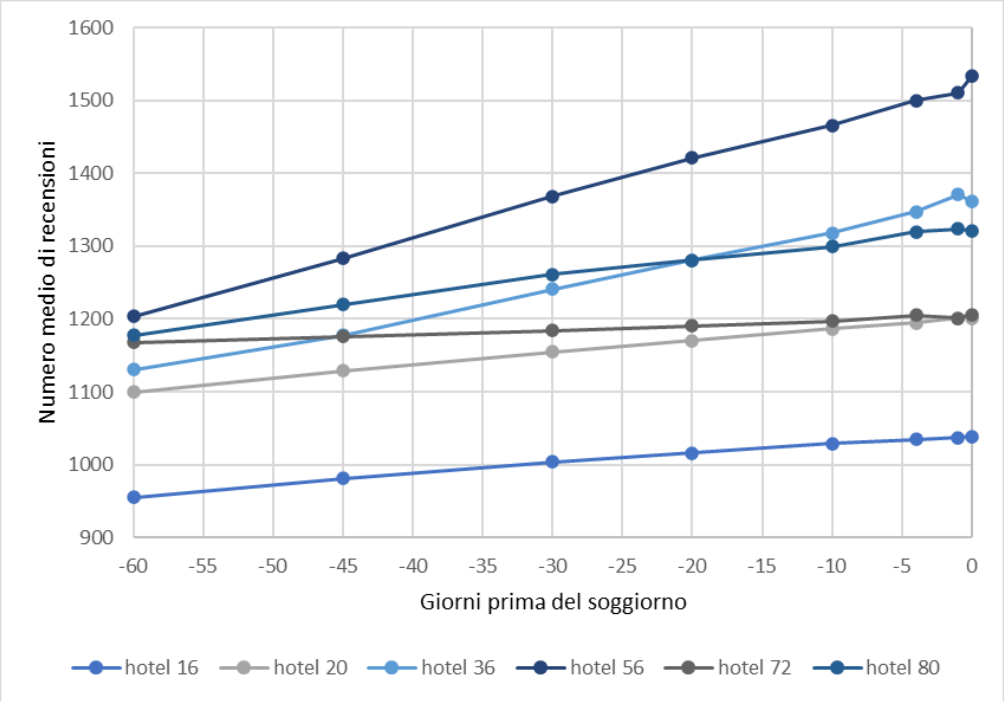

Figura 2-11: Andamenti della variabile numero recensioni.

Per avere una visione globale di questa variabile, si è deciso di calcolare il valore medio del numero di recensioni di ogni hotel tra gli otto istanti temporali considerati; i risultati sono riportati nella **Tabella 2-4** e nella **Figura 2-12**, dove si può notare che la distribuzione del numero medio di recensioni presenta un range molto ampio; infatti, il valore minimo è pari a 14, ossia esiste almeno un hotel che ha un numero medio di recensioni molto basso; mentre il valore massimo è pari a 8648, il quale rappresenta l'hotel che ha registrato il maggior numero medio di recensioni.

La mediana, la quale rappresenta il valore centrale della distribuzione, è pari a 906: questo significa che il 50% delle osservazioni ha un numero medio di recensioni inferiore a quest'ultimo e il restante 50% ha un numero medio di recensioni superiori a questo valore. Più precisamente, se si considera il terzo quartile, risulta che il 75% delle osservazioni ha un numero medio di recensioni inferiore a 1725, il quale si discosta molto dal valore massimo registrato pari a 8648: questa differenza indica la presenza di outliers, ossia punti estremi che si discostano

in modo significativo dalla maggior parte delle osservazioni. Questi punti sono evidenziati anche nel grafico a scatola e baffi, presente nella **Figura 2-12**, dove appaiono infatti dei punti al di fuori del limite superiore.

VALORE MINIMO	14
PRIMO QUARTILE	385
MEDIANA	906
MEDIA	1324
TERZO QUARTILE	1725
VALORE MASSIMO	8648

Tabella 2-4: Indicatori della variabile numero di recensioni.

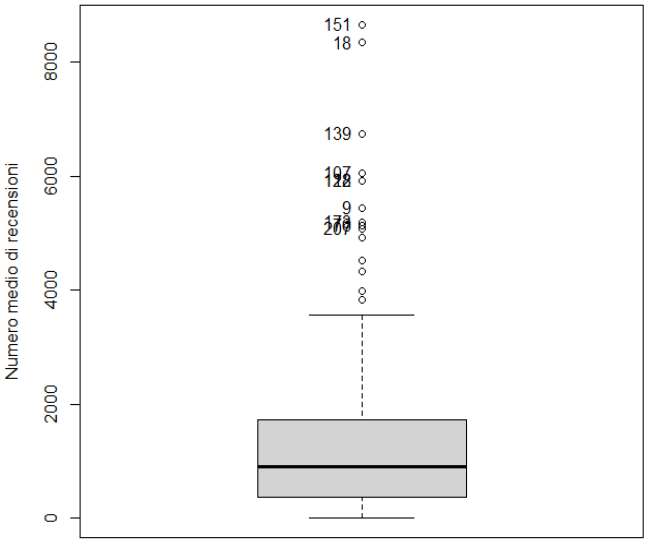

Figura 2-12: Boxplot della variabile numero di recensioni.

Poiché dalla **Figura 2-11** è emerso che il numero di recensioni assume andamenti differenti tra i vari hotel del campione, si è deciso di calcolare anche la deviazione standard e il coefficiente di variazione di ogni hotel, il quale è dato dal rapporto tra deviazione standard e numero medio di recensioni: in questo modo è possibile capire quanto si discostano i dati dalla loro media in termini relativi. La distribuzione del coefficiente di variazione intertemporale del numero di recensioni è rappresentata dalla **Figura 2-13**.

Si può osservare che si tratta di una distribuzione asimmetrica, in quanto la distribuzione ha un picco concentrato sulla parte sinistra del grafico, mentre nella parte destra appare una lunga coda, la quale evidenzia la presenza di outliers e quindi una maggiore variabilità dei dati. Si può inoltre notare che, per la maggior parte degli hotel del campione, il numero di recensioni per i vari istanti temporali considerati varia intorno alla relativa media di una percentuale inferiore al 10%.

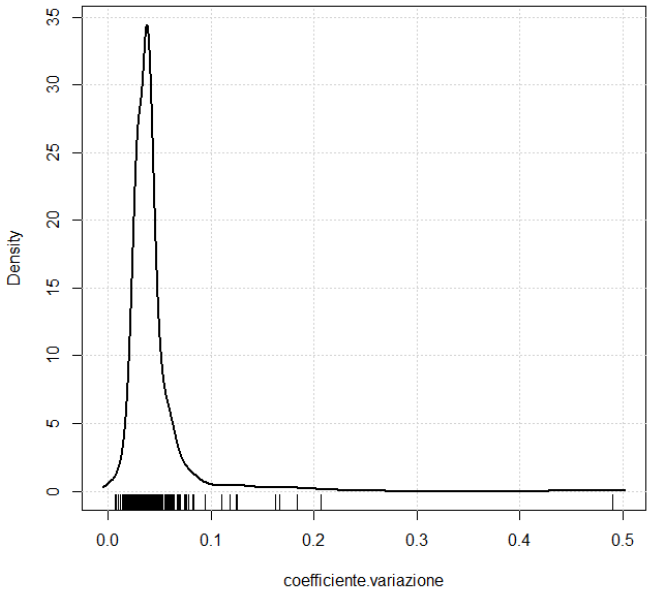

Figura 2-13: Distribuzione del coefficiente di variazione intertemporale del numero di recensioni.

2.1.6 PAGINA E POSIZIONE

Le variabili "pagina" e "posizione" esprimono entrambe, anche se in maniera diversa, la posizione occupata da uno specifico hotel nell'elenco dei risultati della piattaforma Booking.com, al momento della ricerca. Più precisamente, la variabile "pagina" indica il numero della pagina di ricerca in cui appare uno specifico hotel: nel dataset in esame, questa variabile ha assunto, in media, valori compresi tra 1 e 24. Si precisa, inoltre, che il valore di questa variabile dipende da quanti hotel si visualizzano per ogni pagina di ricerca: in questo caso, ogni pagina contiene quindici hotel. La variabile "posizione", invece, indica la posizione occupata da uno specifico hotel nell'elenco totale dei risultati della ricerca e non la posizione all'interno di una determinata pagina di ricerca. Nel dataset analizzato, la variabile in questione ha assunto, in media, valori compresi tra 7 e 359.

Si parla di pagina e posizione media poiché il valore che corrisponde alle variabili in esame, per ogni singolo hotel e per ciascun istante temporale considerato, è in realtà la media tra i valori assunti dalla stessa variabile nei ventotto giorni consecutivi dell'aprile del 2016. Di conseguenza, se ad esempio ad entrambe le variabili dovesse corrispondere il numero uno, per un determinato istante temporale, allora vorrebbe dire che, per quell'istante temporale, l'hotel in questione si trova, in media, alla prima posizione e, quindi, nella prima pagina dei risultati. Se, invece, ad un determinato hotel dovessero corrispondere i seguenti valori: 150 per la variabile "posizione" e 10 per la variabile "pagina", vorrebbe dire che, al momento della ricerca,

l'hotel, in media, è al centocinquantesimo posto nell'elenco dei risultati e, poiché vi sono quindici hotel per pagina, allora si troverebbe nella pagina numero dieci della piattaforma Booking.com.

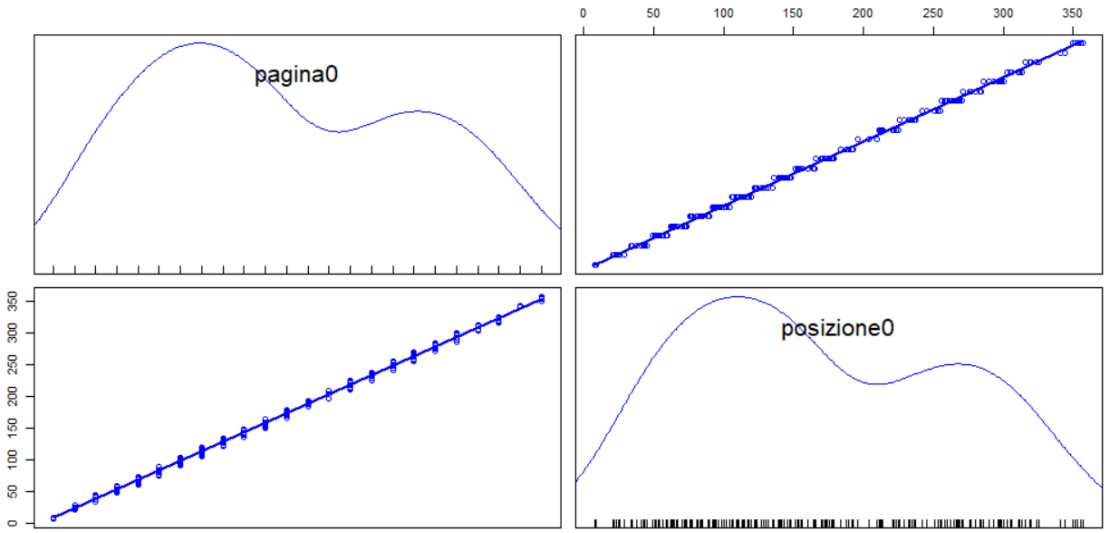

Figura 2-14: Matrice di correlazione con grafici a dispersione.

Come già anticipato, le due variabili in esame esprimono lo stesso concetto, anche se in maniera differente; infatti, osservando la **Errore. L'origine riferimento non è stata trovata.**, dove è r iportata la matrice di correlazione tra le due variabili, si può dedurre che le due variabili sono una la copia dell'altra in quanto la nuvola di punti è perfettamente disposta lungo la linea dei minimi quadrati. Di conseguenza, poiché le due variabili sono strettamente correlate, si è deciso di procedere con l'analisi di una sola delle due variabili in esame, la variabile "posizione" poiché, in caso contrario, sarebbe una ripetizione.

Nella **Figura 2-15**, sono raffigurati, a titolo esemplificativo, sei diversi andamenti della variabile "posizione", i quali possono rappresentare le differenti situazioni verificatesi nel campione analizzato. Facendo un confronto tra gli andamenti riportati, si può osservare che non si ha un comportamento univoco tra i vari hotel considerati, ma si ha piuttosto un'elevata variabilità, infatti per alcuni hotel le variabili in esame hanno un andamento crescente, per altri si ha un andamento decrescente, per altri ancora rimane costante o si ha una variazione minima e infine si hanno hotel caratterizzati da un andamento oscillante: prima decresce, poi aumenta e successivamente decresce ulteriormente, e così via.

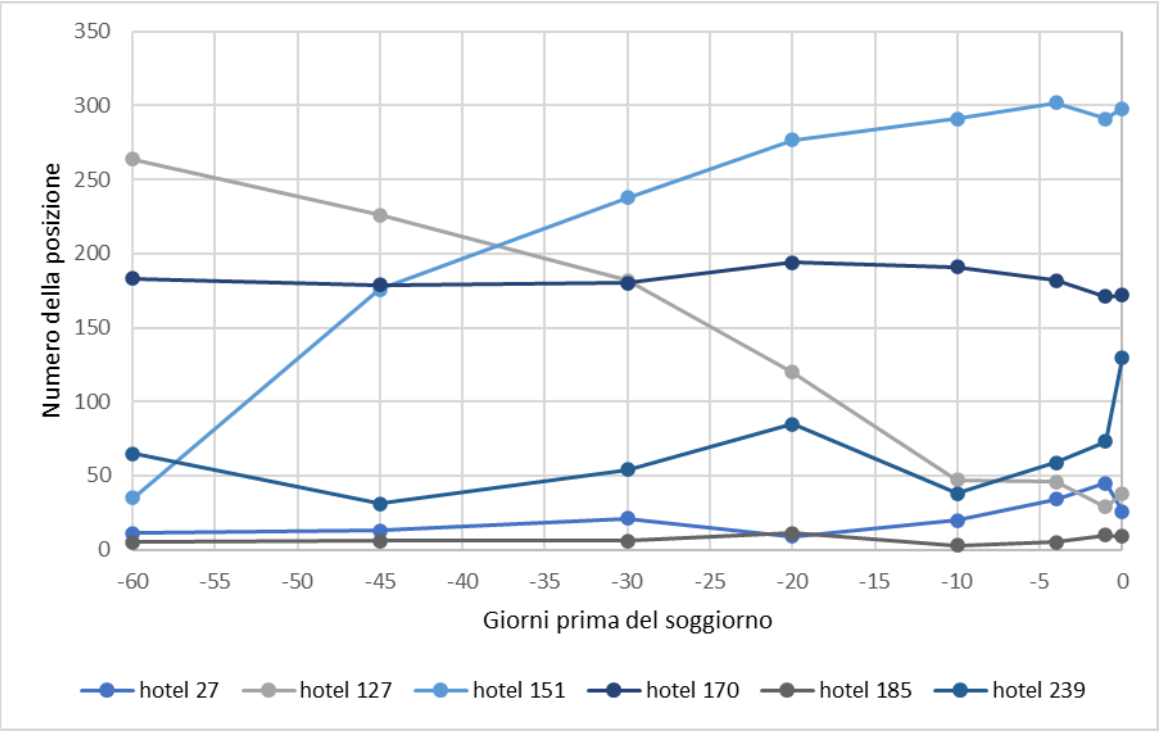

Figura 2-15: Andamenti della variabile posizione.

Poiché le variabili in esame assumono comportamenti differenti tra i vari hotel, per avere una visione più completa, si è deciso di calcolare i valori medi di ognuna delle due variabili per ciascun hotel. Utilizzando i valori medi della variabile "posizione", sono stati calcolati i seguenti indicatori (**Tabella 2-5**), i quali possono essere osservati graficamente nella **Figura 2-16**.

VALORE MINIMO	7
PRIMO QUARTILE	86
MEDIANA	145
MEDIA	166
TERZO QUARTILE	260
VALORE MASSIMO	359

Tabella 2-5: Indicatori della variabile posizione.

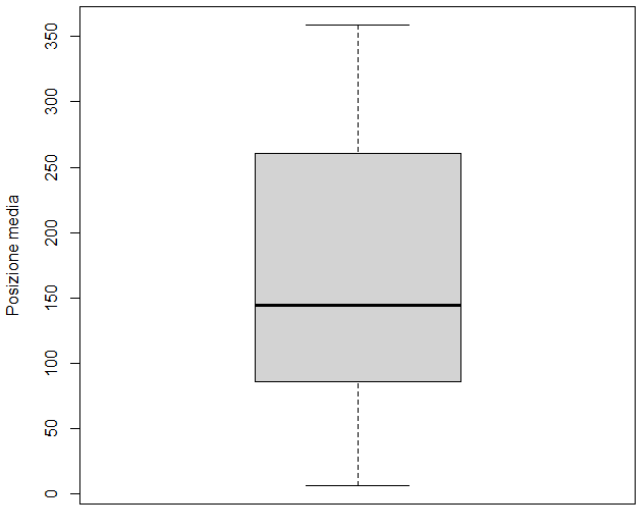

Figura 2-16: Boxplot della variabile posizione.

Si può notare che gli hotel del dataset in esame, nei diversi istanti temporali osservati, hanno assunto in media posizioni comprese tra 7 e 359 che rappresentano rispettivamente la posizione media minima e la posizione media massima; essendo però dei valori medi significa che, nel dataset, ci sono hotel che hanno occupato per uno o più istanti temporali una posizione inferiore alla settima o superiore alla trecentocinquantanovesima.

Poiché per ogni pagina di ricerca ci sono quindici hotel, considerando il valore minimo pari a 7, significa che esiste almeno un hotel che in media si trova nella prima pagina dei risultati e quindi gode di un'ottima visibilità. In particolare, considerando il primo quartile pari a 86, vuol dire che il 25% del dataset ha occupato mediamente una posizione compresa tra le prime sei pagine di ricerca e viene dunque visualizzato maggiormente dagli utenti rispetto agli hotel che occupano posizioni superiori.

La mediana, la quale indica il valore centrale della distribuzione, è pari a 145: questo significa che il 50% delle osservazioni occupa mediamente una posizione inferiore a quest'ultima e, invece, il restante 50% occupa in media una posizione superiore. Di conseguenza, si può affermare che il primo 50% degli hotel gode generalmente di una buona visibilità, mentre la restante parte viene visualizzata meno dagli utenti della piattaforma.

Il terzo quartile risulta essere pari a 260: questo vuol dire che il 75% del campione occupa in media una posizione compresa tra la prima e la diciottesima pagina dei risultati, mentre il restante 25% si trova mediamente nelle ultime pagine dei risultati, fino alla ventiquattresima, siccome la posizione media massima rilevata è pari a 359.

Infine, poiché le osservazioni hanno comportamenti molto diversi tra di loro, si è deciso di calcolare il coefficiente di variazione intertemporale, dato dal rapporto tra deviazione standard e media; in questo modo è possibile osservare la variabilità dei dati rispetto alla loro media. Tali risultati sono riportati nella **Figura 2-17**, la quale rappresenta la distribuzione del coefficiente di variazione intertemporale dei singoli hotel per la variabile in esame. Si può osservare che si tratta di una distribuzione asimmetrica, in quanto si ha un picco elevato nella parte sinistra del grafico e dei picchi minori nella restante parte del grafico. La maggior parte delle osservazioni ha un coefficiente di variazione intertemporale inferiore allo 0.3 quindi, per la maggioranza degli hotel, la variabile "posizione" varia intorno alla relativa media per il 30%. Ci sono poi hotel che sono caratterizzati da una variabilità maggiore, in quanto hanno un coefficiente di variazione intorno allo 0.4. Vi sono infine cinque hotel che sono caratterizzati da un'elevata variabilità, per la variabile "posizione"; infatti, sono caratterizzati da un coefficiente di

variazione intertemporale compreso tra il 50% e l'80%; per questi hotel quindi i dati sono dispersi maggiormente intorno alla media.

