Alma Mater Studiorum · Università di Bologna

**SCUOLA DI SCIENZE**
**Corso di Laurea in Informatica per il Management**

# STRUMENTI DI MACHINE LEARNING IN STUDI DI CRIMINOLOGIA

**Relatore:**
**Dott.**
**ELENA LOLI PICCOLOMINI**

**Presentata da:**
**GERALDA NUSHI**

**Sessione III**
**Anno Accademico 2019/2020**

*Alla mia famiglia,*
*con affetto e gratitudine*

# Indice

# Elenco delle figure

# Introduzione

La presente tesi di laurea nasce da un interesse personale verso l'analisi statistica dei dati e il campo di ricerca in cui rientrano, ovvero la criminologia. Si tratta di un contesto, apparentemente lontano, ma costantemente presente nelle vicende di vita quotidiana. Questo aspetto, insieme alla curiosità nei confronti del tema dell'intelligenza artificiale e in particolare dell'applicazione del machine learning, ha posto le basi per la nascita di questa tesi di ricerca. Quando si parla di machine learning si parla di una particolare branca dell'informatica che può essere considerata un'applicazione dell'intelligenza artificiale. Il termine "Intelligenza artificiale" è stato coniato per la prima volta negli anni '50 e coinvolge tutte quelle macchine computazionali in grado di eseguire compiti caratteristici dell'intelligenza umana. L'apprendimento automatico rappresenta un mezzo per raggiungere l'intelligenza artificiale. Il termine "machine learning" è stato coniato successivamente all'AI, inteso come "la capacità di una macchina di apprendere senza essere programmata esplicitamente"[21]. Definire in maniera lineare il sistema di apprendimento automatico non è sempre possibile, data la differenziazione delle tecniche e lo sviluppo di algoritmi che ne allargano il campo di applicazione[5]. È uno dei settori tecnici in più rapida crescita di oggi, che si trova all'incrocio tra informatica e statistica e al centro dell'intelligenza artificiale e della scienza dei dati [29]. Viene definita come un sottogruppo dell'AI che verte sulla capacità delle macchine di ricevere una serie di dati, estrapolarli e creare conoscenza modificando gli algoritmi man mano che ricevono più informazioni sull'elaborazione in corso. Lo scopo del Machine learning è di migliorare autonomamente l'identificazione di pattern nei dati; si concentra sullo sviluppo di software per l'elaborazione di grosse moli di dati, utilizzate per apprendere nuove informazioni e fornire previsioni su di esse. In questo contesto, è interessante utilizzare tale strumento nell'analisi della realtà criminologica che presenta statistiche sempre in costante aumento. La possibilità di raccolta dei dati, di analisi e statistiche permette di far previsioni , di comprendere la realtà e in più di trovare alternative per restringere questa situazione. L'obiettivo della seguente tesi è quello di comprendere non solo le differenti applicazioni del machine learning ma nello specifico il rapporto tra machine learning e statistica che permette di far inferenza nel campo della criminologia. A tale proposito viene riportato un recente caso di studio con l'applicazione delle tecniche di regressione bayesiana a seguire vengono riportate implementazioni di analisi statistiche di regressione lineare e multivariata su datasets relativi a differenti tipologie di crimine. La struttura della tesi è definita di seguito:

- Il primo capitolo verterà l'approfondimento del concetto di machine learning, partendo da un excursus storico dell'origine , fino alle applicazioni attuali. Si approndisce la relazione fra statista e machine learning , per dar spazio al statistical learing come supporto fondamentale all'analisi dei dati.La definizione del concetto di apprendimento automatico farà riferimento inoltre alle tre categorie di machine learning : apprendimento supervisionato , apprendimento non supervisionato e apprendimento per rinforzo.Infine si propone un confronto con il deep learning, disciplina ad essa strettamente collegata.

- Il secondo capitolo presenterà un esempio di case study recente che descrive l'uso di tecniche di apprendimento automatico per fornire un approccio completamente probabilistico alla modellazione del crimine, tenendo conto dell'incertezza di previsione dei reati nonché sui parametri del modello. Verrà fatta prima un'introduzione sul fenomeno della criminalità e delle sue cause , motivazione che porta a cercar ausilio nelle tecniche automatizzate e statistiche. Nello specifico si fa riferimento all'implementazione di un approccio bayesiano alla modellazione della dipendenza tra i dati del reato e i fattori ambientali come le caratteristiche demografiche e spaziali.

- Nel terzo capitolo si porteranno come applicazione del metodo sopra descritto delle analisi statistiche su datasets recuperati autonomamente, che fanno riferimento a tipi di crimini differenti del territorio italiano e americano.L'analisi si dirama in due: Regressione univariata e regressione multivariata. Gli strumenti utilizzati sono quelli acquisiti nel corso di Studi di Statistica, da qui in seguito verranno fatte delle stime e delle previsioni cosi come il metodo di analisi vuole.

# Capitolo 1

# Machine Learning, storia di algoritmi in evoluzione

## 1.1 Definizione e terminologia

Il Machine Learning è un'applicazione dell'intelligenza artificiale (AI) che fornisce ai sistemi la capacità di apprendere e migliorare automaticamente dall'esperienza senza essere programmati esplicitamente. In particolare, ci si riferisce ad algoritmi che, basati su studi di probabilità e statistica, sviluppano autonomamente le loro conoscenze grazie ai pattern di dati ricevuti, senza il bisogno di avere degli input iniziali specifici da parte dello sviluppatore umano. Il processo di apprendimento inizia con osservazioni o dati, come esempi, esperienza diretta o istruzioni, al fine di cercare modelli nei dati e prendere decisioni migliori in futuro sulla base degli esempi forniti.Questo significa accantonare la classica programmazione esplicita, sistema in cui la componente umana programma un modello in base a comandi del tipo "if-then", a favore di un metodo in cui la macchina è in grado di stabilire da sola gli schemi da seguire per ottenere il risultato desiderato[2]. L'obiettivo principale è consentire ai computer di apprendere automaticamente senza intervento o assistenza umana e regolare le azioni di conseguenza. Pertanto, il vero fattore che distingue l'intelligenza artificiale è l'autonomia. Tom M. Mitchell ha fornito la definizione più citata di apprendimento automatico nel suo libro "Machine Learning": "Si dice che un programma apprende dall'esperienza E con riferimento a alcune classi di compiti T e con misurazione della performance P, se le sue performance nel compito T, come misurato da P, migliorano con l'esperienza E."[29] In poche parole, si potrebbe semplificare dicendo che un programma apprende se c'è un miglioramento delle prestazioni dopo un compito svolto. Questa definizione è rilevante poiché