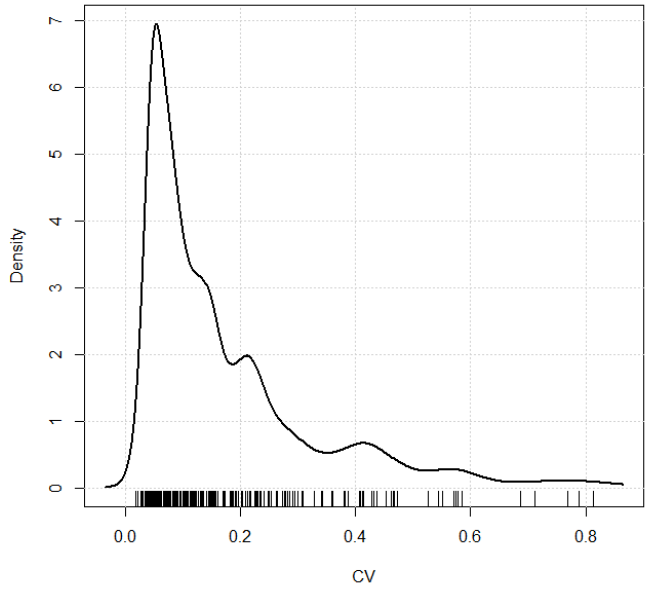

Figura 2-17: Distribuzione del coefficiente di variazione della variabile posizione.

2.1.7 PREZZO

La variabile "prezzo" indica il prezzo medio del soggiorno di una persona per una sola notte per gli hotel di Londra nell'aprile 2016. Si parla di prezzo medio poiché i prezzi presenti nel dataset analizzato corrispondono alla media tra i valori registrati per quattro settimane consecutive. Questa variabile, come tutte le altre descritte in precedenza, è stata inoltre studiata per un intervallo di tempo pari a due mesi e in particolare nei seguenti momenti: 60, 45, 30, 20, 10, 4, 1 giorno prima e lo stesso giorno prima del soggiorno. La variabile "prezzo" rappresenta anche la variabile di risposta Y per i modelli che si vanno a studiare ed analizzare.

Nella **Figura 2-18** sono riportati a scopo esemplificativo gli andamenti della variabile di risposta di sei hotel selezionati all'interno del dataset. Tra gli hotel del campione non si verifica uno stesso andamento per la variabile in esame, anzi si verificano situazioni differenti: per alcuni hotel, il prezzo aumenta con l'avvicinarsi della data del soggiorno, per altri invece diminuisce nell'orizzonte temporale considerato, per altri ancora resta costante o varia leggermente.

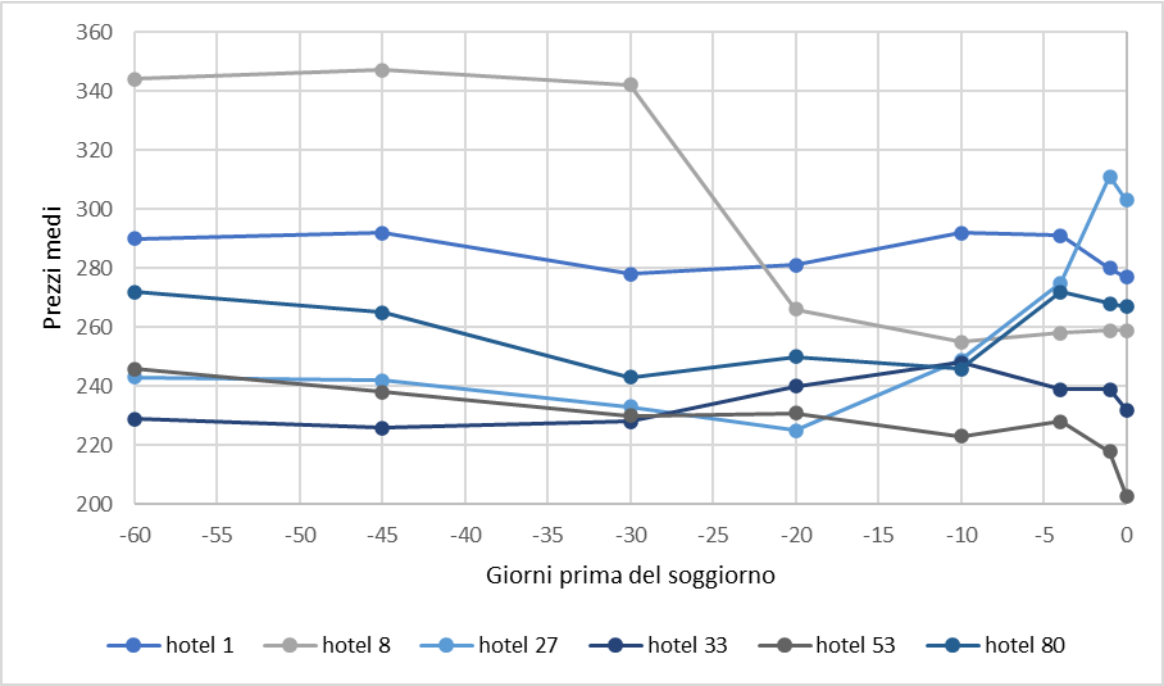

Figura 2-18: Andamenti della variabile prezzo.

Per avere un'analisi più completa, si è deciso di calcolare, per ogni hotel, la media tra i prezzi medi registrati negli otto istanti temporali considerati, suddividendoli poi in base al numero di stelle; sono stati in seguito calcolati gli indicatori presenti nella **Tabella 2-6** che sono visibili anche nella **Figura 2-19**, dove sono riportati i boxplot dei prezzi medi degli hotel osservati, suddivisi in tre gruppi, in base al numero di stelle.

	HOTEL 3 STELLE	HOTEL 4 STELLE	HOTEL 5 STELLE
VALORE MINIMO	116	147	249
PRIMO QUARTILE	139	196	343
MEDIANA	156	217	395
MEDIA	158	223	422
TERZO QUARTILE	183	250	488
VALORE MASSIMO	206	340	862

Tabella 2-6: Indicatori della variabile prezzo in base al numero di stelle.

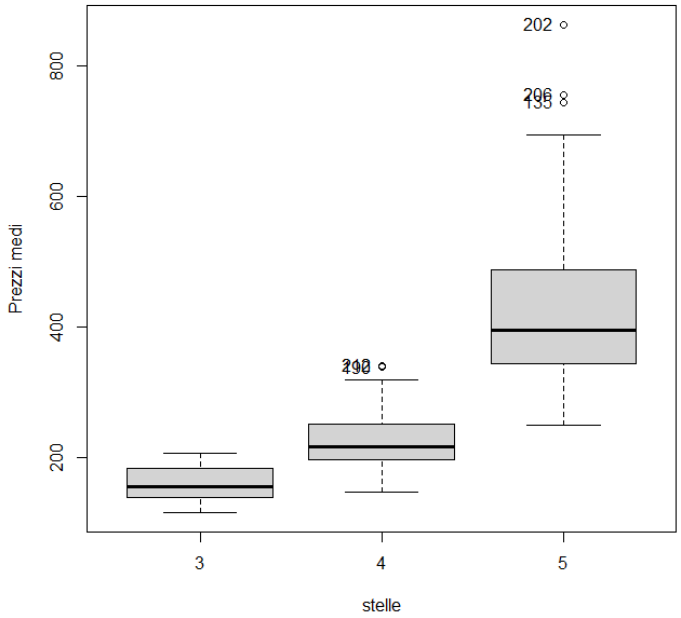

Figura 2-19: Boxplot dei prezzi medi degli hotel suddivisi in base al numero di stelle.

Confrontando i risultati, è possibile notare che i tre boxplot non sono perfettamente separati, ma per alcuni valori si sovrappongono, soprattutto per quanto riguarda i prezzi degli hotel a tre stelle e quelli degli hotel a quattro stelle. Considerando quindi che la mediana dei prezzi degli hotel a quattro stelle è pari a 217, significa che il 50% di questi hotel ha un prezzo medio compreso tra 147 e 217, ma anche la maggior parte degli hotel a tre stelle applica mediamente un prezzo compreso in questo range; infatti, il primo quartile di quest'ultima categoria è pari a 139 e il valore massimo è pari a 206.

Per quanto riguarda invece il confronto tra i prezzi medi degli hotel a quattro stelle e quelli degli hotel a cinque stelle, è possibile notare una piccola sovrapposizione tra i relativi boxplot, in quanto il primo quartile degli hotel a cinque stelle coincide con il valore massimo registrato per gli hotel a quattro stelle. Inoltre, sempre per queste due categorie di hotel, è possibile osservare la presenza di outliers, al di fuori dei limiti superiori, i quali indicano valori anomali che si discostano in modo significativo dalla maggior parte dei valori.

Le sovrapposizioni individuate implicano, quindi, l'esistenza di un'area del mercato in cui, a parità di prezzo, si hanno hotel con livelli di qualità differenti: questo potrebbe dipendere da diversi fattori quali eventuali sconti oppure presenza di elevata concorrenza. Considerando, invece, la differenza tra il valore massimo e il valore minimo, per i prezzi degli hotel a tre stelle è pari a 90 (= 206 – 116), per quelli degli hotel a quattro stelle è 194 (=340 – 147) e, infine, per

quelli degli hotel a cinque stelle è 613 (=862 – 249); questo significa che per gli alberghi a cinque stelle si ha una maggiore variabilità dei dati rispetto alle altre due categorie.

Poiché gli andamenti dei prezzi dei vari hotel sono differenti tra di loro, si è deciso di calcolare il coefficiente di variazione intertemporale, dato dal rapporto tra deviazione standard e media; questo coefficiente rappresenta un indice di dispersione che si utilizza per confrontare la variabilità dei dati rispetto alla loro media e si calcola rapportando la deviazione standard alla media. La distribuzione della densità del cv di questa variabile è rappresentata nella **Figura 2-20**, dove sono evidenziati tre punte, le quali indicano i tre gruppi di hotel suddivisi in base al numero di stelle.

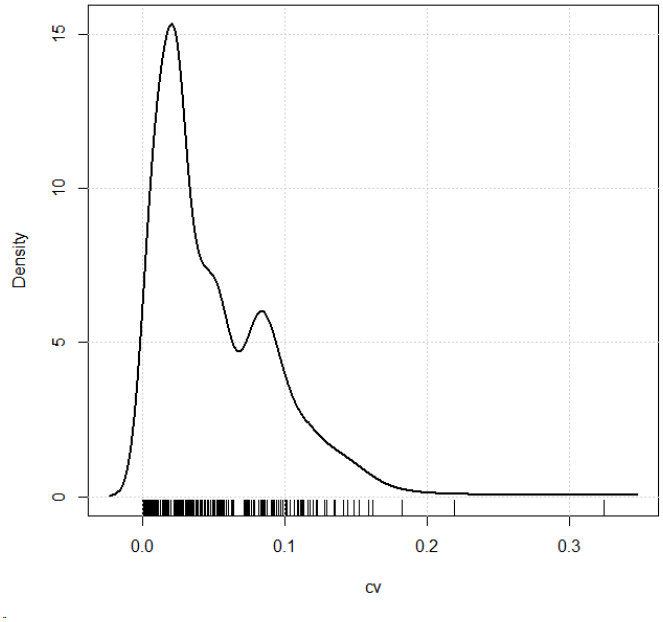

Figura 2-20: Distribuzione del coefficiente di variazione intertemporale del prezzo per i vari hotel.

È possibile anche osservare che la maggior parte degli hotel ha una variazione dei propri prezzi medi intorno alla relativa media inferiore allo 0.1, tuttavia si hanno hotel con una variazione prossima allo zero, questo significa che nel campione esiste almeno un hotel con prezzi medi che rimangono costati per l'intero orizzonte temporale osservato. Infatti, il minore coefficiente di variazione che si registra per questa variabile è pari a 0.000424: questo vuol dire che l'hotel che ha questo cv mantiene costanti i livelli di prezzo per l'intero periodo di tempo osservato. Nel campione in esame, questo albergo ha fissato i seguenti prezzi medi: €401 sessanta giorni prima, €416 quarantacinque giorni prima, €437 trenta giorni prima, €426 venti giorni prima,

€401 dieci giorni prima, €417 quattro giorni prima, €423 il giorno prima e, di nuovo, €417 lo stesso giorno del soggiorno.

Esiste infine un hotel i cui prezzi variano dello 0.3 intorno alla relativa media: ovvero significa che l'hotel in questione varia i propri prezzi medi fino al 30%, nell'orizzonte temporale considerato. Nel campione, infatti, questo hotel assume i seguenti prezzi: €167 sessanta giorni prima, €176 quarantacinque giorni prima, €181 trenta giorni prima, €164 venti giorni prima, €234 dieci giorni prima, €225 quattro giorni prima, €231 il giorno prima e €100 lo stesso giorno del soggiorno.

Sono stati successivamente utilizzati i coefficienti di variazione intertemporale della variabile "prezzo" per calcolare gli indicatori presenti nella **Tabella 2-7** che sono anche visibili nella **Figura 2-21**.

VALORE MINIMO	0.000424
PRIMO QUARTILE	0.018165
MEDIANA	0.038337
MEDIA	0.052052
TERZO QUARTILE	0.078829
VALORE MASSIMO	0.32437

Tabella 2-7: Indicatori del coefficiente di variazione intertemporale dei prezzi medi.

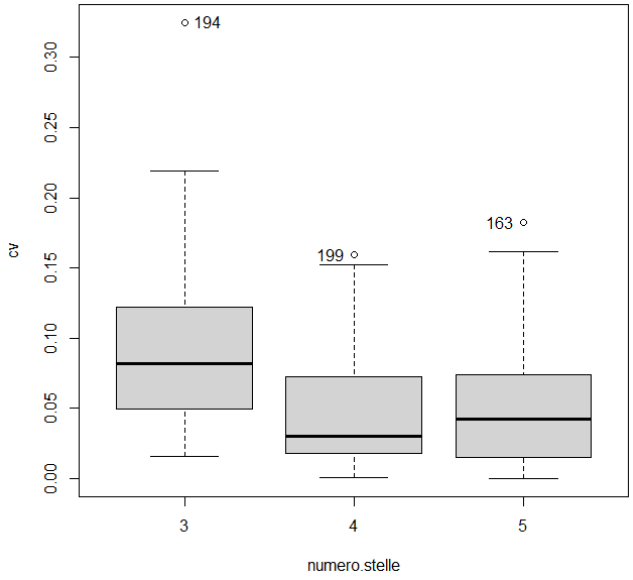

Figura 2-21: Boxplot del coefficiente di variazione intertemporale dei prezzi medi.

Il primo quartile è pari a 0.018, questo significa che il 25% degli hotel osservati ha una variabilità bassa dei prezzi rispetto alla relativa media; quindi, una parte del campione ha una variazione limitata del livello dei prezzi nell'intervallo di tempo considerato.

La mediana risulta essere pari a 0.038, il quale rappresenta il valore centrale tra i coefficienti di variazione calcolati. Inoltre, questo significa che, se si considera una stanza di un hotel, il cui prezzo medio sulla piattaforma Booking.com è pari a €100, allora il prezzo definitivo potrebbe

variare di circa 4 (= 100 * 0.038) e dunque tra €96 (= 100 – 4) e €104 (=100 + 4); ciò dipende dall'orizzonte temporale di prenotazione.

Considerando il terzo quartile significa che il 75% degli hotel ha un coefficiente di variazione intertemporale del prezzo al massimo pari a 0.0788, di conseguenza la maggior parte del campione ha una bassa variabilità dei dati rispetto alla media e quindi il livello dei prezzi non ha una variazione eccessiva per il periodo di tempo osservato.

Nella **Figura 2-21** appare che nel campione esaminato gli hotel, i cui prezzi medi variano maggiormente rispetto alla relativa media, sono quelli a tre stelle: questo conferma quanto riportato nell'articolo "Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model." (Wang, Sun, & Wen, Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model, 2019), dove si afferma che precedenti studi hanno dimostrato che gli hotel con un'elevata reputazione, come quelli che appartengono alle catene, dipendono meno dalla stagionalità.

2.2 ANALISI DELLA CORRELAZIONE TRA LE VARIABILI

Prima di procedere con la stima dei modelli, si analizza l'eventuale presenza di correlazione tra le variabili in esame, in modo da individuare ed escludere le variabili copie, ossia quelle che apportano lo stesso tipo di informazione sulla variabile di risposta. Per analizzare la correlazione sono stati utilizzati i valori delle variabili al t0, ossia il giorno stesso del soggiorno, tuttavia, si precisa che si registrano gli stessi comportamenti anche per tutti gli altri istanti temporali considerati.

Dall'analisi grafica, oltre all'elevata correlazione tra le variabili "pagina" e "posizione", evidenziata in precedenza (**Errore. L'origine riferimento non è stata trovata.**), risultano s trettamente correlate tra di loro le variabili "n_preferiri" e "n_rev", le quali sono riportate nella **Figura 2-22**. Di conseguenza, nei modelli stimati bisogna considerare una sola variabile per ognuna di queste due coppie: in particolare, per la prima coppia, si è deciso di considerare la variabile "posizione", in quanto il numero della pagina di ricerca dipende da quanti hotel si visualizzano per ogni pagina; per quanto riguarda la seconda coppia, invece, si è deciso di tener presente la variabile "n_preferiti", poiché può essere considerato come indice di qualità per gli hotel.

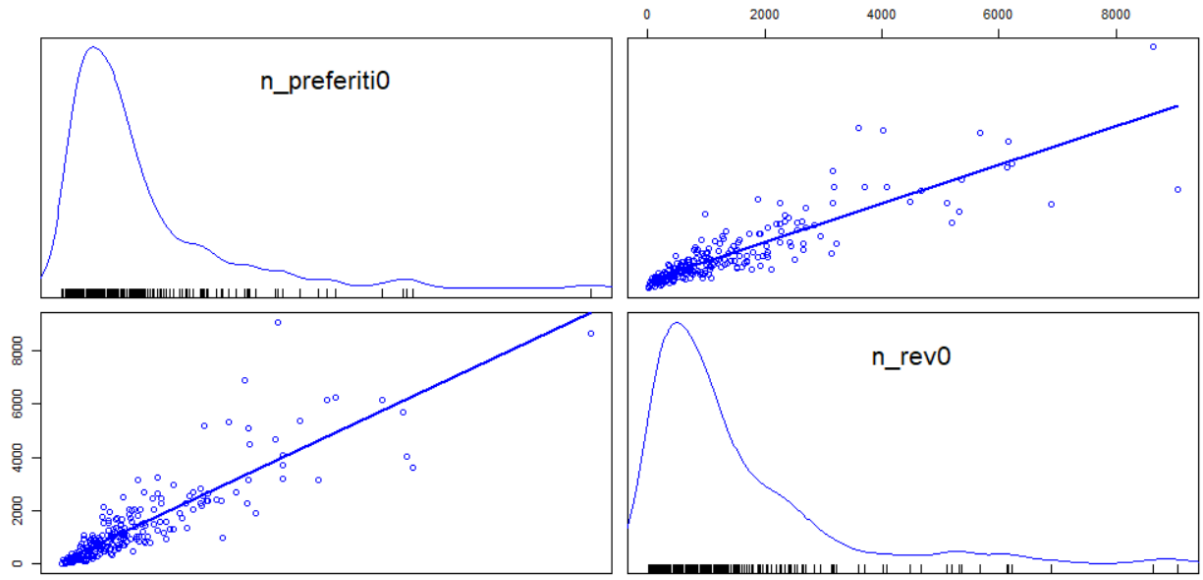

Figura 2-22: Matrice di correlazione con grafico a dispersione.

Nella **Figura 2-23** sono messe in relazione le variabili "n_preferiti" e "posizione" con la variabile di risposta "prezzo"; tra le due variabili esplicative si registra una tendenza negativa. Nonostante il numero di preferiti rappresenti il numero di utenti che hanno indicato uno

specifico hotel come preferito e quindi potrebbe rappresentare un indice di qualità, si evidenzia una tendenza negativa tra questa variabile e il prezzo, di conseguenza all'aumentare del numero di preferiti, il prezzo tende a diminuire. Tra le variabili "posizione" e "prezzo" invece non si osserva alcuna correlazione, infatti, la linea dei minimi quadrati risulta piatta.

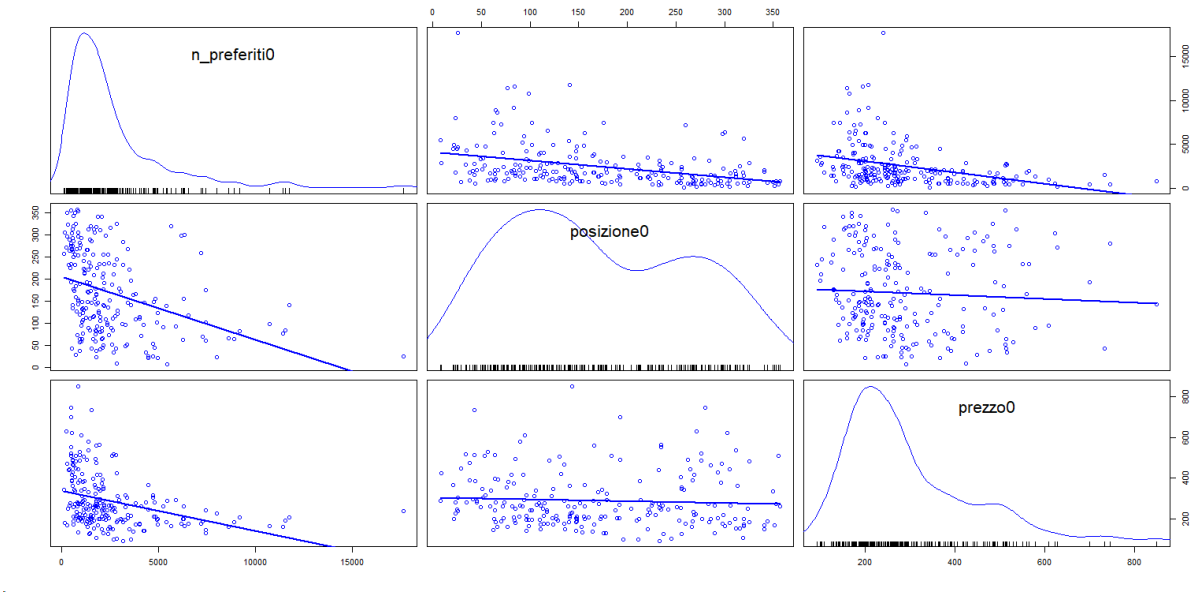

Figura 2-23: Matrice di correlazione con la variabile di risposta.

Si può valutare la presenza di correlazione anche da un punto di vista numerico attraverso la Matrice di Correlazione di Pearson (**Tabella 2-8**) e la Matrice di Correlazione Parziale (**Tabella 2-9**), quest'ultima permette di calcolare la correlazione tra le coppie di variabili, al netto dell'effetto delle altre variabili.

	n_preferiti0	n_rev0	pagina0	posizione0	prezzo0	valutazione0
n_preferiti0	1.0000	**0.8671**	-0.3642	-0.3674	-0.3573	-0.2116
n_rev0		1.0000	-0.2531	-0.2552	-0.3968	-0.2575
pagina0			1.0000	**0.9990**	-0.0625	0.0099
posizione0				1.0000	-0.0609	0.0095
prezzo0					1.0000	**0.5359**
valutazione0						1.0000

Tabella 2-8: Matrice di Correlazione di Pearson.

	n_preferiti0	n_rev0	pagina0	posizione0	prezzo0	valutazione0
n_preferiti0	0.0000	**0.8310**	0.0571	-0.0726	-0.1049	0.0557
n_rev0		0.0000	-0.0287	0.0337	-0.1061	-0.0718
pagina0			0.0000	**0.9988**	-0.0282	0.0333
posizione0				0.000	0.0183	-0.0307
prezzo0					0.0000	**0.4902**
valutazione0						0.000

Tabella 2-9: Matrice di Correlazione Parziale.

Per valutare se due variabili sono strettamente correlate tra di loro, si considera come valore soglia lo 0.5. Di conseguenza, anche da un punto di vista numerico, si confermano le correlazioni evidenziate attraverso i grafici presenti nelle **Errore. L'origine riferimento non è stata trovata.** e **Figura 2-22**: si registra dunque una forte correlazione tra le variabili "pagina" e "posizione" (correlazione = 0.99) e tra le variabili "n_preferiti" e "n_rev" (correlazione = 0.8).

Tuttavia, attraverso le matrici di correlazione, è possibile notare anche una tendenza positiva tra la variabile "valutazione" e la variabile di risposta "prezzo": questa tendenza riflette dunque la relazione positiva tra la qualità di uno specifico hotel percepita dagli utenti e il relativo prezzo.

CAPITOLO 3: DETERMINANTI DEL PREZZO, APPROCCIO NAÏF

In questo capitolo si vuole andare ad individuare le determinanti del prezzo attraverso un approccio naïf, ovvero si va a studiare come varia, negli istanti temporali considerati, l'impatto dei predittori (n_preferiti, posizione, stelle e valutazione) sulla variabile di risposta Y (prezzo). Per fare ciò, si è deciso di procedere con la realizzazione e l'analisi di otto modelli di regressione lineare, uno per ogni istante temporale osservato. Successivamente, si vuole riportare i valori dei coefficienti di una stessa variabile in un grafico a dispersione, in modo tale da studiarne l'andamento e capire quindi quando impatta maggiormente sul prezzo.

Il modello lineare che si va ad analizzare, per i diversi istanti di tempo osservati, è identificato dalla seguente equazione:

$$Y_{i(t)} = \beta_0 + \sum \beta_{j(t)} X_{ji(t)} + \varepsilon_{i(t)} \quad \textbf{(3.1)}$$

dove: $X_{ji(t)}$= {n_preferiti; posizione; stelle (dummy); valutazione} rappresenta l'insieme di predittori, $Y_{i(t)}$ è la variabile di risposta, ossia il prezzo e $\varepsilon_{i(t)}$ indica l'errore. Nell'equazione **(3.1)**, β_0 è una costante e rappresenta l'intercetta del modello, mentre $\beta_{j(t)}$ sono i coefficienti delle variabili esplicative per l'hotel i-esimo (con i = 1, …, 226) al tempo t (con t = 0, 1, 4, 10, 20, 30, 45, 60, ossia i giorni prima della data del soggiorno) e rappresentano quindi di quanto aumenta o diminuisce la variabile di risposta in seguito ad una variazione unitaria dei singoli predittori.

3.1 MODELLI DI REGRESSIONE LINEARE

In questa sezione si inizia con un'analisi dettagliata del primo modello che rappresenta la situazione finale ossia quella del giorno del soggiorno, per poi passare ad analizzare in maniera più rapida gli altri sette modelli e fare un commento generale di tutti i modelli analizzati, evidenziandone le caratteristiche comuni.

3.1.1 STIMA DEL MODELLO DI REGRESSIONE LINEARE PER LA DATA DEL SOGGIORNO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-106.943	76.919	-1.390	0.165828	
n_preferiti0	-0.0051	0.0024	-2.166	0.031410	*
stelle0 [T.4]	59.527	18.899	3.150	0.001860	**
stelle0 [T.5]	232.522	21.854	10.640	< 2e-16	***
valutazione0	34.538	9.554	3.615	0.000372	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 3-1: Regressione lineare a t0.

Il modello finale è quindi composto dal seguente insieme di predittori: X = {n_preferiti; stelle; valutazione}; in particolare, la variabile "stelle", essendo una variabile qualitativa, viene gestita da un'opportuna dummy a tre livelli: tre, quattro e cinque stelle; tuttavia, nei risultati ne appaiono solo due poiché il primo livello è implicitamente accettato dal modello.

Osservando i coefficienti stimati delle variabili quantitative, si può notare che la variabile "n_preferiti" impatta negativamente sul prezzo: infatti, se si aumentasse di un'unità il numero di preferiti, allora il prezzo diminuirebbe, in media, di 0.0051, ferme restando le altre variabili; la variabile "valutazione", invece, impatta positivamente sul prezzo: infatti, se si aumentasse di un'unità la variabile "valutazione", il prezzo aumenterebbe, in media, di 34.54, ferme restando le altre variabili.

Con riferimento alla variabile qualitativa "stelle", invece, si può notare come varia il prezzo degli hotel a quattro e a cinque stelle rispetto al prezzo degli hotel a tre stelle, che rappresenta il livello base implicitamente assunto dal modello; in particolare, se si trattasse di un hotel a

quattro stelle, allora, il prezzo aumenterebbe, in media, di 59.527 rispetto al prezzo di un hotel tre stelle, mentre se si trattasse di un hotel cinque stelle, allora, il prezzo sarebbe, in media, superiore di 232.522, rispetto a quello di un hotel a tre stelle.

Considerando le statistiche t e i corrispondenti p-value, si può notare che il modello stimato comprende solo le variabili significative, ovvero quelle con una statistica t maggiore di due e il corrispondente p-value minore di 0.05. Inoltre, il modello risulta nel suo complesso significativo, infatti, la statistica F è pari a 103 (>2) e il relativo p-value è molto basso (< 2.2e-16).

Il coefficiente di determinazione R-quadro o Multiple R-squared è pari a 0.6509: questo significa che il modello è in grado di spiegare il 65.09% della variabilità della risposta sul campione. L'R-quadro aggiustato o Adjusted R-squared, invece, è pari a 0.6446 e permette di fare valutazioni sulla bontà di adattamento del modello fuori campione.

Le seguenti statistiche: AIC, "Akaike Information Criterion" e BIC, "Bayesian Information Criterion", rappresentano delle misure di errore che vengono utilizzate per la selezione tra più modelli; in particolare, per questo modello, le due statistiche sono pari rispettivamente a 2627.739 e 2648.262.

Per completare l'analisi del modello, si procede con l'implementazione delle seguenti analisi diagnostiche: Influence Plot, studio della collinearità ed analisi dei residui.

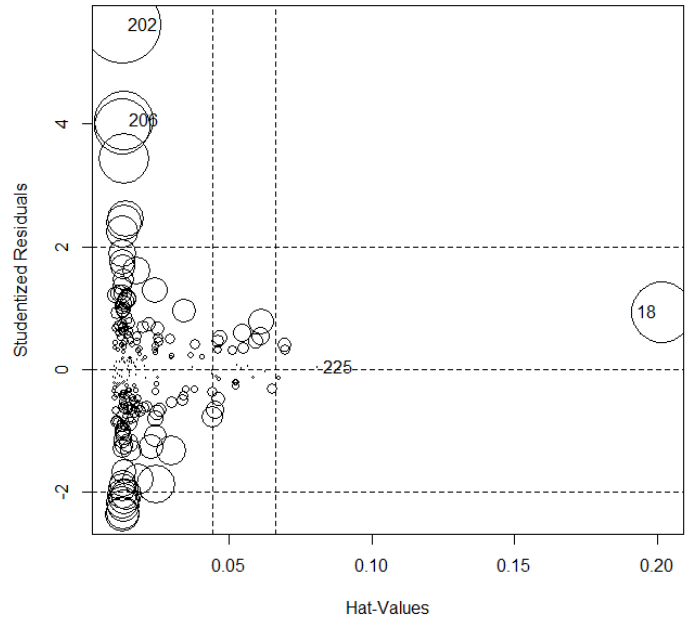

Figura 3-1: Influence Plot a t0.

L'Influence Plot (**Figura 3-1**) permette di identificare i punti anomali presenti nel dataset, ovvero i cosiddetti ouliers, dati troppo distanti dalla media della variabile di risposta, e i punti con High Leverage (HLP), valori di X troppo lontani dalla media di tali predittori. Si possono identificare come punti di leverage cattivo, quei punti che hanno un residuo maggiore di più o meno due e con un leverage medio elevato; in questo caso, non sono presenti "HLP cattivi" e quindi non vi sono osservazioni da eliminare dal dataset.

Calcolando il Fattore di Inflazione della Varianza (VIF) per lo studio della collinearità, si può osservare nella **Tabella 3-2** che le tre variabili rimaste non sono correlate tra di loro, infatti, il VIF è minore di due, ovvero il valore soglia per calcolare la correlazione.

	GVIF	Df	GVIF^(1/(2*Df))
n_preferiti0	1.152897	1	1.073730
stelle0	1.547735	2	1.115383
valutazione0	1.411078	1	1.187888

Tabella 3-2: Fattore di Inflazione della Varianza.

Una delle ipotesi standard della regressione lineare è la normalità dei residui: questi ultimi, i quali rappresentano le differenze tra i valori osservati e quelli predetti dal modello, devono seguire una distribuzione normale. Nel grafico in alto della **Figura 3-2** è rappresentato uno scatterplot che confronta i percentili osservati degli errori studentizzati e quelli che si avrebbero se ci fosse normalità; poiché la nuvola di punti è disposta lungo la diagonale, allora significa che l'ipotesi di normalità è compatibile coi dati. Dal secondo grafico, inoltre, si può dedurre dalla linea piatta che il modello è ben specificato; tuttavia, si può notare la presenza di due clusters, rappresentati da due nuvole di punti ben separate, i quali erano già stati apparsi nell'analisi della variabile "n_preferiti".

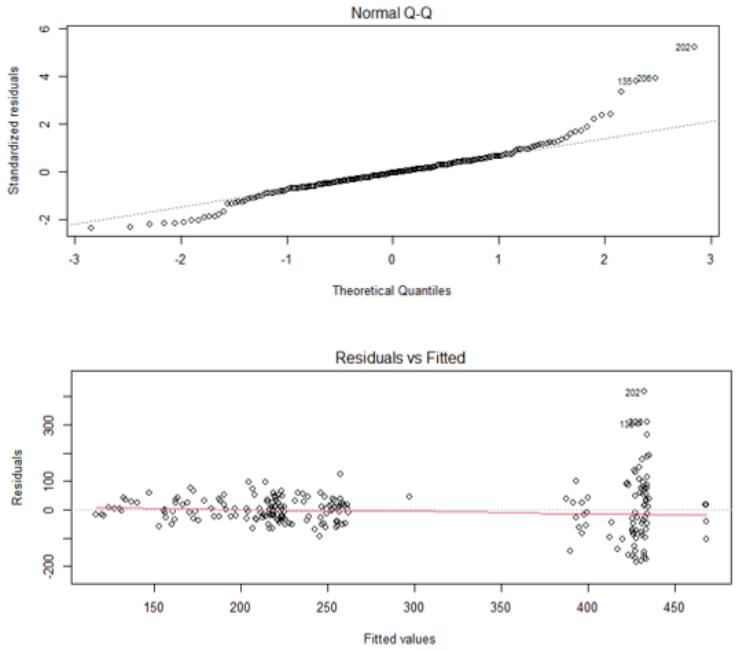

Figura 3-2: Grafici diagnostici di base.

3.1.2 STIMA DEL MODELLO DI REGRESSIONE LINEARE PER GLI ALTRI ISTANTI TEMPORALI

UN GIORNO PRIMA DEL SOGGIORNO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-94.422	77.272	-1.222	0.223033	
n_preferiti1	-0.0052	0.0024	-2.215	0.027780	*
stelle1 [T.4]	51.864	18.879	2.747	0.006508	**
stelle1 [T.5]	226.359	21.902	10.335	< 2e-16	***
valutazione1	34.328	9.598	3.576	0.000428	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 3-2: Regressione lineare a t1.

L'R-quadro è pari a 0.6448: il modello spiega il 64.48% della variabilità del prezzo; mentre l'R-quadro aggiustato è pari a 0.6448. Il modello risulta, nel suo complesso, significativo poiché la statistica F è pari a 103.1 (>2) con un p-value < 2.2e-16. Le statistiche AIC e BIC sono rispettivamente pari a 2627.05 e 2647.57.

QUATTRO GIORNI PRIMA DEL SOGGIORNO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-348.774	87.195	-4.000	8.64e-05	***
n_preferiti4	-0.0061	0.0024	2.516	0.0126	*
stelle4 [T.4]	39.814	19.375	2.055	0.0411	*
stelle4 [T.5]	199.811	22.9006	8.725	6.52e-16	***
valutazione4	67.0303	10.858	6.173	3.16e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 3-3: Regressione lineare a t4.

Il modello, oltre a spiegare il 67.17% della variabilità della risposta (R-quadro = 0.6717 e R-quadro aggiustato = 0.6657), risulta significativo nel suo complesso, infatti, la statistica F è pari a 113 (> 2) con un p-value < 2.2e-16. Le statistiche AIC e BIC sono rispettivamente pari a 2638.193 e 2658.716.

DIECI GIORNI PRIMA DEL SOGGIORNO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-155.314	76.628	-2.027	0.04388	*
n_preferiti10	-0.0067	0.0024	-2.812	0.00537	**
stelle10 [T.4]	40.575	18.827	2.155	0.03223	*
stelle10 [T.5]	204.063	21.773	21.773411	< 2e-16	***
valutazione10	42.875	9.518	4.505	0.0000108	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 3-4: Regressione lineare a t10.

Poiché l'R-quadro è pari a 0.6468, il modello spiega il 64.68% della variabilità del prezzo; l'R-quadro aggiustato, invece, è pari a 0.6405. Il modello nel complesso è significativo, infatti, la statistica F è pari a 101.2 (> 2) con un p-value < 2.2e-16. Le statistiche AIC e BIC sono rispettivamente pari a 2626.039 e 2646.562.

VENTI GIORNI PRIMA DEL SOGGIORNO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-159	76.283	-2.084	0.03828	*
n_preferiti20	-0.0065	0.0024	-2.733	0.00678	**
stelle20 [T.4]	41.014	18.777	2.184	0.03000	*
stelle20 [T.5]	199.318	21.684	9.192	< 2e-16	***
valutazione20	43.389	9.476	4.579	0.0000078	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 3-5: Regressione lineare a t20.

Il modello, oltre a spiegare il 63.69% della variabilità della risposta (R-quadro = 0.6369 e R-quadro aggiustato = 0.6303), risulta complessivamente significativo (statistica F = 96.92 > 2 con un p-value < 2.2e-16). Le statistiche AIC e BIC sono rispettivamente pari a 2625.515 e 2646.038.

TRENTA GIORNI PRIMA DEL SOGGIORNO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-122.804	75.289	-1.631	0.10429	
n_preferiti30	-0.0066	0.0024	-2.750	0.00645	**
stelle30 [T.4]	39.257	18.53	2.119	0.03524	*
stelle30 [T.5]	197.761	21.402	9.240	< 2e-16	***
valutazione30	38.944	9.352	4.164	0.0000448	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 3-6: Regressione lineare a t30.

L'R-quadro è pari a 0.6332: il modello spiega il 63.32% della variabilità della risposta, ossia la variabilità del prezzo; l'R-quadro aggiustato, invece, è pari a 0.6266. Il modello è nel complesso significativo: la statistica F è pari a 95.39 (> 2) con un p-value < 2.2e-16; le statistiche AIC e BIC sono rispettivamente uguali a 2619.536 e 2640.06.

QUARANTACINQUE GIORNI PRIMA DEL SOGGIORNO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-120.909	74.436	-1.624	0.1057	
n_preferiti45	-0.006	0.0024	-2.477	0.0140	*
stelle45 [T.4]	41.721	18.32	2.277	0.0237	*
stelle45 [T.5]	201.983	21.111	9.567	< 2e-16	***
valutazione45	38.355	9.237	4.152	0.000047	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 3-7: Regressione lineare a t45.

Il modello, oltre a spiegare il 63.79% della variabilità della risposta (R-quadro = 0.6379 e R-quadro aggiustato = 0.6314), è complessivamente significativo (statistica F = 97.33 > 2 con un p-value < 2.2e-16). Le statistiche AIC e BIC sono rispettivamente pari a 2614.971 e 2635.494.

SESSANTA GIORNI PRIMA DEL SOGGIORNO

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-98.225	73.305	-1.340	0.181635	
n_preferiti60	-0.0065	0.0025	-2.656	0.008490	**
stelle60 [T.4]	42.815	18.209	2.351	0.019583	*
stelle60 [T.5]	204.59	20.858	9.808	< 2e-16	***
valutazione60	36.021	9.0933	3.961	0.000101	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 3-8: Regressione lineare a t60.

L'R-quadro è pari a 0.6397: il modello spiega il 63.97% della variabilità del prezzo; l'R-quadro aggiustato, invece, è pari a 0.6332. Il modello risulta nel complesso significativo, infatti, la statistica F è uguale a 98.09 (> 2) con un p-value < 2.2e-16. Le statistiche AIC e BIC sono rispettivamente pari a 2612.59 e 2633.114.

3.1.2 CONFRONTO TRA I MODELLI DI REGRESSIONE LINEARE STIMATI

In generale, si può affermare che le variabili "n_preferiti", "stelle" (dummy) e "valutazione" sono sempre significative (da sessanta giorni prima al giorno stesso del soggiorno), inoltre, la variabile "n_preferiti" influenza sempre la variabile di risposta, il "prezzo", in maniera negativa, mentre le altre due variabili impattano positivamente sul prezzo delle camere degli hotel. Oltre alle singole variabili evidenziate, anche i vari modelli risultano complessivamente significativi: infatti, presentano sempre una statistica F maggiore di due, considerato il valore soglia, e un p-value minore dello 0.05. I modelli stimati presentano inoltre un R-quadro piuttosto elevato, compreso tra lo 0.6332 e lo 0.6717.