fornisce una definizione operativa dell'apprendimento automatico, invece che in termini cognitivi. Fornendo questa definizione, Mitchell di fatto segue la proposta che Alan Turing fece nel suo articolo "Computing Machinery and Intelligence", sostituendo la domanda "Le macchine possono pensare?" con la domanda "Le macchine possono fare quello che noi (in quanto entità pensanti) possiamo fare?[40]
Turing portava avanti l'ideazione del suo metodo astratto di calcolo. Lo dimostra il fatto che paragonò gli stati interni della sua macchina agli stati mentali delle persone. L'intuizione era che i sistemi naturali potessero essere descritti computazionalmente da un insieme finito di funzioni, e lo scopo della scienza artificiale fosse quello di ricercarle e formalizzarle all'interno di un programma, che sarebbe stato computato da un calcolatore digitale.[19] Esistono vari algoritmi di apprendimento automatico , riportiamone alcuni dei più utilizzati: Regressione lineare, Regressione logistica,Classification and Regression Tree , Alberi decisionali,Naive Bayes,Random Forest,K-Nearest Neighbors,K-means clustering,Support Vector Machine. Questi algoritmi sono classificabili in due principali categorie di problemi: supervised learning e unsupervised learning.Tra quelli citati, solo il K-means rientra nei problemi di apprendimento non supervisionato, i restanti rientrano nella classe di apprendimento supervisionato .
Ci poniamo il quesito relativamente a come sia possibile che il computer apprenda automaticamente. La risposta è sicuramente attraverso i dati. Si inseriscono dati con attributi o caratteristiche differenti che gli algoritmi devono comprendere e dare un limite decisionale basato sui dati forniti. Una volta che l'algoritmo ha appreso e interpretato i dati, ovvero si è addestrato da solo, si può mettere l'algoritmo in fase di test e senza programmarlo esplicitamente, inserire un punto dati di test e aspettare dei risultati[2].
Il processo di creazione di un modello in genere prevede tre fasi principali [29]:

1. Fase di addestramento: questa è la fase in cui i dati di addestramento vengono utilizzati per addestrare il modello accoppiando l'input fornito con l'output previsto. Il risultato di questa fase è il modello di apprendimento stesso.

2. Convalida e fase di test: questa fase serve a misurare quanto è buono il modello di apprendimento che è stato addestrato e stimare le proprietà del modello, come misure di errore, richiamo, precisione. Questa fase utilizza un set di dati di convalida e l'output è un modello di apprendimento sofisticato.

3. Fase dell'applicazione: in questa fase, il modello è soggetto ai dati del mondo reale per i quali è necessario derivare i risultati.

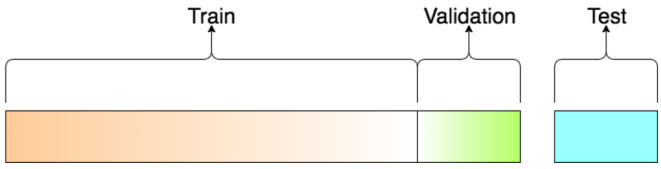



Figura 1.1: Rappresentazione del processo di costruzione del modello a partire dai dati

I dati utilizzati per costruire il modello finale di solito provengono da più set di dati ,in particolare se ne considerano tre. Un set di addestramento che viene implementato per creare un modello, un set di convalida e un set di test per convalidare il modello creato. L'addestramento applica un metodo di apprendimento supervisionato, ad esempio metodi di ottimizzazione come la discesa del gradiente o la discesa del gradiente stocastico [37]. In pratica, il set di dati di addestramento è spesso costituito da coppie di un vettore di input e il corrispondente vettore di output , dove la chiave di risposta è comunemente indicata come obiettivo (o etichetta ). Il modello corrente viene eseguito con il set di dati di addestramento e produce un risultato, che viene quindi confrontato con l' obiettivo. In base al risultato del confronto e all'algoritmo di apprendimento specifico utilizzato, i parametri del modello vengono adeguati. Successivamente, il modello adattato viene utilizzato per prevedere le risposte per le osservazioni in un secondo set di dati chiamato set di dati di convalida . La fase di convalida/test permette di stimare il livello di addestramento del modello che dipende dalla dimensione dei dati, dal valore che si desidera prevedere e immettere. Inoltre stima le proprietà del modello come ad esempio l'errore medio per predittori numerici. É importante sottolineare come il set di dati di convalida fornisce una valutazione imparziale di un adattamento del modello sul set di dati di addestramento durante l'ottimizzazione degli iperparametri del modello. Segue infine.la fase di applicazione in cui il modello appena sviluppato viene applicato ai dati del mondo reale. La fase di validazione è spesso suddivisa in due parti : nella prima parte, si guardano solo i modelli creati e si seleziona l'approccio più performante utilizzando i dati di convalida. Quindi si stima l'accuratezza dell'approccio selezionato. Da qui si separano i set di test da quelli di convalida per analizzare la stima del tasso di errore del modello finale[18]. I modelli per di Machine learning , così come quello di deep learing, hanno la

stessa finalità dell'intelligenza artificiale ovvero partire da determinati input ottenere e calcolare l'output.

## 1.2 Evoluzione del Machine Learning

Al giorno d'oggi parlare di apprendimento automatico, di intelligenza artificiale, di computer e macchine intelligenti sembra quasi la normalità, per arrivare ai risultati odierni la strada è stata molto complessa, perché divisa tra sperimentazioni e scetticismo. Le prime sperimentazioni per la realizzazione di macchine intelligenti risalgono agli inizi degli anni Cinquanta del Novecento, quando alcuni matematici e statistici iniziarono a pensare di utilizzare i metodi probabilistici per realizzare macchine che potessero prendere decisioni proprio tenendo conto delle probabilità di accadimento di un evento. Il primo grande nome legato al machine learning è sicuramente quello di Alan Turing, che ipotizzò la necessità di realizzare algoritmi specifici per applicarle a macchine in grado di apprendere.Nel 1952 Arthur Samuel scrive il primo programma "computer learning". Si tratta del gioco della dama: i computer migliorano le strategie di gioco quanto più gli umani interagiscono con essi.È stato condotto un grande lavoro per verificare il fatto che un computer possa essere programmato in modo da imparare a giocare una partita a dama migliore di quella giocata dalla persona che ha scritto il programma stesso.Ci riuscì, e nel 1962 il suo programma superò il campione di dama dello stato del Connecticut.[35] Nel 1957 vengono attivati i primi Neural Networks: la simulazione del cervello umano. In quegli stessi anni, anche gli studi sull'intelligenza artificiale, sui sistemi esperti e sulle reti neurali vedevano momenti di grossa crescita alternati da periodi di abbandono, causati soprattutto dalle molte difficoltà riscontrate nelle possibilità di realizzazione dei diversi sistemi intelligenti, nella mancanza di sussidi economici e dallo scetticismo che circondava spesso chi provava a lavorarci. Il più simile algoritmo a quello dei nostri tempi viene scritto nel 1967 utilizzando uno schema di riconoscimento che permette di costruire percorsi di visite per le città. Nel 1979 degli studenti della Stanford University inventano "Stanford Cart", una sorta di robot dotato di software che evita gli ostacoli in una stanza in maniera autonoma mappando i percorsi.A partire dagli anni Ottanta, una serie di interessanti risultati ha portato alla rinascita di questo settore della ricerca resa possibile da nuovi investimenti. Alla fine degli anni Novanta l'apprendimento automatico trova nuova linfa vitale in una serie di innovative tecniche legate ad elementi statistici e probabilistici: si trattava di un importante passo che permise quello sviluppo che ha portato oggi l'apprendimento automatico ad essere un ramo della ricerca riconosciuto e altamente richiesto. Il termine