Come già anticipato all'inizio del capitolo, le statistiche AIC e BIC rappresentano due misure di errore e vengono utilizzate per fare un confronto tra i modelli; in questo caso, il modello migliore, ossia quello con un valore minore per entrambe le statistiche è il modello che rappresenta la situazione a sessanta giorni prima del soggiorno.

Considerando il Fattore di Inflazione della Varianza, invece, non si registra, mai, alcuna collinearità tra le variabili esaminate. I modelli risultano, inoltre, tutti ben specificati e non presentano alcuno punto di leva cattivo (HLP cattivi) da dover eliminare.

Si analizza ora come variano, nell'orizzonte temporale considerato, i coefficienti dei modelli di regressione lineare stimati, attraverso la realizzazione di grafici a dispersione.

3.3 COEFFICIENTI DELLE SINGOLE VARIABILI NEGLI ISTANTI TEMPORALI OSSERVATI

	-60	-45	-30	-20	-10	-4	-1	0
n_preferiti	-0.0065	-0.0060	-0.0066	-0.0065	-0.0067	-0.0061	-0.0052	-0.0051
stelle [T.4]	42.82	41.72	39.26	41.01	40.58	39.81	51.86	59.53
stelle [T.5]	204.59	201.98	197.76	199.32	204.06	199.81	226.36	232.52
valutazione	36.02	38.36	38.94	43.39	42.88	67.03	34.33	34.54

Tabella 3-9: Stime dei coefficienti delle variabili significative.

Nella **Tabella 3-9** sono riportati in ordine temporale i valori dei coefficienti delle variabili significative ("n_preferiti", "stelle[T.4]", "stelle[T.5]" e "valutazione") assunti negli otto modelli analizzati, uno per ogni istante temporale considerato. Tali valori sono stati inseriti in quattro grafici a dispersione, uno per ogni variabile, in modo tale da poter analizzare come variano nell'intervallo considerato.

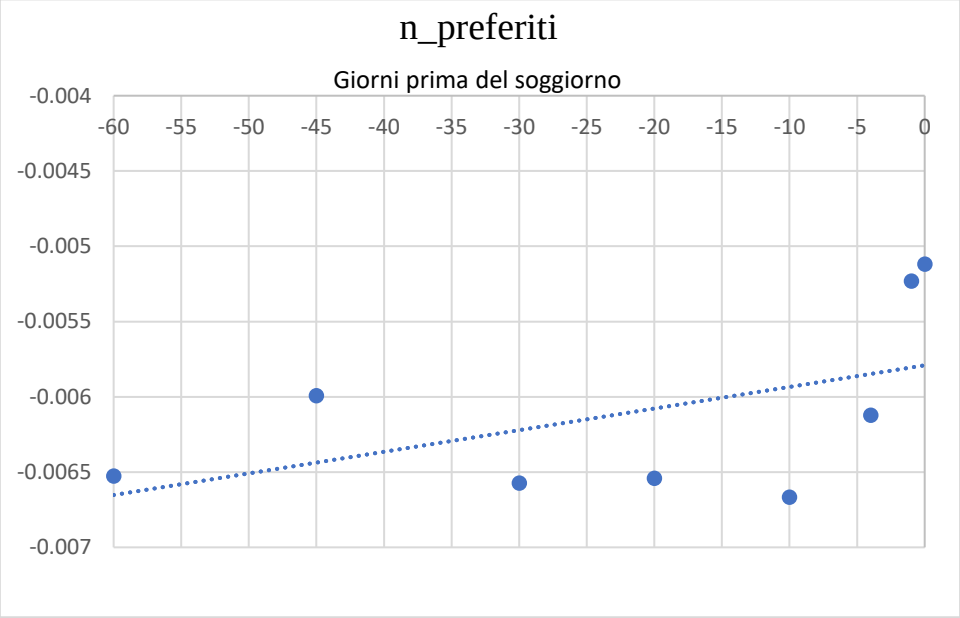

Figura 3-3: Andamento variabile n_preferiti.

Nella **Figura 3-3** sono rappresentati, attraverso un grafico a dispersione, i valori dei coefficienti della variabile "n_preferiti" assunti negli otto modelli analizzati. Si può notare che, nell'intervallo temporale considerato, questa variabile impatta sempre negativamente sul prezzo, tuttavia, gli impatti che si registrano sono di lieve entità: infatti, la maggiore variazione negativa che si osserva è pari a 0.0067 e avviene dieci giorni prima del soggiorno e la minore variazione, invece, si registra il giorno del soggiorno (-0.0051). In generale, si osserva una tendenza positiva, ovvero l'impatto di questa variabile sul prezzo diminuisce con l'avvicinarsi della data del soggiorno; tuttavia, è interessante osservare che fino a quattro giorni prima del soggiorno, questa variabile assume valori compresi tra (-0.006) e (-0.0067), invece, il giorno prima e il giorno del soggiorno assume valori intorno allo (-0.005).

Per quanto riguarda la variabile "stelle", essendo una variabile qualitativa a tre livelli (tre, quattro e cinque), viene trasformata in una dummy; di conseguenza, un livello è implicitamente assunto nei modelli e, per questo motivo, nei risultati appaiono solo gli altri due livelli. In generale, si può affermare che la variabile stelle impatta positivamente sul prezzo, ovvero gli hotel con quattro e cinque stelle hanno un prezzo maggiore degli hotel a tre stelle.

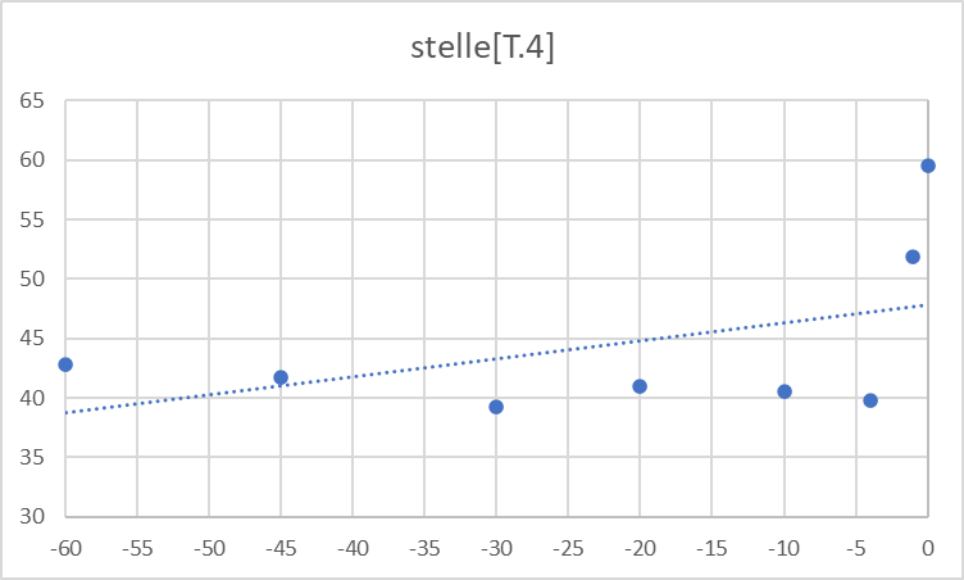

Figura 3-4: Andamento della variabile stelle [T.4].

Nella **Figura 3-4**, si può osservare come variano i prezzi degli hotel a quattro stelle rispetto a quelli degli hotel a tre stelle nell'orizzonte temporale considerato, ferme restando le altre variabili. In generale, si registra una tendenza positiva, ovvero si ha un impatto maggiore sul

prezzo con l'avvicinarsi della data del soggiorno; infatti, fino a quattro giorni prima del soggiorno, la differenza tra i prezzi degli hotel a quattro stelle e quelli degli hotel a tre stelle varia da €39 a €43, invece, il giorno prima e il giorno del soggiorno, si nota una differenza maggiore di €50: in particolare, il prezzo per una notte alla data del soggiorno di un hotel a quattro stelle supera, in media, di €60 il prezzo di un hotel a tre stelle. È interessante osservare che nonostante la tendenza crescente, la differenza media minore, pari a €39, tra i prezzi degli hotel a quattro stelle e quelli degli hotel a cinque stelle si registra a trenta e quattro giorni prima.

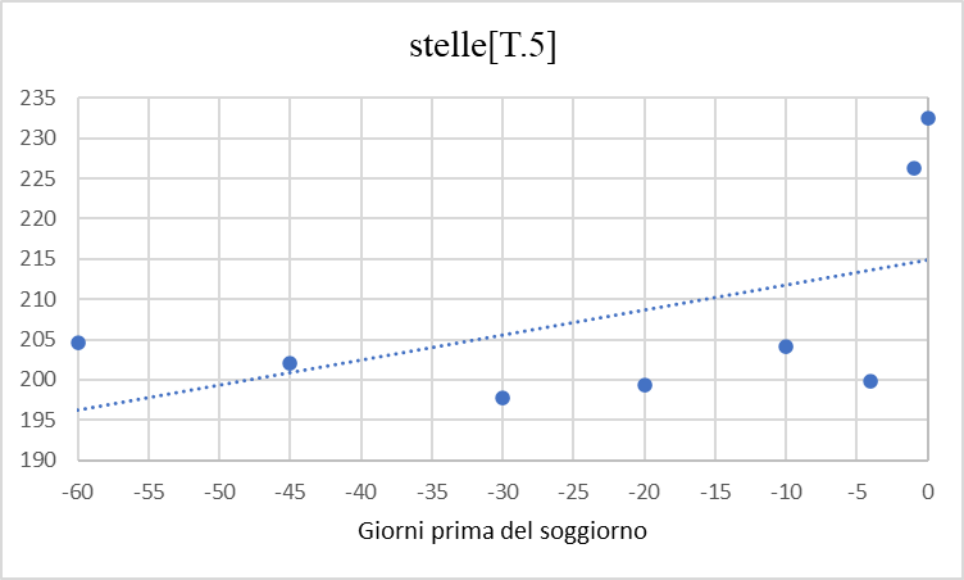

Figura 3-5: Andamento della variabile stelle [T.5].

Nella **Figura 3-5** si osservano i coefficienti della variabile "stelle" per il livello cinque. In generale, si osserva una tendenza crescente molto più marcata rispetto a quella del livello quattro, tuttavia si nota un comportamento simile, infatti, in media si registra una differenza minore, tra i prezzi degli hotel a cinque stelle e quelli degli hotel a tre stelle, a trenta e a quattro giorni prima, rispettivamente pari a €197 e €199. Come accade per le altre due variabili, anche in questo caso, i valori assunti fino a quattro giorni prima del soggiorno sono minori rispetto a quelli assunti il giorno prima e il giorno del soggiorno, ma in questo caso si verifica una differenza maggiore rispetto ai casi precedenti, infatti, per la maggior parte dell'orizzonte temporale osservato, la differenza dei prezzi degli hotel a cinque stelle e quelli degli hotel a tre stelle varia da €197 a €205, invece, per gli ultimi istanti studiati la differenza è, in media, superiore a €225. Di conseguenza, se si prenota uno stesso tipo di camera trenta giorni prima

del soggiorno per un hotel a cinque stelle si paga, in media, €197 in più rispetto ad un hotel a tre stelle, invece, se si prenota lo stesso giorno del soggiorno la camera dell'hotel a cinque stelle, in media, costa fino a €233 in più di quella a tre stelle.

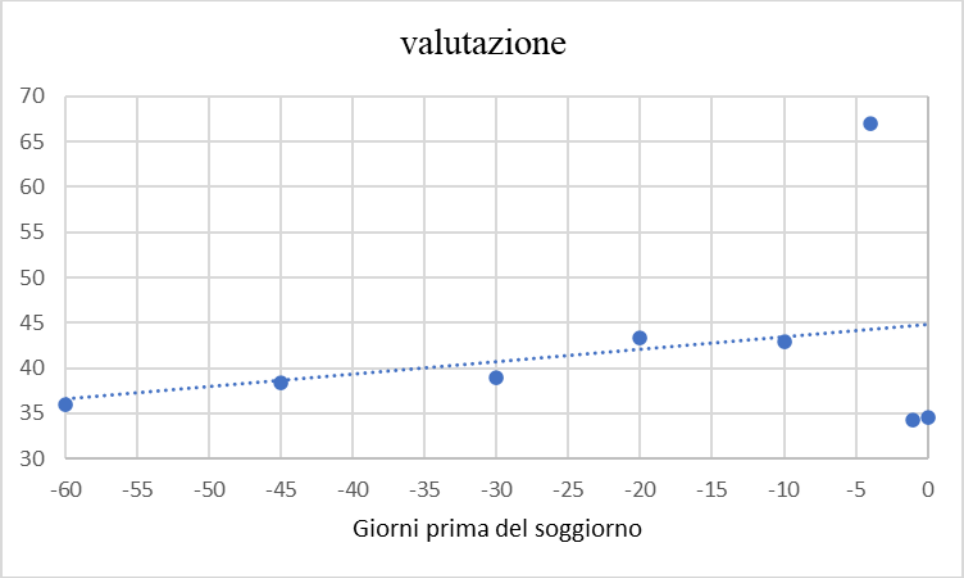

Figura 3-6: Andamento della variabile valutazione.

La variabile "valutazione" (**Figura 3-6**) presenta una tendenza positiva, ovvero l'impatto cresce nel tempo, tuttavia, a differenza delle altre variabili analizzate, si può osservare che in questo caso l'effetto minore si registra proprio il giorno prima e il giorno del soggiorno, infatti, se si aumenta di un'unità questa variabile, per i giorni indicati, si ha mediamente un aumento rispettivamente pari a €34.3 e €34.5, ferme restando le altre variabili. L'impatto maggiore, invece, si verifica quattro giorni prima del soggiorno, in cui se si aumenta di un'unità questa variabile, allora, in media, il prezzo di una camera di un hotel a tre stelle aumenta di €67, ferme restando le altre variabili; se, invece, si trattasse di un hotel a quattro stelle o di un hotel a cinque stelle, allora il prezzo in media aumenterebbe rispettivamente di €107 (= 67 + 40) e di € 267 (= 67 + 200), rispetto al prezzo di un hotel a tre stelle.

CAPITOLO 4: STIMA DEL MODELLO CON RIDUZIONE DI DIMENSIONALITA'

L'obiettivo di questo lavoro, come già anticipato in precedenza, è quello di studiare i prezzi del settore alberghiero per capire da quali fattori presenti sulla piattaforma Booking.com sono influenzati maggiormente. Nel capitolo precedente, infatti, attraverso un approccio naïf sono stati realizzati otto modelli di regressione lineare, uno per ogni istante temporale osservato, da sessanta giorni prima al giorno stesso del soggiorno. Da questa analisi è emerso che i prezzi delle camere degli hotel sono influenzati significativamente dalle variabili "numero di preferiti", "numero di stelle" e "valutazione"; inoltre, gli otto modelli stimati presentano un R-quadro superiore a 0.6.

Si vuole ora costruire un unico modello che comprenda tutte le variabili in esame, dove la variabile di risposta, ossia il prezzo della data del soggiorno, venga messa in relazione sia con le componenti autoregressive, ovvero i prezzi registrati negli istanti temporali precedenti, che con le altre variabili, anch'esse osservate nel tempo. Il modello in esame è identificato dall'equazione **(4.1)**:

$$Y_i = \beta_0 + \sum_j \beta_j X_{ij} + \varepsilon_i \qquad \textbf{(4.1)}$$

dove Y_i indica la variabile di risposta, mentre X_{ij} rappresenta:

- la variabile "prezzo", in particolare i prezzi osservati negli istanti temporali precedenti, da sessanta a un giorno prima;
- la variabile qualitativa "stelle", la quale viene considerata al tempo zero, ossia alla data del soggiorno poiché rimane costante per l'intero orizzonte temporale osservato;
- la variabile "valutazione" al tempo zero in quanto anch'essa non varia nel tempo;
- la variabile "n_preferiti" considerata in tutti gli otto istanti temporali analizzati;
- la variabile "posizione" osservata per l'intero orizzonte temporale, da sessanta giorni prima al giorno stesso del soggiorno;

infine, ε_i indica l'errore.

Tale modello però è caratterizzato da variabili altamente correlate tra di loro, quindi per poter gestire questo problema, si procede in due modi differenti ed alternativi: per cominciare si applica una riduzione di dimensionalità con le componenti principali (capitolo quattro), e successivamente si effettua una selezione delle variabili rilevanti con un metodo shrinkage, il quale permette di aggiungere, alle stime dei minimi quadrati, una penalità che riduce a zero i coefficienti delle variabili che sono altamente correlate con le altre (capitolo cinque).

In questo capitolo si procede in primo luogo con un'introduzione teorica dei metodi di riduzione delle dimensioni, soffermandosi in particolare sulla regressione con le componenti principali; successivamente si procede con l'applicazione di tale metodo al dataset in esame, per poi confrontare i risultati ottenuti con quelli dei modelli precedenti.

4.1 REGRESSIONE DELLE COMPONENTI PRINCIPALI

Per l'introduzione teorica relativa alla regressione delle componenti principali si analizzano i seguenti testi: "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" (Hastie, Tibshirani, & Friedman, 2016) e "An Introduction to Statistical Learning with Applications in R" (James, Witten, Hastie, & Tibshirani, 2023).

L'analisi delle componenti principali (PCA) è un approccio molto utilizzato per ricavare da un ampio insieme di variabili, un insieme di minori dimensioni; la PCA infatti è una tecnica per ridurre la dimensione di una matrice di dati (n x p) X, dove n indica il numero di osservazioni e p il numero di variabili. Tutti i metodi di riduzione delle dimensioni funzionano in due fasi: in primo luogo, si ottengono i predittori trasformati: Z_1, Z_2, …, Z_M (con M < p) e successivamente il modello viene stimato utilizzando gli M predittori ricavati.

L'idea chiave che sta alla base dell'approccio di regressione delle componenti principali (PCR) è che spesso un piccolo numero di componenti principali è sufficienti a spiegare la maggior parte della variabilità dei dati, nonché la relazione con la risposta. Se l'ipotesi alla base della PCR è valida, allora la stima del modello ai minimi quadrati con le componenti principali Z_1, Z_2, …, Z_M fornisce risultati migliori rispetto alla stima del modello ai minimi quadrati con i predittori originari X_1, X_2, …, X_p, poiché la maggior parte o la totalità delle informazioni presenti nei dati che si riferiscono alla risposta è contenuta in Z_1, Z_2, …, Z_M e quindi stimando solo i coefficienti M « p si può ridurre l'overfitting.

Considerando una matrice dati costituita da n osservazioni e p predittori, le componenti principali Z_1, Z_2, …, Z_M sono costruite come combinazioni lineari dei p predittori originali, ovvero:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

(4.2)

per alcune costanti ϕ_{1m}, ϕ_{2m}, …, ϕ_{pm} e m = 1, …, M. In seguito, utilizzando i minimi quadrati si procede con la stima del modello di regressione lineare che al posto dei predittori originali presenta le componenti principali:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m Z_{im} + \varepsilon_i \qquad i = 1, \dots, n$$

(4.3)

Si noti che nell'equazione **(4.3)** i coefficienti di regressione sono dati da θ_0, θ_1, …, θ_M. Se le costanti ϕ_{1m}, ϕ_{2m}, …, ϕ_{pm} sono scelte con attenzione, allora gli approcci di riduzione delle dimensioni, come la regressione delle componenti principali, possono essere migliori della regressione lineare ai minimi quadrati.

Il numero di componenti principali M viene tipicamente scelto mediante la tecnica della cross-validation. In genere quando si esegue la PCR, prima di generare le componenti principali, si consiglia di standardizzare ogni predittore: la standardizzazione assicura che tutte le variabili abbiano lo stesso ordine di grandezza, mentre in assenza di standardizzazione le variabili ad alta varianza tendono a giocare un ruolo maggiore nelle componenti principali ottenute, e la scala su cui le variabili sono misurate avrà un effetto sul modello finale di PCR. Tuttavia, se le variabili fossero tutte misurate nelle stesse unità si potrebbe scegliere di non standardizzarle.

4.2 APPLICAZIONE DELL'ANALISI DELLE COMPONENTI PRINCIPALI AL CASO IN STUDIO

Con riferimento al dataset in esame, osservando la **Tabella 4-1**, è possibile notare la percentuale di variabilità spiegata dalle singole componenti principali, definite come Z_1, Z_2, ecc. In particolare, è possibile osservare che la prima componente spiega da sola il 52.86% della varianza totale nei dati, mentre la seconda spiega circa il 23.73% e la terza il 13.42%; invece, queste tre componenti principali, insieme, spiegano il 90% della varianza totale, quindi, possono bastare queste tre componenti principali per comprendere le informazioni principali contenute nel dataset.

	EIGENVALUE	PROPORTION %	CUMULATIVE
Z_1	17.97	52.86	52.86
Z_2	8.066	23.72	76.58
Z_3	4.57	13.43	90.012
Z_4	1.19	3.503	93.52
Z_5	0.82	2.42	95.93

Tabella 4-1: Componenti principali.

4.2.1 COMPOSIZIONE DELLE COMPONENTI PRINCIPALI OTTENUTE

Per poter interpretare quali variabili rientrano principalmente nella composizione delle componenti principali (PC) conviene osservare le figure: **Figura 4-1**, **Figura 4-2** e **Figura 4-3**, ma anche i valori presenti nella **Tabella 4-2**, dalla quale è possibile osservare che le variabili che hanno un impatto maggiore in ogni PC sono il "numero di stelle" e la "valutazione". In particolare, per tutte e tre le PC si evidenzia una contrapposizione tra la variabile "valutazione" insieme al livello cinque della variabile qualitativa "stelle" con i livelli tre e quattro della variabile "stelle". Più precisamente, si può osservare che le prime due rientrano negativamente nella prima e nella seconda componente principale, mentre i livelli tre e quattro stelle rientrano positivamente; nella terza PC, invece, la situazione è opposta, ovvero "valutazione" e il livello cinque stelle hanno segno positivo, mentre i livelli tre e quattro della variabile "stelle" impattano negativamente. Si registra quindi una correlazione positiva tra la valutazione e il livello più alto della variabile qualitativa; questa relazione è stata già evidenziata nella sezione 2.1.3.

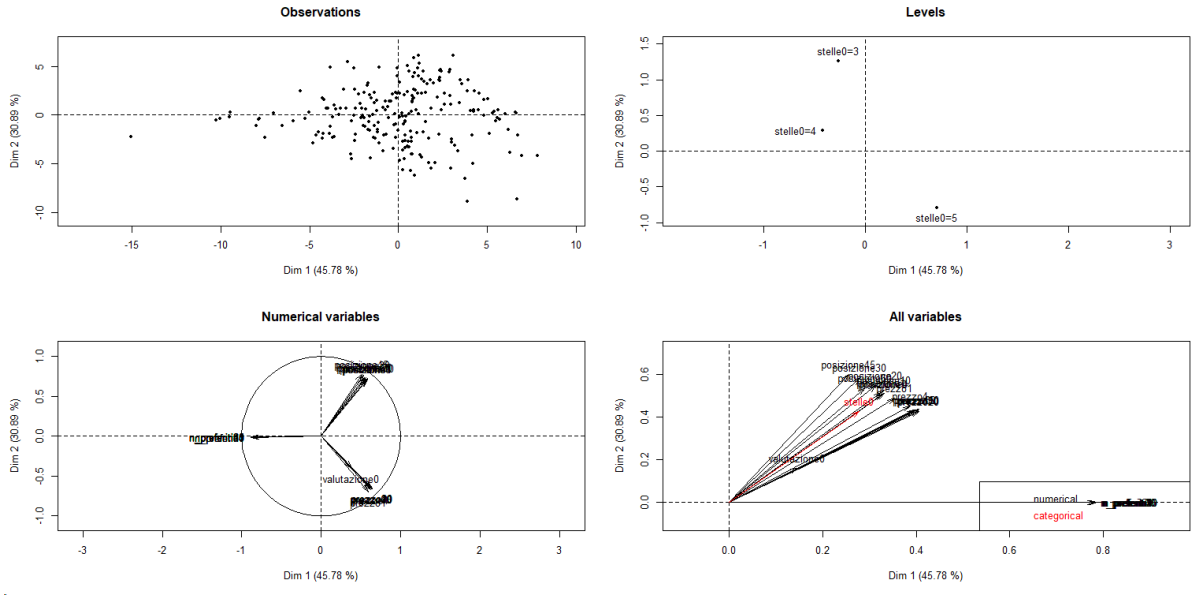

Figura 4-1: Componente principale 1 e Componente principale 3.

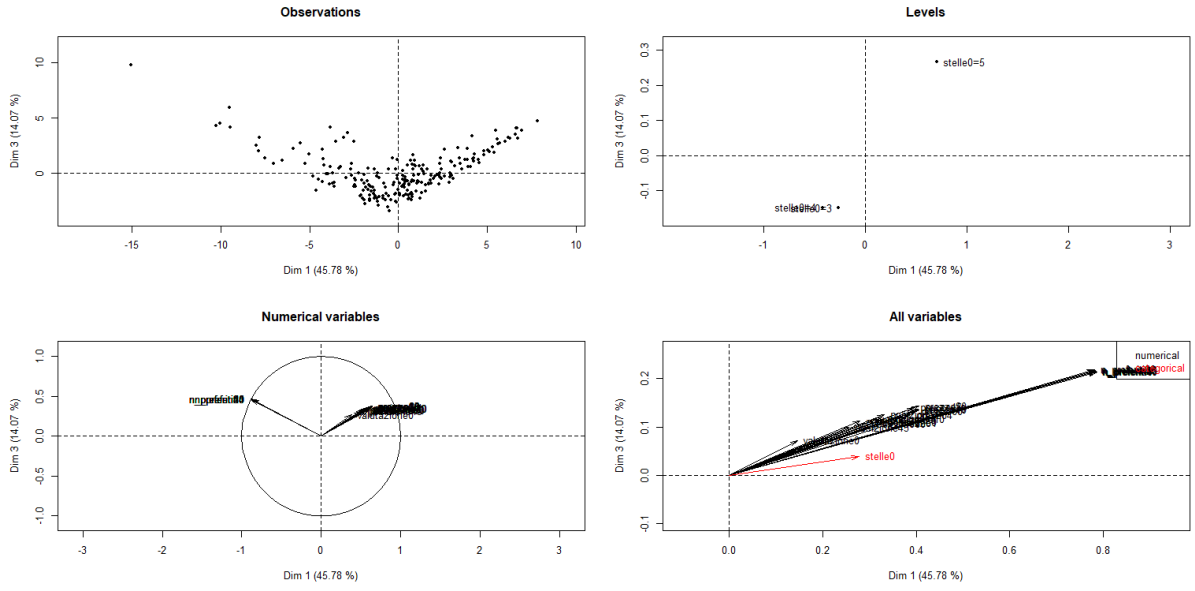

Figura 4-2: Componente principale 1 e Componente principale 3.

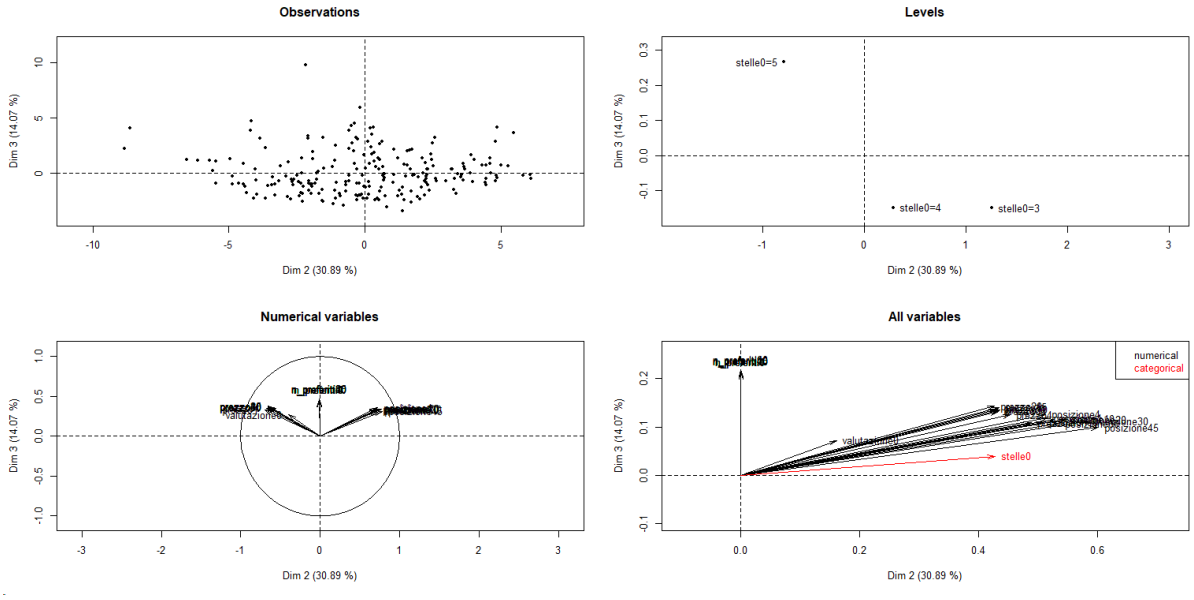

Figura 4-3: Componente principale 2 e Componente principale 3.

Z_1		Z_2		Z_3	
stelle0 [T.5]	-0.159	stelle0 [T.5]	-0.254	stelle0 [T.4]	-0.134
valutazione0	-0.125	valutazione0	-0.194	stelle0 [T.3]	-0.048
stelle0 [T.4]	0.0866	stelle0 [T.4]	0.086	stelle0 [T.5]	0.216
stelle0 [T.3]	0.1005	stelle0 [T.3]	0.436	valutazione0	0.249

Tabella 4-2: Composizione delle componenti principali.

Osservando inoltre le variabili comprese nelle componenti principali, emerge che le variabili significative sono: il "numero di stelle" e la "valutazione"; si evidenzia dunque una differenza rispetto ai singoli modelli del capitolo tre, dove i predittori significativi sono il "numero di preferiti", il "numero di stelle" e la "valutazione". Di conseguenza, utilizzando l'approccio di riduzione delle dimensioni, perde importanza la variabile "numero di preferiti" che non appare fondamentale all'interno delle componenti principali.

La **Figura 4-4** riporta la matrice di correlazione, dalla quale è possibile osservare che, nei grafici in cui si mettono in relazione due, tra le tre dimensioni studiate, la linea dei minimi quadrati è perfettamente piatta, quindi le tre componenti principali non sono correlate tra di loro; osservando, invece, i grafici a dispersione dell'ultima colonna, o dell'ultima riga, i quali mettono in relazione una delle componenti principali con la variabile di risposta, si può notare che si registra una tendenza negativa tra la variabile di risposta con la prima e la seconda

componente principale, invece, si registra una tendenza positiva tra la variabile di risposta e la terza PC, ovvero si evidenziano gli stessi effetti indicati dai coefficienti del modello di regressione lineare (**Tabella 4-3**).

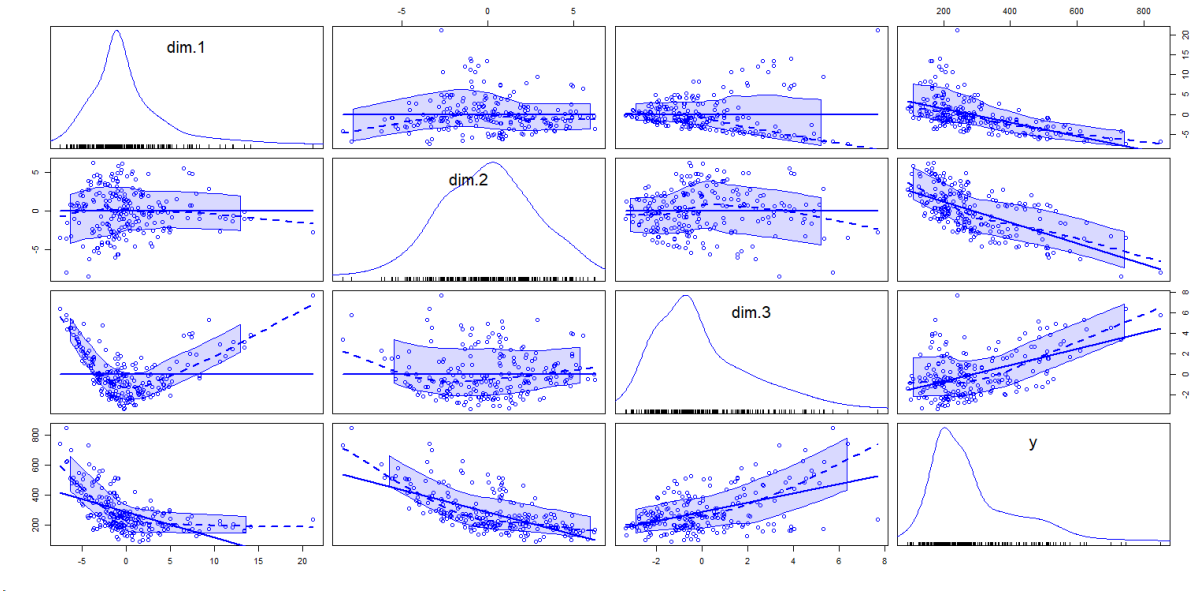

Figura 4-4: Matrice di correlazione tra le componenti principali (dim.1, dim.2 e dim.3) e la variabile di risposta (y).

4.2.2 STIMA DEL MODELLO

Nella **Tabella 4-3** sono invece riportati i risultati del modello di regressione lineare che ha come variabile di risposta il prezzo della data del soggiorno e come predittori le prime tre componenti principali, le quali risultano tutte e tre significative. Inoltre, le PC uno e due impattano negativamente sulla variabile di risposta, in particolare, un aumento unitario della prima componente porta in media ad una diminuzione del prezzo di 17.18, ferme restando le altre variabili, mentre se la seconda dimensione aumentasse di uno, allora la variabile di risposta diminuirebbe mediamente di 29.78, ferme restando le altre variabili. La terza componente principale, invece, impatta positivamente sulla variabile di risposta, infatti, se si avesse un aumento unitario della terza dimensione, allora il prezzo della data del soggiorno aumenterebbe in media di 30.87, ferme restando le altre variabili.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	288.4557	2.1375	134.95	< 2e-16	***
Z_1	-17.1867	0.5042	-34.09	< 2e-16	***
Z_2	-29.7817	0.7526	-39.57	< 2e-16	***
Z_3	30.8748	1.0003	30.87	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 4-3: Modello di regressione lineare con le componenti principali.

Il modello risulta, inoltre, significativo nel suo complesso (statistica F è pari a 1227) ed il coefficiente di determinazione R-quadro mostra che il modello è in grado di spiegare il 94.31% della variabilità della risposta sul campione, mentre l'R-quadro aggiustato è pari 0.9423.

Confrontando il valore dell'R-quadro ottenuto in questo modello con quello dei singoli modelli del capitolo precedente, si può concludere che il modello con le componenti principali spiega meglio la variabilità della risposta rispetto ai singoli modelli di regressione lineare, i quali descrivono la situazione rispetto ad un determinato istante temporale, poiché per questi modelli l'R-quadro non supera lo 0.6717.

4.2.3 DIAGNOSTICA

L'Influence plot permette di identificare i punti di leva presenti nel dataset, in particolare, i cosiddetti punti di leverage cattivo, ossia quei punti che hanno un residuo maggiore di più o meno due e con leverage molto elevato. Dalla **Figura 4-5** non risultano punti di leverage cattivo: non ci sono quindi osservazioni da eliminare dal dataset.

Una delle ipotesi standard della regressione lineare è la normalità dei residui: quest'ultimi, i quali rappresentano le differenze tra i valori osservati e quelli predetti dal modello, devono seguire una distribuzione normale. Nel grafico in alto della **Figura 4-6** è rappresentato un QQ plot: siccome la nuvola di punti è disposta lungo la diagonale, allora l'ipotesi di normalità è compatibile coi dati. Dal secondo grafico si può dedurre che il modello nel suo complesso è ben specificato, nonostante la linea non sia completamente piatta.

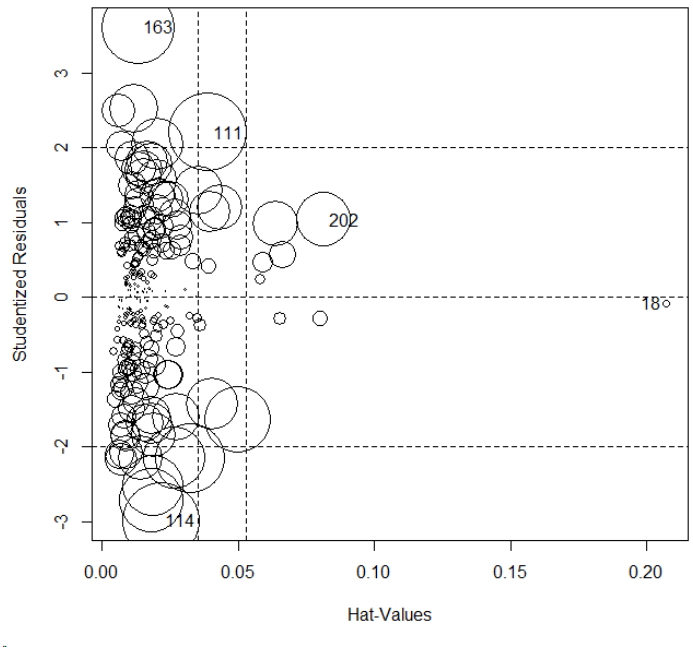

Figura 4-5: Influence plot del modello con le componenti principali.

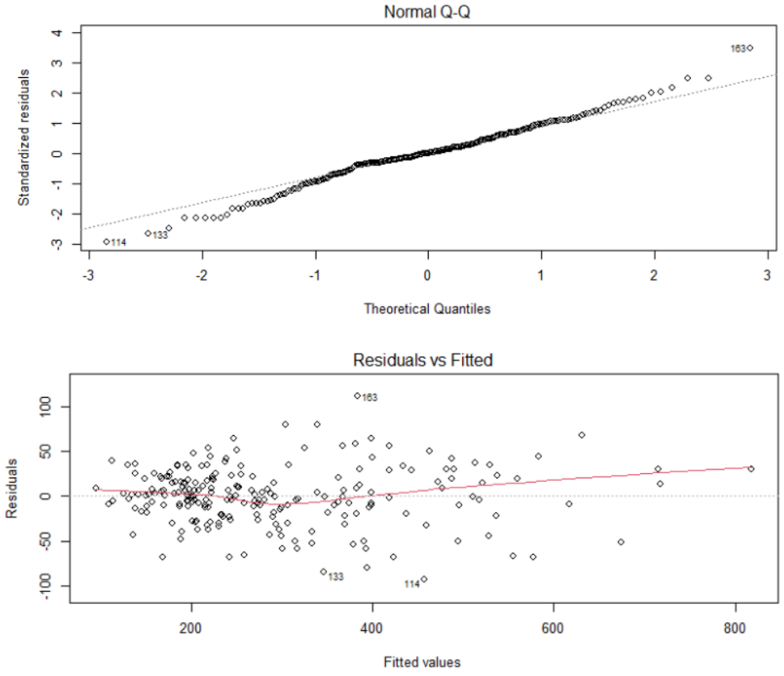

Figura 4-6: Grafici diagnostici di base.

CAPITOLO 5: STIMA DEL MODELLO CON METODO DI COMPRESSIONE

Nel capitolo precedente è stato stimato un unico modello contenente tutte le variabili osservate nel dataset in esame; per evitare però il problema dell'elevata correlazione tra i predittori si è utilizzato un approccio di riduzione della dimensionalità del dataset con l'analisi delle componenti principali, ottenute in seguito alla trasformazione delle variabili originali. Da questa analisi, è emerso che il modello unico con le componenti principali spiega meglio la variabilità della risposta rispetto ai singoli modelli stimati nel terzo capitolo attraverso un metodo naïf.

Un approccio alternativo alla riduzione della dimensionalità consiste nello stimare un unico modello contenente tutti i predittori originali, utilizzando però un metodo di compressione che consente di aggiungere alle stime dei minimi quadrati una penalità che riduce a zero i coefficienti delle variabili controllando così la correlazione tra le variabili in esame. In questo modo, il modello stimato seleziona direttamente le variabili significative.

Dopo un'introduzione teorica dei metodi di shrinkage o compressione, in particolare della regressione ridge e del lasso, si procede con la stima e l'analisi del modello, il quale viene infine confrontato con gli approcci precedenti, con lo scopo di identificare quello che spiega meglio la variabilità della risposta per il dataset in esame.