fu coniato per la prima volta da Arthur Lee Samuel, scienziato americano pioniere nel campo dell'Intelligenza Artificiale, nel 1959 anche se, ad oggi, la definizione più accreditata dalla comunità scientifica è quella fornita da un altro americano, Tom Michael Mitchell, direttore del dipartimento Machine Learning della Carnegie Mellon University: «si dice che un programma apprende dall'esperienza E con riferimento a alcune classi di compiti T e con misurazione della performance P, se le sue performance nel compito T, come misurato da P, migliorano con l'esperienza E» [40]. Ciò segue la proposta di Alan Turing nel suo documento " Computing Machinery and Intelligence ", in cui la domanda "Le macchine possono pensare?" è sostituito dalla domanda "Le macchine possono fare ciò che noi (come entità pensanti) possiamo fare?" Sebbene possa essere utilizzato in diversi contesti, il machine learning di solito viene utilizzato per risolvere problemi trovando modelli nei dati che non possiamo vedere noi stessi; la quantità grezza e la crescita costante dei dati raccolti continuamente da tutte le piattaforme web, le applicazioni e i server crea la necessità di metodi in grado di influire su una serie di professioni e stili di vita[5].Una grande svolta si attua nel 1990 , quando nascono i primi algoritmi che analizzano grandi quantità di dati da cui poter trarre conclusioni.Si passa da un approccio basato sulla conoscenza ad uno basato sui dati. Nel 1997 IBM Deep Blue è il primo calcolatore che riesce a sconfiggere il campione del mondo a scacchi. Nel 2006 Geoffrey Hinton conia il termine "Deep Learning" per spiegare i nuovi algoritmi che permettono al computer di "vedere" e distinguere oggetti e testi in immagini e video. Nel 2010 Microsoft introduce software che possono tenere traccia di 20 caratteristiche umane ad un tasso di 30 volte al secondo, permettendo alle persone di interagire con il computer tramite movimenti e gesti. Le svolte vere e prorpie nel campo del Machine Learning si riconoducono all'utlimo decennio : Apple con Siri, Microsoft con Cortana e Amazon con Alexa, hanno introdotto gli assistenti virtuali. Sono servizi che interagiscono direttamente con gli umani, grazie al Natural Language Processing. Nel 2016 il gruppo di ricerca "Deep Mind" di Google e Oxford University applicano il Deep Learning ai programmi della BBC per creare un sistema di lettura labiale che è più preciso rispetto a un lettore professionale lipnet. Ogni anno centinaia di nuovi algoritmi di apprendimento vengono ideati e riversati sul mercato "E' il "capitalismo della sorveglianza", retto sugli enormi profitti generati dall'estrazione di dati che riguardano la quotidianità di tutti voi" sostiene Shoshana Zuboff in *Il capitalismo della sorveglianza*[**?**]. Ma tutti questi algoritmi, scrive Pedro Domingos, "si basano su un numero ristretto di idee fondamentali che rappresentano tutto quello che bisogna effettivamente sapere per capire in che modo il machine learning sta cambiando il mondo"[22]. In altre parole: il Machine Learning permette ai computer di imparare dall'esperienza : c'è apprendi-

mento quando le prestazioni del programma migliorano dopo lo svolgimento di un compito o il completamento di un'azione anche errata, partendo dall'assunto che anche per l'uomo vale il principio "sbagliando di impara".[15] È indubbio che negli ultimi anni la ricerca abbia ottenuto progressi enormi per quanto riguarda le forme di apprendimento intelligente.Gli smartphone, che accompagnano ogni momento della nostra giornata, sono dotati di un'applicazione classica di machine learning che è quella del riconoscimento vocale. Questo consente di attivare comandi tramite la voce.I vari Alexa e Google Home sono ormai sempre più presenti nelle nostre case. La profilazione che avviene quando navighiamo su internet non è altro che un ulteriore utilizzo dell'apprendimento automatico. Ci stupiamo spesso di aver appena parlato di un determinato prodotto e di trovarcelo subito dopo nelle pubblicità consigliate sui vari social network. Oppure, l'analisi delle ricerche effettuate in rete riconosce i nostri gusti e ci offre prodotti e servizi simili a quelli cercati. Anche l'ambiente ne può giovare, grazie a modelli predittivi che permettano una migliore gestione delle risorse.É evidente come le possibilità di sviluppo sono ancora potenzialmente infinite, legate a diversi settori di applicazione anche di uso comune, non solo scientifici o legati alla ricerca,l'unico ostacolo, tuttavia, al loro massimo sviluppo può essere l'uomo e la sua paura di venire rimpiazzato al lavoro dai robot e dall'intelligenza artificiale.

## 1.3 Statistica e Machine Learning

L'apprendimento automatico e la statistica sono discipline strettamente collegate. Senza statistica, non si può costruire un modello e non è possibile sviluppare una profonda comprensione e applicazione dell'apprendimento automatico. [39] Secondo Michael I.Jordan, le idee di machine learning, dai principi metodologici agli strumenti teorici, sono stati sviluppati prima in statistica non a caso identifica nel termine data science una nozione per definire l'intero campo di studi in questione.[17]. D'altra parte Leo Breiman ha distinto due paradigmi statistici di modellazione: modello basato sui dati e modello basato sugli algoritmi, dove "modello basato sugli algoritmi" indica approssimativamente algoritmi di apprendimento automatico come la foresta casuale.
Queste due discipline tuttavia non sono interscambiabili come erroneamente viene percepito. La principale differenza viene individuata nello scopo.[34] I modelli di machine learning sono progettati per fornire previsioni più accurate possibili. I modelli statistici, invece, sono progettati per far inferenza sulle relazioni tra le variabili. Molti modelli statistici, tuttavia, possono fare previsioni, ma l'accuratezza predittiva non è il loro punto di forza [23]. Allo

stesso modo, i modelli di apprendimento automatico forniscono vari gradi di interpretabilità, dalla regressione lazo altamente interpretabile alle reti neurali impenetrabili , ma generalmente sacrificano questo aspetto per il potere predittivo. Il compromesso tra potere esplicativo (Interpretabilità) e potere predittivo (Flessibilità) dei modelli è illustrato dalla relazione negativa in questa figura proposta da Gareth James nel manuale "An Introduction to Statistical Learning.