5.1 METODI DI SHRINKAGE O COMPRESSIONE

In questa sezione si introducono teoricamente i metodi di shrinkage attraverso l'analisi dei seguenti testi: "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" (Hastie, Tibshirani, & Friedman, 2016) e "An Introduction to Statistical Learning with Applications in R" (James, Witten, Hastie, & Tibshirani, 2023).

I metodi di shrinkage consentono di stimare un modello contenente un elevato numero di predittori, utilizzando una tecnica che vincola o regolarizza le stime dei coefficienti o, in modo equivalente, che le comprime verso lo zero; è infatti dimostrato che la "riduzione" delle stime dei coefficienti può ridurre significativamente la loro varianza. Le due tecniche più conosciute per ridurre i coefficienti di regressione verso zero sono la regressione ridge e il lasso.

5.1.1 REGRESSIONE RIDGE

La regressione ridge riduce i coefficienti di regressione imponendo una penalità sulle loro dimensioni; infatti, i coefficienti della regressione ridge minimizzano la somma dei quadrati dei residui penalizzata:

$$\hat{\beta}^{R}{}_{argmin} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \quad \textbf{(5.1)}$$

dove $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$ indica il Residual Sum of Squares (RSS) e $\lambda \geq 0$ è un parametro di regolarizzazione che bisogna determinare separatamente; ergo è un parametro di complessità che controlla l'entità della compressione: maggiore è il valore di λ, maggiore è lo shrinkage. Come per i minimi quadrati, la regressione ridge cerca stime coerenti che adattino bene i dati, rendendo l'RSS piccolo. Tuttavia, il secondo termine $\lambda \sum_j \beta_j^2$, chiamato penalità di restringimento, è piccolo quando $\beta_1, \ldots, \beta_p$ sono vicini a zero ed ha l'effetto di comprimere le stime β_j verso zero. Il parametro di regolarizzazione λ, invece, serve a controllare l'impatto relativo di questi due termini sulle stime dei coefficienti di regressione; in particolare, quando $\lambda = 0$, il termine di penalità non ha alcun effetto e quindi la regressione ridge produce le stesse stime dei minimi quadrati, invece, man mano che λ cresce, l'impatto della penalità aumenta e le stime dei coefficienti della regressione ridge si avvicinano a zero. Inoltre a differenza dei minimi quadrati che generano un solo insieme di stime dei coefficienti, la regressione ridge produce un diverso insieme di stime dei coefficienti $\hat{\beta}_\lambda^R$, uno per ogni valore di λ; di

conseguenza, è fondamentale la selezione di un buon valore per λ. Si noti, inoltre, che l'intercetta β_0 è stata esclusa dal termine di penalità, questo perché si vuole ridurre l'associazione stimata di ogni variabile con la risposta, mentre non si vuole ridurre l'intercetta, che è semplicemente una misura del valore medio della risposta quando $x_{i1} = x_{i2} = ... = x_{ip} = 0$.

Un modo equivalente per scrivere il problema della regressione ridge è:

$$\underset{\beta}{minimize} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \;\; subject\ to \sum_{j=1}^{p} \beta_j^2 \leq t \qquad \textbf{(5.2)}$$

C'è una corrispondenza biunivoca tra i parametri λ in **(5.1)** e t in **(5.2)**: infatti per ogni valore di λ c'è un corrispondente valore t tale che le equazioni **(5.1)** e **(5.2)** danno le stesse stime per i coefficienti della regressione ridge. Quando ci sono molte variabili correlate in un modello di regressione lineare, i loro coefficienti possono essere scarsamente determinati e mostrare un'elevata variabilità. Imponendo un vincolo dimensionale ai coefficienti, come in **(5.2)**, questo problema viene alleviato.

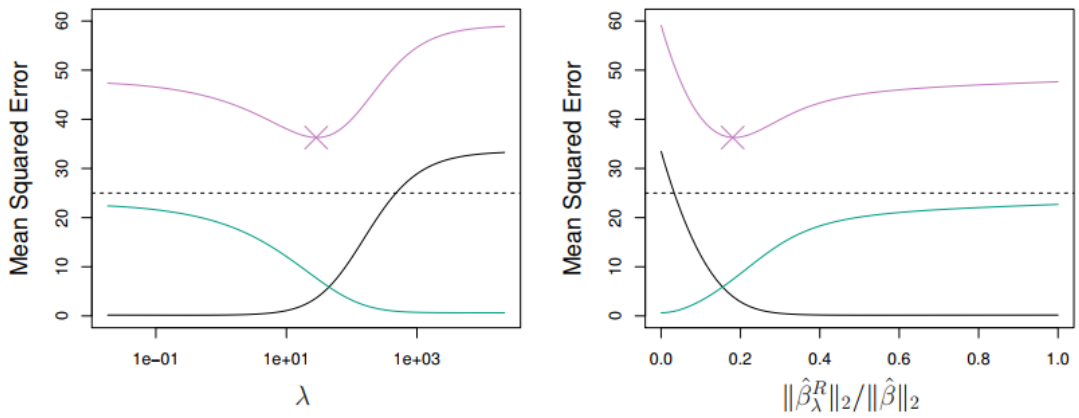

Figura 5-1: Bias al quadrato (nero), varianza (verde), ed errore quadratico medio di prova (viola) per le previsioni della regressione ridge su un set di dati simulato, come funzione di λ e $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. Le linee orizzontali tratteggiate indicano il minore MSE possibile. Le croci viola indicano i modelli di regressione ridge per i quali l'MSE è minore.
Fonte: James, Witten, Hastie, & Tibshirani, 2023, pag 240.

Il vantaggio della regressione ridge rispetto ai minimi quadrati è radicato nel trade-off bias-varianza: all'aumentare di λ la flessibilità di adattamento della regressione ridge diminuisce, portando ad una diminuzione della varianza e ad un aumento della distorsione (bias). Questo è illustrato nella **Figura 5-1**, dove sono riportati bias quadratico (nero), varianza (verde) ed errore quadratico medio di prova (viola) per le previsioni della regressione ridge su un set di dati simulati contenente p = 45 predittori e n = 50 osservazioni, in funzione di λ (nel grafico di sinstra) e in funzione di $\left\|\hat{\beta}_\lambda^R\right\|_2/\left\|\hat{\beta}\right\|_2$ (nel grafico di destra). Le linee orizzontali tratteggiate indicano l'errore quadratico medio minimo possibile (Mean Square Error, MSE) e le croci viola indicano i modelli di regressione ridge per i quali l'MSE è minore.

Nel grafico di sinistra, dove le curve sono tracciate in funzione del parametro di regolarizzazione, per le stime dei minimi quadrati, che corrispondono alla regressione ridge con $\lambda = 0$, la varianza è elevata ma non c'è distorsione, ma all'aumentare di λ la contrazione delle stime dei coefficienti ridge porta ad una sostanziale riduzione della varianza delle previsioni a scapito di un leggero aumento della distorsione. Si ricorda che l'errore quadratico medio del test, tracciato in viola, è strettamente correlato alla varianza più la distorsione al quadrato. Per valori di λ fino a 10 circa, la varianza diminuisce rapidamente, con un aumento molto piccolo della distorsione e di conseguenza l'MSE diminuisce considerevolmente all'aumentare di λ. Oltre questa soglia, la diminuzione della varianza dovuta all'aumento di λ rallenta e la contrazione dei coefficienti fa sì che questi ultimi siano significativamente sottostimati, con conseguente grande aumento della distorsione; l'MSE minimo si ottiene a circa $\lambda = 30$. Inoltre, è interessante notare che, a causa della sua elevata varianza, l'MSE associato ai minimi quadrati quando $\lambda = 0$ è quasi grande come quello del modello nullo, per il quale tutte le stime dei coefficienti sono nulle, ossia quando $\lambda = \infty$. Per un valore intermedio di λ, l'MSE è notevolmente inferiore.

Nel grafico di destra, invece, le stesse curve sono tracciate rispetto a $\left\|\hat{\beta}_\lambda^R\right\|_2/\left\|\hat{\beta}\right\|_2$, dove $\hat{\beta}$ denota il vettore delle stime dei minimi quadrati, mentre $\|\beta\|_2$ denota la norma euclidea "ℓ_2" dei residui di β, che è definita come $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$, e misura la distanza di β da zero. All'aumentare di λ, la norma ℓ_2 di $\hat{\beta}_\lambda^R$ diminuirà sempre, così come $\left\|\hat{\beta}_\lambda^R\right\|_2/\left\|\hat{\beta}\right\|_2$; quest'ultima quantità però varia da 1 (quando $\lambda = 0$, ossia nel caso in cui la stima del coefficiente di regressione ridge è la stessa della stima dei minimi quadrati) a 0 (quando $\lambda = \infty$, nel caso in cui la stima del coefficiente di regressione ridge è un vettore di zeri, con norma ℓ_2 uguale a zero).

Nel grafico di sinistra della **Figura 5-1** quindi l'asse x può essere considerata come la quantità in cui le stime dei coefficienti della regressione ridge sono state ridotte a zero; inoltre, rispetto al grafico di destra, più ci si sposta verso destra più gli adattamenti sono flessibili e di conseguenza il bias diminuisce e la varianza aumenta.

In generale, in situazioni in cui la relazione tra la risposta e i predittori è quasi lineare, le stime dei minimi quadrati hanno una bassa distorsione, ma potrebbero avere un'alta varianza; ciò significa che una piccola modifica nei dati del training set può causare una grande modifica nelle stime dei minimi quadrati. In particolare, quando il numero di variabili p è grande quasi quanto il numero di osservazioni n, le stime dei minimi quadrati sono estremamente variabili; se p > n, allora le stime dei minimi quadrati non hanno una soluzione univoca, mentre la regressione ridge può ancora funzionare bene, se si scambia un piccolo aumento della distorsione con una grande diminuzione della varianza. Di conseguenza, la regressione ridge funziona meglio in situazioni in cui le stime dei minimi quadrati hanno una varianza elevata.

5.1.2 LASSO

La regressione ridge presenta un inconveniente poiché la penalità $\lambda \sum_j \beta_j^2$ riduce tutti i coefficienti verso lo zero, ma non ne fissa nessuno esattamente a zero (a meno che $\lambda = \infty$), di conseguenza il modello finale della regressione ridge comprende sempre tutte le variabili, in quanto l'aumento di λ tende a ridurre le grandezze dei coefficienti, ma non comporta l'esclusione di nessuna delle variabili. Ciò potrebbe rappresentare un problema per l'interpretazione del modello in contesti in cui il numero di variabili p è piuttosto grande. Il LASSO (Least Absolute Shrinkage and Selection Operator Regression) è un'alternativa alla regressione ridge che supera questo svantaggio; infatti, i coefficienti lasso $\hat{\beta}_\lambda^L$ minimizzano la seguente quantità:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2 = RSS + \lambda \sum_{j=1}^{p}|\beta_j| \qquad \textbf{(5.3)}$$

dove $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$ indica il Residual Sum of Squares (RSS) e $\lambda \geq 0$ è un parametro di regolarizzazione. Il lasso traduce ogni coefficiente per un fattore costante λ, troncando a zero; questo processo è chiamato "soft thresholding".

Confrontando **(5.1)** con **(5.3)**, si può notare che l'unica differenza è rappresentata dal secondo termine, la penalità: infatti, per la regressione di ridge è pari a $\lambda \sum_j \beta_j^2$, mentre per il lasso viene sostituito da $\lambda \sum_j |\beta_j|$; si dice, quindi, che il lasso utilizzata una penalità "ℓ_1", invece, di una penalità "ℓ_2". La norma euclidea "ℓ_1" di un vettore coefficiente β è data da $\| \beta \|_1 = \sum |\beta_j|$; il lasso, così come la regressione ridge, riduce le stime dei coefficienti verso zero, ma la penalità "ℓ_1" ha l'effetto di forzare alcune delle stime dei coefficienti ad essere esattamente uguali a zero, quando il parametro di regolarizzazione λ è sufficientemente grande. Di conseguenza, si può affermare che il lasso effettua una selezione delle variabili e, quindi, i modelli finali generati tramite la tecnica lasso sono più facili da interpretare rispetto a quelli prodotti dalla regressione ridge, in quanto comprendono un sottoinsieme delle variabili.

Così come avviene per la regressione ridge, anche per il lasso è fondamentale la selezione di un buon valore di λ: quando $\lambda = 0$, allora il lasso fornisce come risultati i minimi quadrati, mentre, quando λ è sufficientemente grande, il lasso produce come risultato il modello nullo in cui tutte le stime dei coefficienti sono uguali a zero. La cross-validation fornisce un modo semplice per affrontare questo problema: si sceglie una griglia di valori per λ e si calcola l'errore di cross-validation per ogni valore di λ, viene quindi selezionato il valore del parametro di regolarizzazione per il quale l'errore di cross-validation è più piccolo.

Anche in questo caso esiste un'altra formulazione:

$$\underset{\beta}{minimize} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \; subject\ to \sum_{j=1}^{p} |\beta_j| \leq t \qquad \textbf{(5.4)}$$

ovvero, per ogni valore di λ, esiste un valore di t tale che le due equazioni **(5.3)** e **(5.4)** danno le stesse stime dei coefficienti lasso. Inoltre, a causa della natura del vincolo, rendendo t sufficientemente piccolo si fa in modo che alcuni dei coefficienti siano contemporaneamente uguali a zero; in questo modo, il lasso esegue una sorta di selezione continua di sottoinsiemi; bisogna quindi scegliere adeguatamente t, per ridurre al minimo una stima dell'errore di previsione atteso.

5.1.3 CONFRONTO TRA REGRESSIONE RIDGE E LASSO

La **Figura 5-2** mostra i contorni delle funzioni di errore e di vincolo per la regressione lasso (a sinistra) e per la regressione ridge (a destra) quando sono presenti solo due parametri. Le aree blu sono le regioni di vincolo, $|\beta_1| + |\beta_2| \leq t$ e $\beta_1^2 + \beta_2^2 \leq t$, mentre le linee rosse sono i contorni dell'RSS; $\hat{\beta}$ rappresenta la soluzione dei minimi quadrati. Nelle equazioni **(5.2)** e **(5.4)**, se t è sufficientemente grande, allora le regioni di vincolo contengono $\hat{\beta}$ e, quindi, le stime delle regressioni ridge e lasso sono le stesse delle stime dei minimi quadrati. Tuttavia, nella **Figura 5-2**, le stime dei minimi quadrati si trovano al di fuori del rombo e del cerchio, quindi, le stime dei minimi quadrati non sono le stesse delle regressioni ridge e lasso. Ciascuna delle ellissi centrate attorno a $\hat{\beta}$ rappresenta un contorno: ciò significa che tutti i punti di una particolare ellisse hanno lo stesso valore di RSS; man mano che le ellissi di espandono, allontanandosi dalle stime dei minimi quadrati, l'RSS aumenta. Le equazioni **(5.2)** e **(5.4)** indicano che le stime dei coefficienti delle regressioni ridge e lasso sono date dal primo punto in cui l'ellisse entra in contatto con la regione del vincolo. Poiché la regressione ridge ha un vincolo circolare, senza punti acuti, questa intersezione in genere non si verificherà su un asse e quindi le stime dei coefficienti della regressione ridge sono esclusivamente diversi da zero. Il lasso, invece, ha un vincolo a forma di rombo con angoli in corrispondenza di ciascuno degli assi quindi l'ellisse spesso interseca la regione del vincolo in corrispondenza di un asse: quando ciò si verifica, uno dei coefficienti è uguale a zero. Nelle dimensioni superiori, quando p > 2, il diamante diventa un romboide con molti angoli e, quindi, molti coefficienti stimati possono essere uguali a zero contemporaneamente.

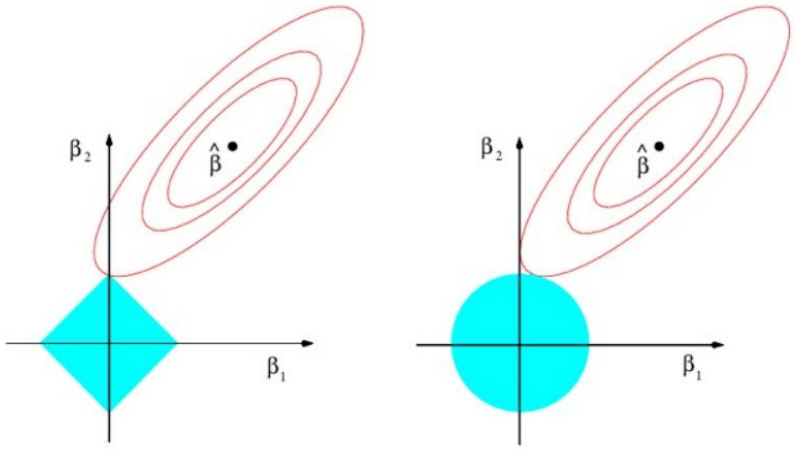

Figura 5-2: Contorni delle funzioni di errore e di vincolo per la regressione lasso (a sinistra) e la regressione ridge (a destra). L'area blu sono le regioni di vincolo $|\beta_1| + |\beta_2| \leq s$ e $\beta_1^2 + \beta_2^2 \leq s$, mentre le linee rosse sono i contorni dell'RSS.
Fonte: James, Witten, Hastie, & Tibshirani, 2023, pag 245.

Nella **Figura 5-3** sono riportati, a sinistra, il grafico del bias quadratico (nero), della varianza (verde) e dell'MSE del test-set per il lasso su un set di dati simulati, gli stessi della **Figura 5-1**; a destra, invece, c'è un confronto del bias quadratico, della varianza e dell'MSE del test-set tra il lasso (linea continua) e la regressione ridge (linea tratteggiata), dove entrambi sono tracciati rispetto al loro R-quadro sui dati di addestramento, come forma comune di indicizzazione. Le croci in entrambi i grafici indicano il modello lasso per il quale l'MSE è minore. Si può osservare che il lasso porta ad un comportamento qualitativamente simile a quello della regressione ridge, in quanto all'aumentare di λ, la varianza diminuisce e la distorsione aumenta. In questo caso, lasso e regressione ridge restituiscono distorsioni identiche, anche se la varianza della regressione ridge è leggermente inferiore a quella del lasso e dunque l'MSE minimo della regressione ridge è leggermente inferiore a quello del lasso; questo succede perché sono stati considerati tutti i predittori. Se invece si considera solo un sottoinsieme di predittori, allora il lasso tende a sovraperformare la regressione ridge in termini di distorsione, varianza e MSE (**Figura 5-4**). Il lasso quindi funziona meglio quando un numero relativamente piccolo di predittori ha coefficienti sostanziali e i predittori rimanenti hanno coefficienti molto piccoli o uguali a zero; la regressione ridge invece funziona meglio quando la risposta è una funzione di molti predittori, tutti con coefficienti di dimensioni approssimativamente uguali.

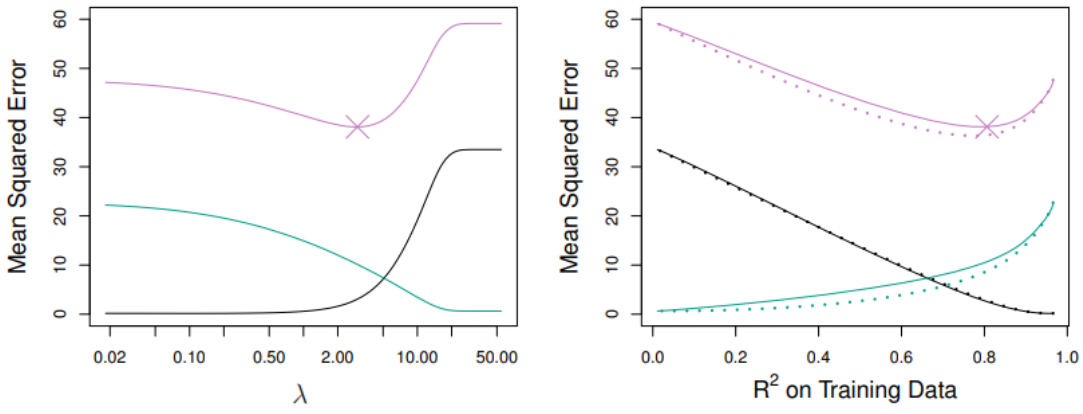

Figura 5-3: A sinistra: grafici del bias quadratico (nero), della varianza (verde) e dell'MSE di prova (viola) per il lasso su un set di dati simulati. A destra: confronto tra bias quadratico, varianza e MSE del test tra lasso (linea continua) e regressione ridge (linea tratteggiata). Entrambi sono tracciati rispetto al loro R^2 sui dati di addestramento, come forma comune di indicizzazione. Le croci in entrambi i grafici indicano il modello lasso per il quale l'MSE è minore.
Fonte: James, Witten, Hastie, & Tibshirani, 2023, pag 246.

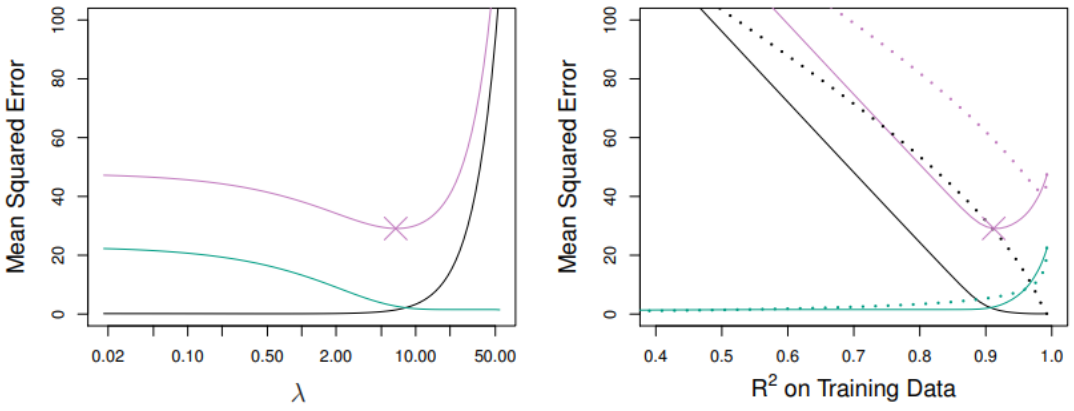

Figura 5-4: A sinistra: grafici del bias quadratico (nero), della varianza (verde) e dell'MSE di prova (viola) per il lasso su un set di dati simulati. A destra: confronto tra bias quadratico, varianza e MSE del test tra lasso (linea continua) e regressione ridge (linea tratteggiata). Entrambi sono tracciati rispetto al loro R^2 sui dati di addestramento, come forma comune di indicizzazione. Le croci in entrambi i grafici indicano il modello lasso per il quale l'MSE è minore.
Fonte: James, Witten, Hastie, & Tibshirani, 2023, pag 247.

In generale, si può affermare che la regressione ridge restringe, più o meno, ogni stima dei coefficienti dei minimi quadrati della stessa proporzione, mentre il lasso riduce, più o meno, verso lo zero tutti i coefficienti di una quantità simile e i coefficienti sufficientemente piccoli sono ridotti fino a zero.

5.2 APPLICAZIONE DEL LASSO AL CASO IN STUDIO

Per la realizzazione del modello si è deciso di utilizzare come metodo di shrinkage il lasso poiché, a differenza della regressione ridge, fissa i coefficienti di alcuni predittori esattamente uguali a zero, rendendo così il modello finale più semplice da interpretare, in quanto risulta formato da un sottoinsieme di predittori; la composizione del sottoinsieme finale, però, dipende dal valore del parametro di regolarizzazione λ scelto. Infatti, nel grafico di sinistra della **Figura 5-5**, dove sono riportate le curve dei coefficienti lasso in funzione di λ, presa in forma logaritmica, si può osservare che, quando $\lambda = 0$, il modello contiene gli stessi risultati dei minimi quadrati, ma all'aumentare di λ, il numero di coefficienti del modello diminuisce. Nel grafico di destra, invece, le curve dei coefficienti lasso sono tracciate in funzione della norma "ℓ_1" (ossia di $\|\hat{\beta}_\lambda^L\|_1 / \|\hat{\beta}\|_1$), in questo caso, la situazione è opposta: infatti spostandosi da sinistra a destra si può notare che, inizialmente solo un predittore rientra nel modello finale, ma all'aumentare della norma "ℓ_1", il numero di coefficienti non nulli aumenta. Di conseguenza, a seconda del valore del parametro di regolarizzazione λ, il lasso può produrre un modello che comprende un qualsiasi numero di variabili.

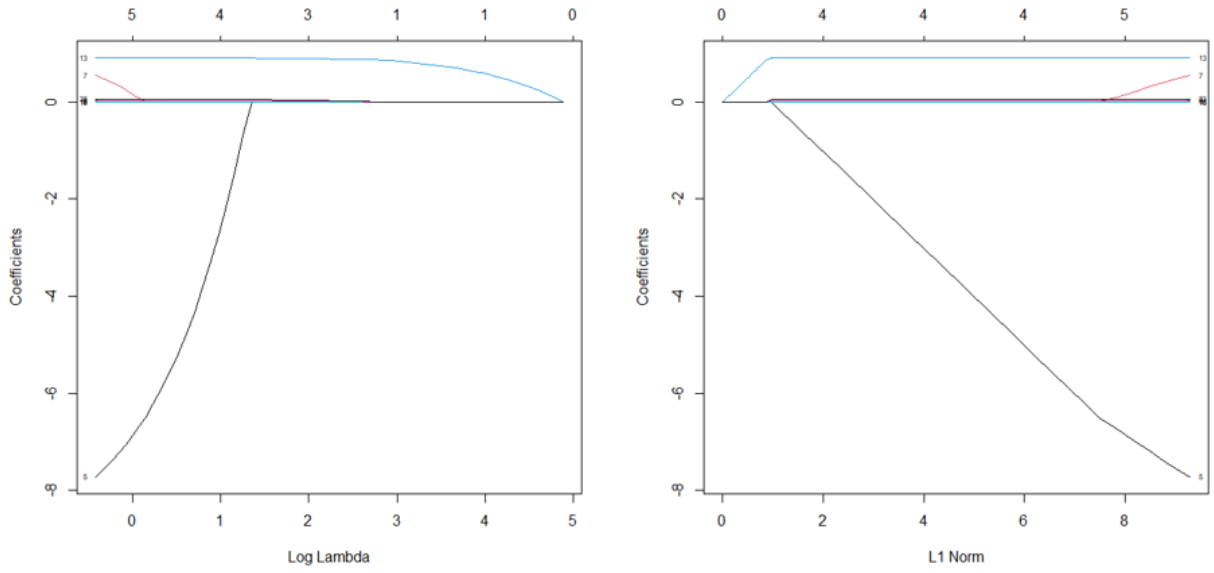

Figura 5-5: Si mostrano i coefficienti lasso in funzione di λ, preso in forma logaritmica, e in funzione di $\|\hat{\beta}_\lambda^L\|_1 / \|\hat{\beta}\|_1$

Per selezionare il valore del parametro di regolarizzazione λ si può utilizzare la tecnica della cross-validation o convalidazione incrociata. Nel caso in esame, λ assume valori che variano principalmente da 0.871 a 1.522; per questo motivo, si è scelto di utilizzare un valore intermedio, ovvero $\lambda = 1.151$. Utilizzando la tecnica della cross-validation è possibile ottenere il grafico presente nella **Figura 5-6**, dove i punti rossi indicano i vari valori assunti dall'errore quadratico medio in funzione del parametro di regolarizzazione λ, preso in forma logaritmica. Nel caso specifico, osservando la curva si può notare che l'errore quadratico medio aumenta all'aumentare di λ, inoltre le linee verticali tratteggiate indicano due valori specifici di λ: quella a sinistra rappresenta il valore di λ a cui corrisponde il più piccolo errore quadratico medio ottenuto in seguito alla cross-validation; la linea di destra, invece, indica un valore di λ tale che l'errore quadratico medio non superi l'errore minimo più di una deviazione standard. Questi due valori di λ forniscono due differenti modelli: il primo presenta un numero maggiore di coefficienti diversi da zero, mentre il secondo risulta essere più regolarizzato e, quindi, con meno predittori significativi; infatti, il numero di coefficienti non nulli in corrispondenza dei valori assunti da λ è visibile nella parte alta del grafico.

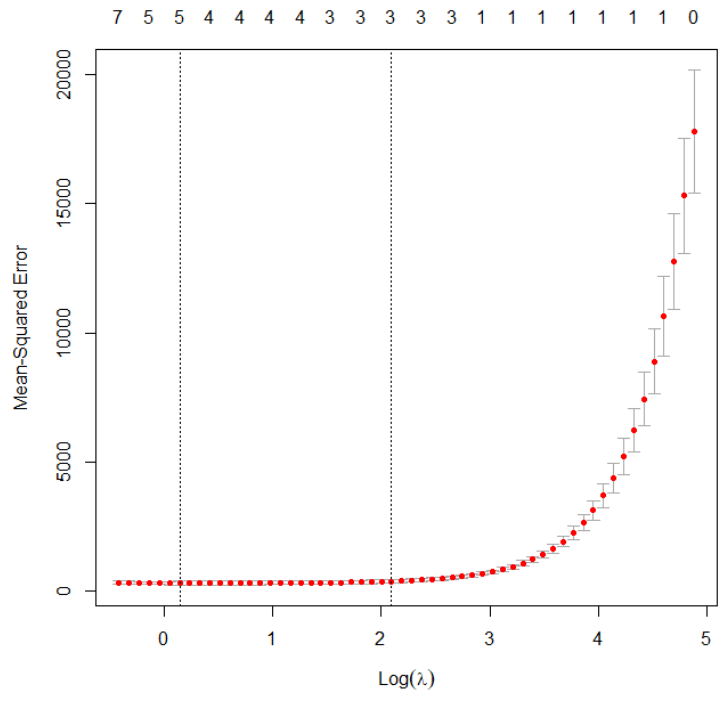

Figura 5-6: Mean Squared-Error in funzione di Log(λ).

5.2.1 STIMA DEL MODELLO

Scegliendo 1.151, come migliore valore di λ, si ottiene un modello finale che comprende un'intercetta e quattro predittori con coefficienti non nulli (**Tabella 5-1**). Dalla matrice dei coefficienti, si può osservare che le variabili, che influenzano principalmente il prezzo della data del soggiorno, sono due: "stelle" e "prezzi", precedenti alla data del soggiorno.

	STIME
(Intercept)	1.729e-00
stelle0 [T.3]	-6.483e-00
stelle0 [T.5]	5.744e-03
prezzo1	9.077e-01
prezzo4	1.677e-02
prezzo30	5.899e-02

Tabella 5-1: Modello con tecnica lasso.

In particolare, si può notare che i predittori significativi impattano tutti positivamente sulla risposta, tranne il livello tre della variabile dummy "stelle": questo significa che, a parità di altre condizioni, gli hotel con tre stelle hanno, in media, un prezzo finale inferiore di 6.48 rispetto a quello degli hotel con quattro stelle, livello accettato implicitamente dal modello. Il livello cinque stelle, invece, impatta positivamente sulla risposta anche se in modo limitato, in particolare, si può affermare che gli hotel con cinque stelle hanno un prezzo finale superiore di 0.0057 rispetto a quello degli hotel a quattro stelle, ferme restando le altre variabili.

Per quanto riguarda i prezzi precedenti, invece, si può notare che i prezzi registrati il giorno prima, quattro giorni prima e trenta giorni prima impattano positivamente sul prezzo della data del soggiorno; più precisamente, l'impatto maggiore si ha per il prezzo1, ossia il prezzo del giorno precedente la data del soggiorno, infatti, se quest'ultimo aumentasse di una unità, allora il prezzo finale aumenterebbe di 0.908, ferme restando le altre variabili. Con riferimento al dataset in esame, è possibile affermare che, se i prezzi delle camere aumentassero trenta, quattro o un giorno prima della data del soggiorno, allora si avrebbe un aumento anche del prezzo finale ossia del prezzo della data del soggiorno.

Confrontando le variabili selezionate con la tecnica lasso, con quelle dei modelli precedenti, emerge una differenza importante: la variabile "valutazione", la quale appare sia nei modelli naïf che nel modello con le componenti principali, non è più significativa: questo significa che, per il modello stimato con la tecnica lasso, il punteggio medio delle recensioni di ogni hotel lasciate dagli utenti sulla piattaforma Booking.com non impatta più in modo significativo sul prezzo finale. L'unica variabile che rimane significativa in tutti i modelli analizzati è la variabile "stelle" che indica il numero di stelle degli hotel osservati e rappresenta quindi un indice di qualità; infatti, in tutti i modelli stimati i livelli quattro e cinque stelle hanno un impatto positivo maggiore rispetto agli hotel con tre stelle.

Per il modello in esame, l'R-quadro è pari a 0.9795. Confrontando questo valore con l'R-quadro del modello con le componenti principali e con quello dei singoli modelli del terzo capitolo, si può notare che, tra i tre approcci esaminati, quello che spiega meglio la variabilità della risposta è quest'ultimo approccio ossia il modello stimato con la tecnica lasso; tuttavia, anche il modello con le componenti principali spiega bene la variabilità del prezzo in quanto ha un R-quadro pari a 0.9428. Si verifica, invece, un'elevata differenza rispetto ai singoli modelli dell'approccio naïf, dove l'R-quadro non supera lo 0.6717.

Per visualizzare la bontà di adattamento si veda la **Figura 5-7**, dove è rappresentato un grafico a dispersione che mette in relazione i valori effettivi della variabile di risposta con i valori predetti dal modello ottenuto con la cross-validation. Dal grafico si deduce che il modello si adatta bene ai dati osservati; di conseguenza si può affermare che il modello, stimato tramite la tecnica del lasso, produce una buona predizione in-sample.

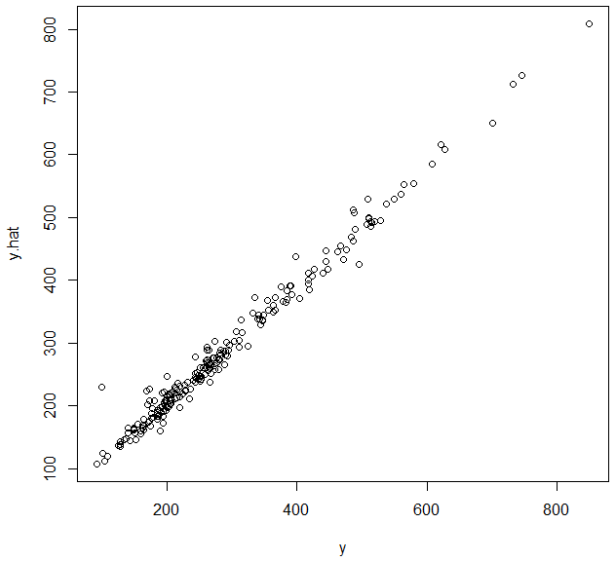

Figura 5-7: Grafico a dispersione tra valori effettivi (y) e valori predetti (y.hat).

CAPITOLO 6: UN'ULTERIORE ANALISI

I modelli stimati nei capitoli precedenti hanno, come predittori, le variabili osservate negli otto istanti temporali analizzati; tuttavia, proprio perché queste variabili vengono osservate più volte nel tempo, potrebbero presentare una correlazione seriale legata alla serie storica delle variabili. Per questo motivo, si è deciso di procede con un'ulteriore analisi che consiste nello stimare un modello avente, come variabile di risposta, il prezzo della data del soggiorno e, come predittori, le variabili osservate considerate, però, come variazioni tra due istanti temporali successivi, ad eccezione delle variabili "stelle" e "valutazione", le quali vengono considerate con riferimento alla data del soggiorno poiché rimangono costanti per l'intero orizzonte temporale studiato.

Calcolando per le variabili "n_preferiti", "posizione" e "prezzo" le variazioni per due istanti temporali successi, si eliminano i trend legati alla serie storica di queste variabili, che cambiano velocemente durante il periodo osservato, evitando così le eventuali correlazioni spurie.

In questo capitolo si procede inizialmente con la stima di un modello di regressione lineare, per poi stimare un ulteriore modello utilizzando la tecnica lasso; infine, si commentano e confrontano i risultati ottenuti, al fine di identificare il modello che spiega meglio il dataset in esame.

6.1 MODELLO DI REGRESSIONE LINEARE

Il modello di regressione lineare, che viene stimato in questa sezione, è identificato dall'equazione **(4.1)**, dove X_{ij} non indica i predittori originali, ma rappresenta le seguenti variabili esplicative:

- "stelle" e "valutazione" considerate al tempo zero poiché rimangono costanti nel periodo osservato;
- "n_preferiti", "posizione" e "prezzo" considerati come variazioni tra due istanti temporali successivi.

Si precisa che i numeri, inseriti al termine del nome delle variabili, indicano i giorni prima del soggiorno; per esempio, nella **Tabella 6-1**, appare la variabile "n_preferiti (45-60)", la quale indica la variazione del numero di preferiti lasciato dagli utenti della piattaforma Booking.com tra sessanta e quarantacinque giorni prima della data del soggiorno.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-38.496	68.342	-0.563	0.574	
n_preferiti (45-60)	-0.260	0.087	-2.993	0.003091	**
posizione (4-10)	0.841	0.260	3.233	0.001422	**
posizione (20-30)	0.928	0.277	3.350	0.000957	***
posizione (30-45)	1.147	0.243	4.716	0.000004356	***
posizione (45-60)	-0.458	0.238	-1.927	0.055352	.
prezzo (0-1)	0.734	0.270	2.721	0.007051	**
prezzo (1-4)	0.335	0.177	1.887	0.060580	.
prezzo (4-10)	0.996	0.190	5.241	0.000000384	***
prezzo (10-20)	0.576	0.234	2.455	0.014899	*
prezzo (20-30)	0.749	0.266	2.819	0.005275	**
stelle0 [T.4]	35.503	16.908	2.100	0.036938	*
stelle0 [T.5]	195.597	19.448	10.057	<2e-16	***
valutazione0	28.499	8.524	3.343	0.000979	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabella 6-1: Modello di regressione lineare.

Nel modello stimato appaiono tutte le variabili significative, ossia quelle variabili che hanno una statistica t che assume un valore maggiore di due e un relativo p-value minore di 0.05, ad eccezione, dei predittori, “posizione (45-60)” e “prezzo (1-4)”, i quali sono significativi con un p-value minore di 0.10.

Inoltre, tutte le variabili significative impattano positivamente sul prezzo del giorno del soggiorno, tranne le variabili esplicative “n_preferiti (45-60)” e “posizione (45-60)”, le quali impattano negativamente sulla variabile di risposta: questo significa che se il numero di preferiti dovesse aumentare in modo unitario tra quarantacinque e sessanta giorni prima, allora il prezzo finale diminuirebbe, in media, di 0.26, ferme restando le altre variabili; mentre, se un hotel dovesse assumere, nell’elenco dei risultati, una posizione maggiore per lo stesso intervallo temporale, il prezzo del giorno del soggiorno diminuirebbe mediamente di 0.458, a parità di altre condizioni.

Si può anche osservare che i coefficienti delle variabili “numero di preferiti”, “posizione” e “prezzi”, presi come variazioni tra due istanti temporali successivi, assumono un valore relativamente basso, mentre le variabili “stelle” e “valutazione” considerate una sola volta poiché costanti per l’intero orizzonte temporale osservato, hanno un impatto maggiore sulla variabile di risposta: in particolare, gli hotel con quattro stelle hanno, in media, un prezzo che supera quello degli hotel a tre stelle di 35.503, mentre gli hotel con cinque stelle hanno mediamente un prezzo superiore di quello degli hotel con tre stelle di 195.60; infine, se un hotel dovesse aumentare di uno la valutazione online, allora il prezzo del giorno del soggiorno, in media, aumenterebbe di 28.50, ferme restando le altre variabili.

L’R-quadro del modello stimato è pari a 0.7533, ovvero il modello spiega il 75.33% della variabilità del prezzo del giorno del soggiorno, mentre l’R-quadro aggiustato è pari a 0.7382. Il modello risulta inoltre significativo anche nel suo complesso (statistica F pari a 49.8).

Confrontando i risultati ottenuti con quelli dei modelli precedenti si deduce che, in quest’ultimo modello, diventa significativa anche la variabile posizione, presa come variazione tra due istanti successivi; in particolare impatta sul prezzo la variazione di posizione degli hotel nell’elenco dei risultati di Booking.com negli intervalli: da quattro a dieci giorni prima, da venti a trenta giorni prima, da trenta a quarantacinque giorni prima e da quarantacinque a sessanta giorni prima del soggiorno. Inoltre, come accade per i modelli di regressione lineare del capitolo tre, anche in questo caso risulta significativa negativamente la variabile “n_preferiti”, ovvero la

variazione del numero di utenti che hanno indicato un hotel come preferito tra i sessanta e quarantacinque giorni prima impatta linearmente sul prezzo della data del soggiorno.