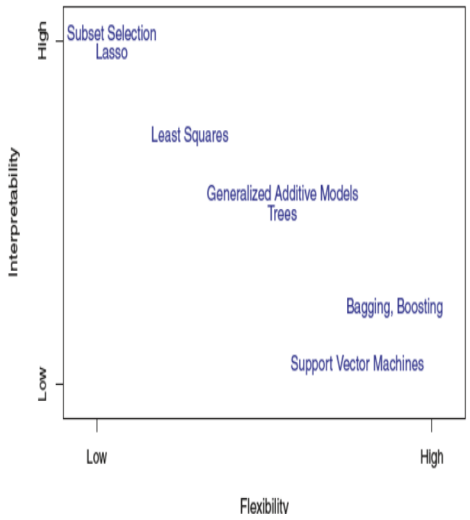



Figura 1.2: Rappresentazione della Flessibilià e Interpretabilità a confronto

**Previsione** e **spiegazione** sono le chiave su cui poggia questa distinzione concettuale.[12]I modelli statistici sono usati per descrivere i modelli così come erano in origine durante il periodo di raccolta dei dati e quantificare la relazione fra le variabili, mentre i modelli di apprendimento automatico sono usati per progettare modelli nel futuro.Un aspetto fondamentale concerne la mole dei dati in campo, difatti il machine learning richiede una ampia gamma di dati rispetto ai modelli statistici poichè l'accuratezza dei modelli predittivi più potenti, come le reti neurali e le foreste casuali ,sono in grado di gestire ulteriori migliaia o milioni di osservazioni. Al contrario, i modelli statistici consentono inferenze e fanno previsioni corrette su centinaia di osservazioni o poco più.[26]
Dal punto di vista dell'analisi dei dati tradizionale è possibile la seguente distinzione:

- L'apprendimento automatico è un algoritmo in grado di apprendere dai dati senza fare affidamento sulla programmazione basata su regole.

- La modellizzazione statistica è una formalizzazione delle relazioni tra variabili nei dati sotto forma di equazioni matematiche.

Tuttavia si può trovare un punto di accordo e lo si ritrova in un obiettivo comune che è quello di apprendere dai dati, come sostiene Larry Wasserman .[41] Egli afferma come gli stessi concetti abbiano nomi diversi nei due campi di analisi:

| Statistiche | Apprendimento automatico |
|---|---|
| Stima | Apprendimento |
| Classificatore | Ipotesi |
| Punto dati | Esempio / istanza |
| Regressione | Apprendimento supervisionato |
| Classificazione | Apprendimento supervisionato |
| Covariate | Caratteristica |
| Risposta | Etichetta |

Figura 1.3: Sintassi a confronto tra statistica e machine learning

Risultano essere relazionate se pensiamo alla finalità di trarre conoscenza o approfondimenti dai dati[6]. Tuttavia provengono da campi differenti :

- L'apprendimento automatico è un sottocampo dell'informatica e dell'intelligenza artificiale. Si occupa di costruire sistemi che possono apprendere dai dati, invece che da istruzioni programmate esplicitamente.

- Un modello statistico è un sottocampo della matematica.

Al giorno d'oggi, sia l'apprendimento automatico che le tecniche statistiche vengono utilizzate nel riconoscimento di modelli, nella scoperta della conoscenza e nel data mining. I due campi stanno convergendo sempre di più anche se la figura sottostante potrebbe mostrarli come quasi esclusivi. Alcuni statistici hanno adottato metodi provenienti dall'apprendimento automatico, il che ha portato alla creazione di una disciplina combinata chiamata "apprendimento statistico".[11]

### 1.3.1 Statistical Learning

La teoria dell'apprendimento statistico affronta il problema di trovare una funzione predittiva basata sui dati.Essa si compone di una serie di strumenti per la modellazione e la comprensione di set di dati complessi. Ci poniamo all'interno di un'area di recente sviluppo in statistica che è strettamente legata a sviluppi nell'informatica e, in particolare, nell'apprendimento automatico.Con l'esplosione dei "Big Data", l'apprendimento statistico è diventato un campo molto utilizzato in molte aree scientifiche, nonché in ambito criminologico per la previsione degli eventi. Gli strumenti utilizzati possono essere classificati come supervisionati o non supervisionati.[23] Tale distinzione viene ereditata dal machine learning; in questo contesto, tuttavia, per apprendimento supervisionato intendiamo la costruzione di un modello statistico per prevedere, o stimare, un output basato su uno o più input. Problemi di questa natura si verificano in campi diversi come affari, medicina, astrofisica, ordine pubblico e criminologia. Con l'apprendimento statistico non supervisionato intendiamo invece il fatto che ci siano input ma nessun risultato di supervisione, nonostante ciò è possibile imparare le relazioni e la struttura da tali dati. La teoria dell'apprendimento statistico ha portato ad applicazioni di successo in campi come la visione artificiale , il riconoscimento vocale e la bioinformatica. L'obiettivo di questa ampia disciplina è l'inferenza, sotto due aspetti rilevanti quali: la comprensione o interpretabilità e la previsione. La comprensione rientra nelle categorie prima descritte , tra cui l'apprendimento supervisionato , apprendimento non supervisionato , l'apprendimento online , e apprendimento per rinforzo . Dal punto di vista della teoria dello statistical learning, l'apprendimento supervisionato è meglio compreso ed implica l'apprendimento da un insieme di dati di formazione . Ogni punto dell'addestramento è una coppia input-output, in cui l'input si associa a un output. Il problema di apprendimento consiste nell'inferenza della funzione che mappa tra l'input e l'output, in modo tale che la funzione appresa possa essere utilizzata per prevedere l'output dall'input futuro. A seconda del tipo di output, i problemi di apprendimento supervisionato sono problemi di regressione o problemi di classificazione. I problemi di classificazione sono quelli per cui l'output sarà un elemento di un insieme discreto di etichette. La classificazione è molto comune per le applicazioni di machine learning: nel riconoscimento facciale , ad esempio, un'immagine del viso di una persona sarebbe l'input e l'etichetta di output sarebbe il nome di quella persona. L'input sarebbe rappresentato da un grande vettore multidimensionale i cui elementi rappresentano i pixel nell'immagine.
Se l'output , invece, assume un intervallo continuo di valori, si tratta di un problema di regressione, un metodo statistico , oggetto e strumento metodo-

logico di applicazione pratica dell'apprendimento automatico, L'esempio più ovvio è il caso della regressione lineare che in questo contesto è possibile addestrare e ottenere lo stesso risultato di un modello di regressione statistica che mira a ridurre al minimo l'errore quadrato tra i punti dati. Generalmente per il modello statistico, troviamo una linea che minimizza l'errore quadratico medio, assumendo che i dati siano un regressore lineare con l'aggiunta di un rumore casuale, che è tipicamente di natura gaussiana. Non sono necessari né addestramento né set di test. Per molti casi, specialmente nella ricerca , il punto del nostro modello è caratterizzare la relazione tra i dati e la nostra variabile di risultato, non fare previsioni sui dati futuri. Si tratta della procedura di inferenza statistica, al contrario di previsione.[32] Tuttavia, possiamo eventualemente utilizzare questo modello per fare previsioni e questo potrebbe essere il suo scopo principale, ma il modo in cui il modello viene valutato non coinvolgerà un set di test e comporterà invece la valutazione della significatività e della robustezza dei parametri del modello stesso.[34] Interessante notare come sono necessari metodi statistici per lavorare efficacemente attraverso un progetto di modellazione predittiva di machine learning . Una modellazione predittiva si compone dei seguenti passi:

1. Inquadratura del problema: è la selezione del tipo di problema, ad esempio regressione o classificazione, e verifica la struttura e i tipi di input e output per il problema.I metodi statistici che possono aiutare nell'esplorazione dei dati durante la definizione di un problema includono:

    - Analisi esplorativa dei dati: Riepilogo e visualizzazione per esplorare visualizzazioni ad hoc dei dati.
    - Data mining : Rilevamento automatico di relazioni e modelli strutturati nei dati.

2. Comprensione dei dati: Comprendere i dati significa avere una conoscenza approfondita sia delle distribuzioni delle variabili sia delle relazioni tra le variabili.

3. Pulizia dei dati: attraverso il rilevamento dei valori anomali,ovver osservazioni lontane dal valore atteso e l'inputazione per inserire valori mancanti

4. Selezione dei dati : selezione dei dati rilevanti per la modellazione attraverso le tecniche di Campionamento dei dati e della selezione delle caratteristiche.

5. Preparazione dei dati: attraverso il ridimensionamento, la codifica, e le trasformazioni.

6. Valutazione del modello : si effettua la stima dell'abilità del modello quando si effettuano previsioni su dati non visti durante l'addestramento del modello

7. Configurazione del modello :un determinato algoritmo di apprendimento automatico ha spesso una serie di iperparametri che consentono di adattare il metodo di apprendimento a un problema specifico.L'interpretazione e il confronto dei risultati tra diverse configurazioni di iperparametri viene effettuato attravesto o test di ipotesi o statistiche di stima.

8. Selezione del modello :uno dei tanti algoritmi di apprendimento automatico può essere appropriato per un dato problema di modellazione predittiva.

9. Presentazione del modello: una volta che un modello finale è stato addestrato, può essere presentato alle parti interessate prima di essere utilizzato o distribuito per fare previsioni effettive su dati reali.

10. Previsioni del modello: utilizzare un modello finale per fare previsioni per nuovi dati di cui non conosciamo il risultato reale.

La macchina ha il compito di costruire un modello probabilistico generale dello spazio delle occorrenze, in maniera tale da essere in grado di produrre previsioni sufficientemente accurate quando sottoposta a nuovi casi. Approfondiamo nello specifico come la statistica interviene nel campo dell'apprendimento automatico. In questo contesto, la componente statistica applicata viene analizzato nel metodo della regressione, regressione bayesiana utilizzata nell'analisi del caso di studio , e regressione lineare e multipla nell'analisi dei datasets proposti per determinare la significatività del modello. La piattaforma a supporto dell'aprendimento autiomatico che viene utilizzato è R.

## 1.4 Modelli di Machine Learning

Le funzioni dell'apprendimento automatico sono tipicamente classificate in tre principali macrocategorie [24].La distinzione poggia sulla natura degli input su cui si basa l'apprendimento o della funzione di valore con la quale il sistema giudica le azioni prese.[38]
Queste cateogrie sono :

- supervised learning: consiste nel fornire al sistema informatico della macchina una serie di nozioni specifiche e codificate, ossia di modelli ed esempi che permettono di costruire un vero e proprio database di informazioni e di esperienze. In questo modo, quando la macchina si trova di fronte ad un problema, non dovrà fare altro che attingere alle esperienze inserite nel proprio sistema, analizzarle, e decidere quale risposta dare sulla base di esperienze già codificate. Gli algoritmi che fanno uso di apprendimento supervisionato vengono utilizzati in molti settori, da quello medico a quello di identificazione vocale: essi, infatti, hanno la capacità di effettuare ipotesi induttive, ossia ipotesi che possono essere ottenute scansionando una serie di problemi specifici per ottenere una soluzione idonea ad un problema di tipo generale. Attraverso l'ottimizzazione iterativa di una funzione oggettiva , gli algoritmi di apprendimento supervisionato apprendono una funzione che può essere utilizzata per prevedere l'output associato a nuovi input.Si possono aver due differenti situazioni: classificazione e regressione, che anticipiamo evidenziando che si parla di classificazione quando la variabile output desiderata è categorica, mentre la regressione si ha quando la variabile output desiderata è quantitativa.[15]

- unsupervised learning: a differenza dell'apprendimento supervisionato, non ha etichette di allenamento per i campioni di formazione.[13] Esso prevede invece che le informazioni inserite all'interno della macchina non siano codificate, ossia la macchina ha la possibilità di attingere a determinate informazioni senza avere alcun esempio del loro utilizzo e, quindi, senza avere conoscenza dei risultati attesi a seconda della scelta effettuata. L'apprendimento senza supervisione offre maggiore libertà di scelta alla macchina che dovrà organizzare le informazioni in maniera intelligente e imparare quali sono i risultati migliori per le differenti situazioni che si presentano. Il compito dell'apprendimento non supervisionato più comune è il clustering: che consiste nel rilevamento di cluster potenzialmente utili di esempi di input; è uno dei metodi utilizzati nell'apprendimento senza supervisione, e consiste nell'analizzare set di dati con alcuni attributi in comune per estrapolare delle relazioni tra di essi.

- reinforcement learning: rappresenta probabilmente il sistema di apprendimento più complesso, che prevede che la macchina sia dotata di sistemi e strumenti in grado di migliorare il proprio apprendimento e, soprattutto, di comprendere le caratteristiche dell'ambiente circostante.[29] In questo caso, quindi, alla macchina vengono forni-

ti una serie di elementi di supporto, quali sensori, telecamere, GPS eccetera, che permettono di rilevare quanto avviene nell'ambiente circostante ed effettuare scelte per un migliore adattamento all'ambiente intorno a loro. Questo tipo di apprendimento è tipico delle auto senza pilota, che grazie a un complesso sistema di sensori di supporto è in grado di percorrere strade cittadine e non, riconoscendo eventuali ostacoli, seguendo le indicazioni stradali e molto altro

Un'ulteriore categorizzazione dell'apprendimento automatico si rileva quando si considera l'output desiderato del sistema di apprendimento automatico:

- Nella classificazione, gli output sono divisi in due o più classi e il sistema di apprendimento deve produrre un modello che assegni gli input non ancora visti a una o più di queste. Questo viene affrontato solitamente in maniera supervisionata. Il filtraggio anti-spam è un esempio di classificazione, dove gli input sono le email e le classi sono "spam" e "non spam".