Osservando il valore del coefficiente di determinazione, si può affermare che questo modello ha un R-quadro minore sia di quello del modello con le componenti principali sia di quello del modello con il lasso.

6.1.1 DIAGNOSTICA

Attraverso l'Influence Plot (**Figura 6-1**) è possibile individuare sia gli eventuali outliers, ossia punti anomali all'interno del dataset in esame, in quanto sono troppo distanti dalla variabile di risposta, sia i punti di High Leverage, ossia dati troppo distanti dalla media dei predittori; in particolare, si vuole evidenziare gli eventuali punti di leverage cattivo, che, in tal caso, andrebbero eliminati dal dataset poiché, altrimenti, porterebbero ad un errore nell'analisi. Dalla **Figura 6-1**, però, è possibile notare che non si hanno osservazioni che costituiscono punti di leverage cattivo.

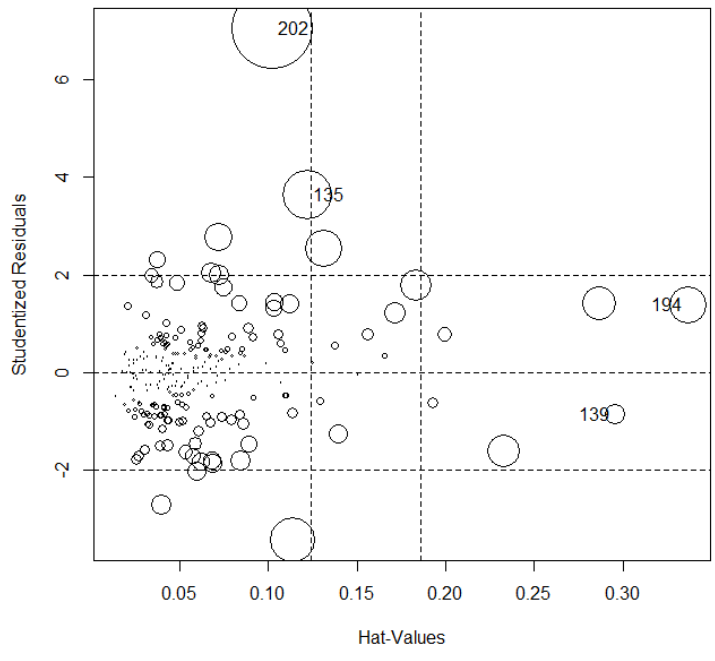

Figura 6-1: Influence plot del modello con le variazioni.

Calcolando il Fattore di Inflazione della Varianza (VIF), per lo studio della collinearità, si può osservare che le variabili significative del modello stimato non sono correlate tra di loro, in quanto il VIF assume sempre valori inferiori a due, il quale rappresenta il valore soglia (**Tabella 6-2**).

	GVIF	Df	GVIF^(1/(2*Df))
n_preferiti (45-60)	1.128	1	1.062
posizione (4-10)	1.078	1	1.038
posizione (20-30)	1.092	1	1.045
posizione (30-45)	1.076	1	1.037
posizione (45-60)	1.132	1	1.064
prezzo (0-1)	1.082	1	1.040
prezzo (1-4)	1.262	1	1.123
prezzo (4-10)	1.259	1	1.122
prezzo (10-20)	1.058	1	1.028
prezzo (20-30)	1.098	1	1.048
stelle0	1.767	2	1.153
valutazione0	1.525	1	1.235

Tabella 6-2: VIF.

Una delle ipotesi standard della regressione lineare è la normalità dei residui, i quali rappresentano le differenze tra i valori osservati e quelli predetti dal modello e devono seguire una distribuzione normale. Quest'ipotesi è confermata dal grafico in alto presente nella **Figura 6-2**, dove è rappresentato uno scatterplot che confronta i percentili osservati degli errori studentizzati e quelli che si avrebbero se ci fosse normalità, in quanto la nuvola di punti è disposta lungo la diagonale. Dal grafico in basso, invece, si può dedurre che il modello è ben specificato poiché la linea appare quasi piatta, anche se devia leggermente nelle due estremità.

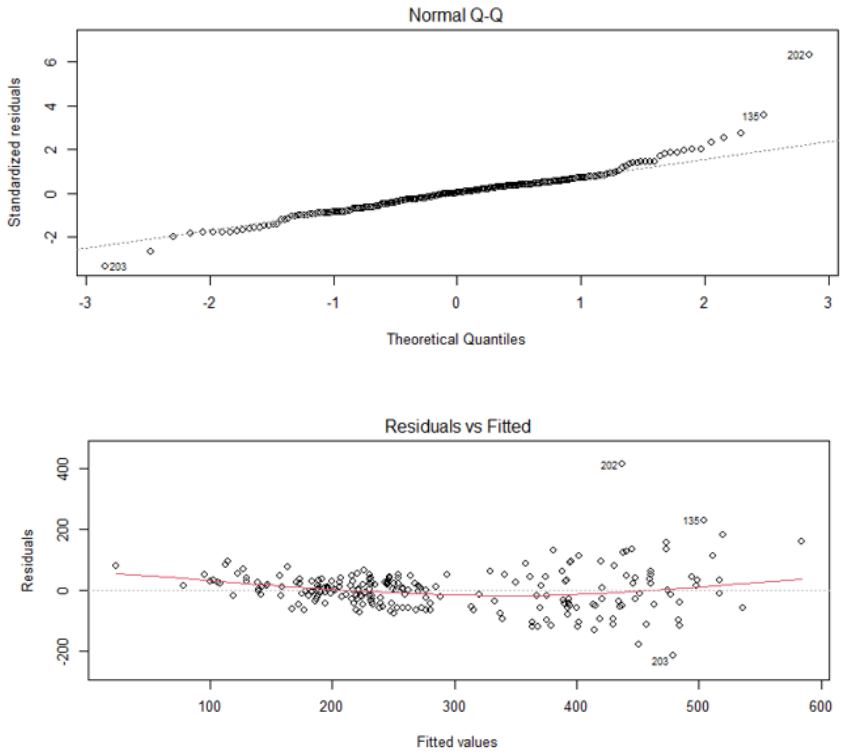

Figura 6-2: Grafici diagnostici di base per il modello con variazioni.

6.2 APPLICAZIONE DEL LASSO

Il modello precedente presenta un R-quadro pari a 0.7533, questo potrebbe derivare dal fatto che le variabili esplicative presentano tra di loro un'elevata correlazione; per questo motivo, si è deciso di procedere con la stima di un altro modello utilizzando nuovamente la tecnica lasso al fine di capire se ci possano essere miglioramenti nell'analisi del dataset.

Per il dataset comprendente le differenze tra istanti successivi, il parametro di regolarizzazione λ varia principalmente da 0.42 a 0.67, si è quindi deciso di utilizzare un valore intermedio, ossia $\lambda = 0.512$. Dalla **Figura 6-3** si può osservare che l'errore quadratico medio aumenta all'aumentare del parametro di regolarizzazione preso in forma logaritmica.

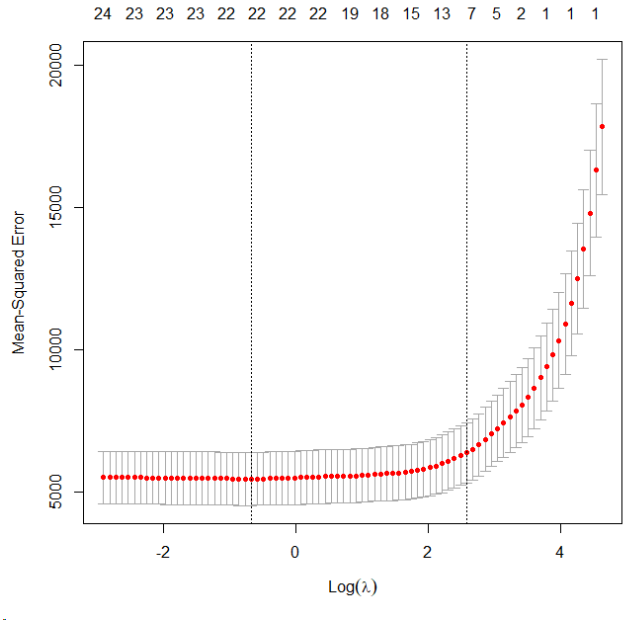

Figura 6-3: Mean Squared-Error in funzione di Log(λ).

Fissando $\lambda = 0.512$ si ottiene un unico modello costituito da un'intercetta e ventuno variabili (**Tabella 6-3**) tra le quali si ha la variabile dummy "stelle" che è suddivisa in tre livelli, di cui uno è accettato implicitamente nel modello. Dalla matrice dei coefficienti si può notare che le variabili che impattano maggiormente sulla variabile di risposta sono "stelle" e "valutazione": in particolare, a parità di condizioni gli hotel con tre stelle hanno in media un prezzo inferiore di 28.84 rispetto agli hotel a quattro stelle, mentre gli hotel con cinque stelle hanno mediamente un prezzo che supera quello degli hotel a quattro stelle di 161.27. La variabile "valutazione" impatta positivamente sul prezzo della data del soggiorno, in particolare se si aumenta di uno la valutazione allora il prezzo della data del soggiorno aumenta di 29, ferme restando le altre

variabili. Dal modello finale appaiono significative anche le altre variabili: “n_preferiti”, “posizione” e “prezzo”, le quali però hanno un impatto limitato sulla variabile di risposta, infatti, presentano tutte un coefficiente non superiore a uno.

	STIME
(Intercept)	-3.13
n_preferiti 0_1	0.37
n_preferiti 1_4	0.505
n_preferiti 4_10	-0.801
n_preferiti 10_20	0.64
n_preferiti 45_60	-0.55
posizione 0_1	-0.14
posizione 1_4	-0.068
posizione 4_10	0.74
posizione 10_20	-0.26
posizione 20_30	0.99
posizione 30_45	1.045
posizione 45_60	-0.42
prezzo 0_1	0.66
prezzo 1_4	0.32
prezzo 4_10	1.02
prezzo 10_20	0.62
prezzo 20_30	0.87
prezzo 30_45	0.38
prezzo 45_60	0.41
stelle0 [T. 3]	-28.84
stelle0 [T. 5]	161.27
valutazione	29

Tabella 6-3: Modello con tecnica lasso.

Il modello stimato presenta un R-quadro pari a 0.6721: il modello spiega quindi il 67.21% della variabilità del prezzo della data del soggiorno. Confrontando questo modello con quello di regressione lineare si può affermare che il modello di regressione lineare ha un R-quadro maggiore e quindi spiega meglio la variabilità della risposta rispetto al modello stimato

utilizzando la tecnica lasso; in questo caso l'utilizzo di questo metodo di shrinkage va a peggiorare il modello. Inoltre, per valutare graficamente la bontà di adattamento del modello stimato si veda la **Figura 6-4**, dove attraverso un grafico a dispersione si mettono in relazione i valori effettivi della variabile di risposta con i relativi valori predetti dal modello. Osservando il grafico si deduce che il modello non si adatta bene ai dati osservati ed evidenzia la presenza di due gruppi; di conseguenza il modello stimato non produce una buona predizione in-sample.

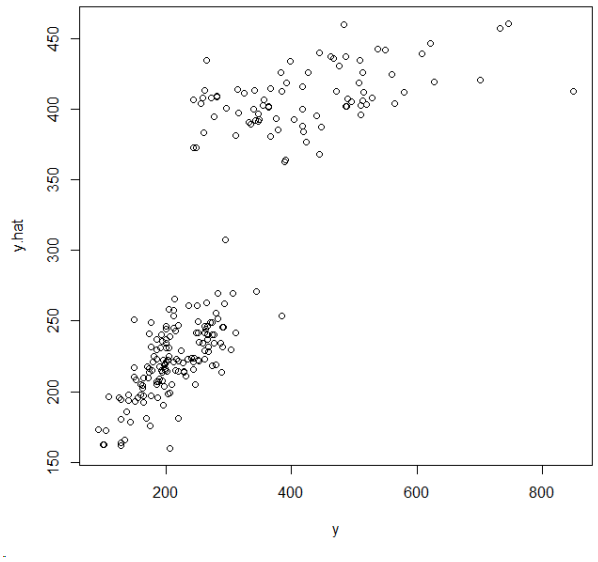

Figura 6-4: Grafico a dispersione tra valori effettivi (y) e valori predetti (y.hat).

Confrontando i due modelli stimati entrambi con la tecnica lasso, che permette di controllare l'elevata correlazione tra i predittori e seleziona direttamente le variabili esplicative significate, si conclude che, con riferimento ai dati analizzati, il metodo di shrinkage utilizzato risulta fondamentale per il dataset che considera le variabili originarie, mentre è meno efficace per il modello che considera come predittori le variabili prese come differenze tra due istanti di tempo successivi, infatti il modello del capitolo cinque presenta un R-quadro pari a 0.9795, mentre quest'ultimo modello stimato ha un R-quadro pari a 0.6721.

CONCLUSIONE

Oggigiorno, nel settore alberghiero, le agenzie di viaggio online sono sempre più importanti per i consumatori poiché permettono di accedere a numerose informazioni e, soprattutto, consentono di confrontare velocemente il prezzo applicato dai vari hotel per un determinato servizio; in questo modo, i consumatori possono valutare rapidamente quale sia l'offerta più conveniente e possono quindi effettuare una scelta più ponderata. Queste piattaforme, inoltre, permettono di raccogliere i dati sui vari utenti, consentendo agli hotel di migliorare la propria offerta.

L'elaborato, infatti, si concentra sull'analisi di un subset del dataset utilizzato nell'articolo "The impact of dynamic price variability on revenue maximization" (Abrate, Nicolau, & Viglia, The impact of dynamic price variability on revenue maximization, 2019), dove si analizzano i prezzi ed altri fattori, presenti sulla piattaforma Booking.com, di 255 hotel della città di Londra per tutto il mese di aprile del 2016, con lo scopo di individuare, attraverso vari approcci e modelli, quali fattori impattano maggiormente sul prezzo delle camere.

Il dataset in esame è costituito dalle seguenti variabili: "camera", "numero di stelle", "numero di preferiti", "numero di recensioni", "pagina", "posizione", "prezzo" e "valutazione", osservate in otto istanti temporali differenti, ipotizzando cioè di prenotare sessanta, quarantacinque, trenta, venti, dieci, quattro giorni prima del soggiorno, ma anche un giorno prima e, infine, il giorno stesso del soggiorno. Per questo motivo, si è deciso di iniziare con un approccio definito naïf, il quale consiste nello stimare otto modelli di regressione lineare differenti, uno per ogni istante temporale osservato, dai quali è emerso che le variabili che impattano sui relativi prezzi sono "numero di preferiti", "numero di stelle" e "valutazione". Successivamente, sono stati analizzati come variano i coefficienti di queste variabili nell'orizzonte temporale studiato, evidenziando un aumento dei coefficienti all'avvicinarsi della data del soggiorno, sottolineando così un impatto lineare maggiore.

In seguito, si prosegue con la stima di un unico modello che consideri, come variabile di risposta, il prezzo del giorno del soggiorno in relazione a tutte le altre variabili; tuttavia, il modello in questione presenta un'elevata correlazione tra le variabili esplicative, in quanto è formato dalle componenti autoregressive, ossia i prezzi registrati negli altri istanti temporali, che sono altamente correlate con la variabile di risposta, e dalle altre variabili esplicative, considerate in tutti gli otto istanti temporali, evidenziando quindi un'ulteriore elevata

correlazione. Per poter controllare questo problema si procede in due modi differenti: il primo consiste nello stimare un modello di regressione lineare con le componenti principali, ossia si creano nuove variabili, la quali sono costruite come combinazioni lineari delle variabili originarie ponderate con pesi differenti; il secondo approccio, invece, si basa su un metodo di shrinkage, il lasso, attraverso il quale è possibile aggiungere, alle stime dei minimi quadrati, una penalità che riduce a zero i coefficienti di alcune variabili, controllando in questo modo le correlazioni esistenti; di conseguenza, si ottiene così un modello che comprende le variabili originarie significative.

Dai risultati emerge che il modello, il quale spiega meglio la variabilità dei dati in esame è quello realizzato con la tecnica lasso, in quanto spiega il 97.95% della variabilità del prezzo del soggiorno, tuttavia, anche il modello con le componenti principali spiega il 94.28% della variabilità della risposta, mentre i singoli modelli identificati con l'approccio naïf hanno un R-quadro non superiore allo 0.6717. Inoltre, è possibile notare che l'unica variabile significativa in tutti e tre gli approcci è il "numero di stelle", questo significa che questa variabile rappresenta un fattore fondamentale nella determinazione del prezzo delle camere d'hotel.

Nell'ultimo capitolo si procede con un'analisi differente, ovvero si stimano due diversi modelli, uno di regressione lineare e uno con la tecnica lasso, che hanno come variabile di risposta il prezzo della data del soggiorno e come predittori considerano le variazioni delle variabili osservate tra due istanti temporali successivi; l'obiettivo consiste nel controllare la correlazione seriale delle variabili osservate in più istanti temporali. In questo caso, però, la tecnica lasso produce un modello che un R-quadro piuttosto basso (0.6721) e addirittura inferiore a quello del modello di regressione lineare (0.7533).

In conclusione, le informazioni ottenute da questa ricerca sono utili per comprendere quali fattori sono rilevanti nella determinazione del prezzo delle camere d'hotel, tuttavia, dai risultati ottenuti è anche emerso che gli approcci analizzati forniscono in parte risultati significativi differenti.

BIBLIOGRAFIA

Abrate, G. (2020). Pricing a creazione di valore. Strumenti e applicazioni manageriali. Aracne Editrice.

Abrate, G., & Viglia, G. (2016). Strategic and tactical price decisions in hotel revenue management. *Tourism Management 55*, 123-132.

Abrate, G., Nicolau, J. L., & Viglia, G. (2019). The impact of dynamic price variability on revenue maximization. *Tourism Management, 74*, 224-233.

Binesh, F., Belarmino, A., & Raab, C. (2021). A meta-analysis of revenue management. *Revenue and Pricing Management 20*, 546-558.

Gavilan, D., Avello, M., & Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management 66*, 53-61.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (edition 2nd).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An Introduction to Statistical Learning with Applications in R. Springer (edition 2nd).

Moro, S., Rita, P., & Oliveira, C. (2018). Factors Influencing Hotels' Online Prices. *Journal of Hospitality Marketing & Management, 27*, 4, 443-464.

Nair, G. K. (2019). Dynamics of pricing and non-pricing strategies, revenue management. *International Journal of Hospitality Management 82*, 287-297.

Simon, H., Zatta, D., & Fassnacht, M. (2013). Price management. I: Strategia, analisi e determinazione del prezzo. Franco Angeli Editore.

Talluri, K. T., & Van Ryzin, G. J. (2004). The Theory and Practice of Revenue Management. International Series in Operations Research & Management Science. Springer.

Wang, X., Sun, J., & Wen, H. (2019). Tourism seasonality, online user rating and hotel price: A quantitive approach based on the hedonic price model. *International Journal of Hospitality Management, 79*, 140-147.

RINGRAZIAMENTI

In primis vorrei ringraziare il mio relatore, il Professor Aldo Goia, per avermi seguita passo dopo passo nella stesura del mio elaborato, dandomi preziosi consigli che mi hanno permesso di approfondire l'analisi del dataset a disposizione, attraverso la scoperta di nuovi modelli.

Vorrei ringraziare anche il mio correlatore, il Professor Graziano Abrate, per avermi dato il permesso di analizzare il subset di un dataset raccolto per una sua ricerca e per avermi presentato, durante il corso di Price Management, il tema del Revenue Management che mia ha incuriosita fin da subito.

Un ringraziamento speciale va ai miei genitori e a mio fratello per essermi sempre stati accanto e per avermi sostenuta soprattutto nei momenti di difficoltà. Grazie per avermi permesso di arrivare fino a qui. Ringrazio anche nonne, zii e cugini per esserci sempre stati e per avermi aiutato quando ne avevo bisogno.

Ringrazio gli amici di una vita che sono rimasti nonostante abbiano preso strade differenti, le nuove amicizie e le compagne dell'università con cui ho trascorso questi ultimi anni. Grazie a tutti per avermi supportata e sopportata soprattutto nei momenti di debolezza. Grazie per tutti i momenti passati insieme, per le risate, per i sorrisi, per i mille consigli, ma anche per gli scleri condivisi.