- Nella regressione, che è anch'essa un problema supervisionato, l'output e il modello utilizzati sono continui. Si dispone di un numero di variabili predittive (descrittive) e una variabile target continua (il risultato). In questo tipo di problema si cerca di trovare una relazione tra queste variabili al fine di prevedere un risultato. Un esempio è la predizione del valore del tasso di cambio di una valuta nel futuro, dati i suoi valori in tempi recenti. Altri esempi saranno proposti sperimentalemente nei capitoli successivi.

- Nel clustering un insieme di input viene diviso in gruppi, non noti prima; viene , per questo motivo, considerato un compito non supervisionato che consiste nell'analizzare set di dati con alcuni attributi in comune per estrapolare delle relazioni tra di essi. Il clustering si occupa di trovare un'etichetta ai dati che ne sono sprovvisti, quindi invece di rispondere al feedback, trova similitudini o elementi comuni nei dati prendendo una decisione in base alla presenza o l'assenza di tali caratteristiche in ogni nuovo dato analizzato. Nel data mining, che è l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di informazioni utili da grandi quantità di dati attraverso metodi automatici o semi-automatici, viene spesso utilizzato il clustering, per esempio per scoprire chi sono i clienti che hanno maggiore propensione di acquisto su certi prodotti o campagne pubblicitarie. Uno degli sviluppi più recenti è il social data mining, ovvero la pratica di analizzare le informazioni generate dalle reti sociali e dai contenuti che gli

utenti generano su di essi per estrapolare degli schemi che poi verranno utilizzati per scopi prettamente pubblicitari.

## 1.5 Deep Learning e Machine Learning

Il Machine Learning e il Deep Learning possono essere considerati come insiemi di algoritmi e metodi di programmazione a servizio dell'intelligenza artificiale. La differenza risiede nell'uso della programmazione in quanto il Deep Learning usa strettamente le reti neurali per simulare il comportamento cellulare del nostro cervello[2]. È una sottocategoria dell'apprendimento automatico da cui eredita i concetti di apprendimento supervisionato, non supervisionato e di rinforzo. Come l'idea dell'intelligenza artificiale è stata ispirata dal cervello umano, il deep learning è stato ispirato da reti neurali artificiali e le reti neurali artificiali comunemente note come ANN sono state ispirate da reti neurali biologiche umane.[13] È uno degli approcci all'apprendimento automatico che ha preso spunto dalla struttura del cervello, ovvero l'interconnessione dei vari neuroni.Il neurone umano è il paradigma computazionale che nutre il deep learning e lo fa attraverso le famose Reti Neurali Artificiali. Comprendere il loro significato è molto semplice. Una rete neurale cerca di riprodurre il funzionamento del neurone umano, ovvero tutti quei processi che avvengono nel cervello durante la fase di apprendimento e quella successiva del riconoscimento.La pura e semplice esperienza guida l'apprendimento e offre al cervello i dati necessari per comprendere.[3] Tale approccio tenta di modellare matematicamente il modo in cui il cervello umano processa luce e suoni e li interpreta in vista e udito: gli stimoli di occhi ed orecchie, attraversando il cervello umano, vengono inizialmente scomposti in concetti semplici e man mano ricostruiti in concetti sempre più complessi ed astratti.La macchina deve poter offrire un valido paradigma, ovvero offrire un modo di "pensare", simile al funzionamento dei neuroni umani. Tant'è vero che una delle applicazioni più affermate del Deep Learning sono appunto la computer vision ed il riconoscimento vocale.

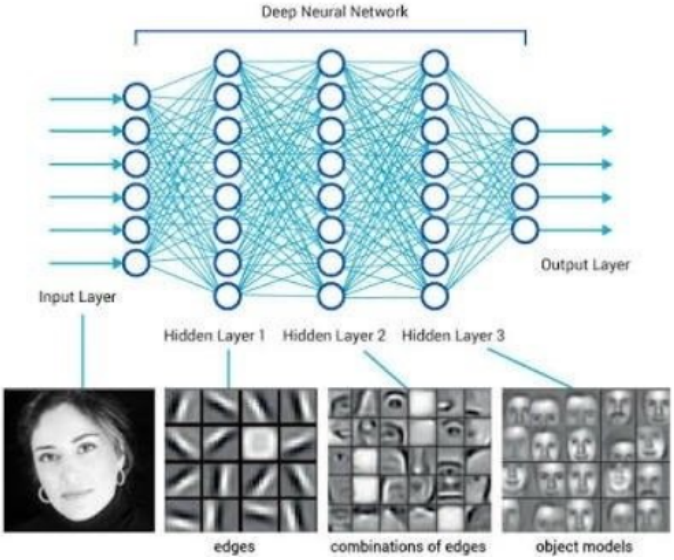



Figura 1.4: Rappresentazione del processo di deep learning

Altri approcci includono la programmazione logica induttiva, il clustering e le reti bayesiane. Il deep learning è uno dei modi per eseguire il machine learning utilizzando enormi modelli di reti neurali con varie unità di elaborazione. L'apprendimento approfondito sfrutta i progressi computazionali e tecniche di allenamento per apprendere modelli complessi attraverso una enorme quantità di dati. Le applicazioni comuni includono l'image e lo speech recognition.[42] Il concetto di apprendimento approfondito viene a volte indicato semplicemente come "rete neurale profonda", in riferimento ai numerosi livelli coinvolti:

- Livello di input: può essere pixel di un'immagine o dati di una serie temporale
- Livello nascosto: comunemente noto come pesi che vengono appresi durante il training della rete neurale
- Livello di output: il livello finale ti fornisce principalmente una previsione dell'input che hai inserito nella tua rete

**Aspetti a confronto tra deep learning e machine learning**
Il primo aspetto da prendere in considerazione è il funzionamento. Il deep learning è un sottoinsieme dell'apprendimento automatico che prende i dati come input e prende decisioni intuitive e intelligenti utilizzando una rete neurale artificiale impilata a livello di livello. D'altra parte, il machine learning,

essendo un super-set di deep learning, prende i dati come input, li analizza, cerca di dar loro un senso (decisioni) in base a ciò che ha appreso durante l'addestramento. In più il deep learning è considerato un metodo adatto per estrarre caratteristiche significative dai dati grezzi. Per quanto concerne la dipendenza dai dati,gli algoritmi di machine learning spesso funzionano bene anche se il set di dati è piccolo, il deep learning, tuttavia, tratta grandi quantità di dati , più dati tratta, meglio è probabile che funzioni. Si dice spesso che con più dati la profondità della rete (numero di strati) aumenta, così come aumentano i calcoli.

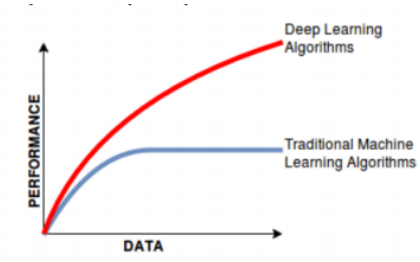



Figura 1.5: Performance degli algoritmi a confronto

Dalla figura si può vedere come all'aumentare dei dati, le prestazioni degli algoritmi di deep learning aumentano rispetto ai tradizionali algoritmi di machine learning in cui le prestazioni si saturano prima nonostante i dati vengono aumentati.
La rappresentazione dei dati avviene a livello gerarchico e soprattutto a livelli diversi tra loro, riuscendo a elaborarli e a trasformarli. Questo fattore è interessante poichè ci consente di assistere ad una macchina che riesce a classificare i dati in entrata (input) e quelli in uscita (output), evidenziando quelli importanti ai fini della risoluzione del problema e scartando quelli che non rilevanti .La rivoluzione apportata dal deep learning consiste essenzialmente nella capacità, simile a quella umana, di elaborare i dati, le proprie conoscenze a livelli che non sono affatto lineari. Grazie a questa facoltà, la macchina apprende e perfeziona funzionalità sempre più complesse.
Dopo aver compreso come agisce il deep learning sorge spontaneo il quesito relativo ai vari campi di applicazione di questo sistema di approfondimento. Pensiamo al campo medico , fino ad arrivare per esempio alla guida automatica.L'applicazione del concetto delle reti neurali in ambito medico è molto semplice perché i medici si servono già degli algoritmi, soprattutto in ambito specialistico.La guida automatica invece consente di riconoscere gli ostacoli in entrambi i lati della carreggiata, grazie all'ausilio di sensori e telecamere in grado di elaborare le immagini. La computer vision in questo caso riproduce

la vista umana, riconoscendo l'ambito nel quale si sta muovendo e fornendo tutte le indicazioni utili per muoversi in sicurezza.Il futuro del deep learning si nutre del concetto di interazione tra uomo e macchina: possibilità che offrire all'uomo la capacità di essere capito dalla macchina, attraverso la comprensione del linguaggio orale e dei gesti. Lo studio e lo sviluppo degli algoritmi intelligenti, difatti, è finalizzato alla creazione di macchine pensanti.[3]

# Capitolo 2

# Applicazione del machine learning alla criminologia

## 2.1 Analisi della Criminalità

L'analisi del fenomeno della criminalità solitamente ha preso in considerazione la figura del criminale e le statistiche inerenti questa tematica a partire dal secolo XIX. Per oltre 150 anni , i criminologi hanno cercato di capire le cause del verificarsi del crimine , le coordinate temporali e geografiche. Nella maggior parte dei casi, questo esercizio prevalentemente di scienze sociali si è incentrato sulla convinzione che comprendere meglio chi commette un crimine significa massimizzare le possibilità che la politica di giustizia sociale e penale possa essere progettata in modo ottimale per migliorare la prevenzione, mitigare i rischi e gestire l'allocazione efficiente delle risorse scarse.[16] Le prime generazioni di indagini si focalizzano soprattutto sul numero oscuro dei reati, mentre le indagini così dette di seconda generazione pongono maggiore enfasi sulla percezione della sicurezza in quanto elemento essenziale nel contribuire alla qualità della vita. Attraverso questa indagine è possibile definire l'entità e la diffusione del fenomeno della criminalità rispetto ai reati rilevati, è possibile inoltre rilevare la percentuale del sommerso, evidenziare quali sono i gruppi della popolazione più a rischio di subire furti, rapine, aggressioni o minacce e violenze. Vi è modo anche di calcolare qual è il danno e la perdita associata a questi reati individuando le modalità con cui si sono verificati, si può conoscere ,inoltre ,la relazione con l'autore del reato e capire cosa espone maggiormente le vittime. Le indagini di vittimizzazione permettono di calcolare tre interessanti indicatori per cogliere la misura del fenomeno della criminalità e della sua pervasività: ovvero la prevalenza, l'incidenza e la concentrazione. Tali indagini permettono anche di evidenziare qual è la

popolazione più a rischio di subire i reati. La distribuzione del rischio non è omogenea, bensì differenziata nel territorio a seconda del tipo di reato preso in considerazione[30].

## 2.2 Previsione del Crimine

La previsione del crimine in tempo reale, con l'intelligenza artificiale, è oggi una priorità per la comunità scientifica che si impegna a sviluppare modelli statistici sempre più precisi ed efficaci. Gli obiettivi degli esperti incaricati di sviluppare gli algoritmi riguardano in primis l'individuazione di luogo e tempo di un probabile crimine, evidenziando "hot post" in una mappa geografica, ciclicamente aggiornata e controllata dai funzionari di polizia per poi porre la concentrazione sugli aspetti di causa e vittima.[20] Analizziamo lo scenario che le nuove tecnologie prospettano in questo ambito e che stanno aprendo nuovi fronti per la sicurezza dei cittadini. I dipartimenti di polizia che utilizzano l'analisi predittiva per la prevenzione del crimine sono attivi nelle nostre città da anni. Tutti i progetti di predictive policing usano dati storici (soprattutto quelli che si riferiscono ad un passato molto vicino) i quali, se incrociati con le tecnologie avanzate nel modo giusto, possono condurre alla definizione di trend dei comportamenti criminali, facilmente prevedibili e individuabili[14].
L'approccio più comune per visualizzare i modelli geografici di criminalità è la mappatura dei punti. In applicazione informatica, se vengono definite opportunamente informazioni come il codice che descrive il tipo di reato, data e ora, possono essere selezionati punti sulla mappa geografica in modo semplice e veloce. Tuttavia,l'interpretazione di modelli spaziali e "hot post" in una mappa geografica non è facile soprattutto se i set di dati sono di grandi dimensioni. Le tecniche che possono essere di supporto in questo contesto rientrano nell'ambito statistico:

- Hierarchical Clustering : metodo che utilizza una tecnica di analisi che consiste nell'identificare i gruppi di un numero minimo più vicini ai punti definiti dall'utente.

- K-means clustering: trova il miglior posizionamento dei centri K e quindi assegna ogni punto al centro che si trova più vicino

Entrambi i metodi sono plausibili e, soprattutto, offrono l'opportunità di esplorare la natura della criminalità in queste aree in modo più dettagliato. Tuttavia, nessuno dei due metodi presenta l'opportunità immediata di dare la priorità a principali "hot post" della criminalità per coaudiuvare il corpo

dei polizia nella prevenzione mirata.
Modelli rilevanti per l'analisi spaziale del crimine sono presentati da Chainey [4] Eck, Leitner, Perry e altri, che includono Kernel Density Estimation (KDE), clustering K-means, ellissi di copertura e altre euristiche per l'identificazione dei punti caldi e nell'analisi spazio-temporale del crimine.[7] Perry insieme ad altri studiosi descrive in dettaglio molte altre tecniche per identificare la stagionalità e la periodicità a diverse risoluzioni in serie temporali di intensità del crimine, tuttavia non riescono a esplorare rappresentazioni multivariate del crimine che accoppiano gli effetti demografici e ambientali.[8] Un altro approccio spazio-temporale è quello di Flaxman che utilizza il processo gaussiano spazio-temporale sull'intensità di una distribuzione di Poisson dei conteggi degli eventi per spiegare il verificarsi del crimine. Egli combina funzioni di covarianza spazio-temporale con componenti periodiche che possono catturare la stagionalità nel dominio temporale.[1] Negli ultimi anni, difatti, i processi gaussiani sono stati ampiamente utilizzati nell'apprendimento automatico come priorità su funzioni sconosciute, per modellare fenomeni correlati spazialmente e temporalmente.
Con riferimento ad alucne delle numerose sperimantazioni in campo criminologico ci poniamo il seguente quesito.
**Il machine learning può prevenire il crimine?**
Nel noto film *Minority Report* i criminali venivano arrestati ancora prima di commettere un reato perché c'era chi li prevedeva in anticipo. Bene, c'è chi ritiene che questo sia parzialmente possibile anche nel mondo reale grazie alle potenzialità del machine learning. È il campo del predictive policing, la "sorveglianza predittiva" che applica algoritmi di intelligenza artificiale ai dati storici della criminalità. Il presupposto di base su cui si regge questa visione è che ci siano regolarità identificabili nella distribuzione dei crimini in una città. Quindi si parte da una base dati storica dei crimini e, arricchendola nel tempo con i nuovi reati, si arriva a fare previsioni su dove un determinato tipo di crimine si verificherà con più probabilità. Più precisamente, si ottiene la probabilità che un dato crimine si verifichi in un dato lasso di tempo in una data area. Su questa previsione, le forze dell'ordine dovrebbero poter organizzare meglio i propri turni di ronda per avere la massima possibilità di cogliere un reato in corso. In realtà non sappiamo quali forze dell'ordine utilizzino già soluzioni di predictive policing, perché quasi nessuna lo comunica chiaramente. E questo impedisce di valutare quanto l'idea di prevedere i crimini sia davvero efficace[36].Gli approcci sopra menzionati modellano il crimine unicamente in funzione dello spazio e del tempo, ignorando altre fonti di spiegazione. Sebbene queste tecniche possano portare a buone prestazioni predittive, non aiutano a comprendere i fattori che guidano la criminalità, necessari per l'allocazione ottimale delle scarse risorse per la prevenzione

del fenomeno. Capire la criminalità ha spesso indirizzato la concentrazione su informazioni longitudinali sulla popolazione, comportamenti e ambienti, tra cui istruzione, occupazione, strutture familiari, salute e contatti con il sistema di polizia e giustizia. Gli ultimi sviluppi nella scienza dei dati e nell'apprendimento automatico offrono nuovi modi per prevedere l'incidenza della criminalità e per comprendere gli impatti delle caratteristiche sociali e individuali sul comportamento criminale.[27] In particolare, studiosi come Liu e Brown propongono un modello di densità di transizione che tiene conto degli attributi demografici, economici, sociali, vittima e spaziali dell'attività criminale.[9] Gli algoritmi statistici e di apprendimento automatico sono sempre più utilizzati per documentare le decisioni di grandi dimensioni che hanno forte impatto sulla vita degli individui come ad esempio la polizia predittiva, il rischio pre-processuale,valutazione della recidiva e del rischio di violenza durante la detenzione.Sono stati proposti diversi approcci per correggere modelli predittivi ingiusti. L'approccio più semplice è escludere le variabili protette dall'analisi;ovviamente, escludere semplicemente una variabile protetta non è sufficiente per evitare previsioni discriminatorie, poiché qualsiasi variabile inclusa correlata con le variabili protette contengono ancora informazioni sulla caratteristica protetta. Un'alternativa è modificare la variabile di risultato usando, ad esempio, un classificatore ingenuo di Bayes per classificare ogni osservazione e perturbare il risultato in modo tale che le previsioni prodotte dall'algoritmo siano indipendenti dalla variabile protetta. A questo proposito i modelli statistici per la gestione della scarsità di dati, come i modelli gerarchici bayesiani, possono essere facilmente implementati per superare questo problema. Utilizzando un approccio di modellazione statistica, osserviamo come applicare tali aggiustamenti: se viene applicato a un set di dati di recidività, il nostro approccio ha successo e crea previsioni di recidività indipendenti dalla variabile protetta, con perdite minime nell'accuratezza predittiva. In questo lavoro mostriamo come costruire modelli completamente probabilistici che siano in grado di rispondere a domande importanti sul crimine, come: qual è la probabilità che si verifichi un crimine in un luogo particolare? Quali sono le caratteristiche della popolazione che influenzano l'incidenza della criminalità? A confermare tale considerazione portiamo un esempio completo.

## 2.3 CASE STUDY

**Regressione bayesiana semi-parametrica spaziale-demografica.**
Il caso di studio preso in analisi fornisce un esempio dell'applicabilità delle tecniche di apprendimento automatico , in questo specifico contesto fornendo

un approccio completamente probabilistico alla modellazione del crimine che porta alla previsione dei reati. I dati che vengono presi in considerazione sono relativi ad aggressioni legate alla violenza domestica , furto con scasso e furto di veicoli a motore, nello stato del New South Wales (NSW), Australia.[27] L'analisi predispone i seguenti contributi:

1. Fornire una metodologia quantitativa basata sull'evidenza che mette in relazione il crimine con le informazioni ambientali e demografiche utilizzando algoritmi di apprendimento automatico nella gestioni dei dati in questione.

2. Combina tecniche parametriche e non parametriche per modellare la dipendenza tra l'incidenza del crimine e fattori specifici della posizione

3. Proporre un modello completamente probabilistico, in grado di quantificare l'incertezza nelle previsioni

La metodologia di analisi applicata si basa sulla regressione lineare Bayesiana, viene utilizzata la media a posteriori per la previsione , viene fatta ,inoltre ,l'inferenza attraverso la distribuzione marginale a posteriori della quantità di interesse. Vengono definiti i parametri indicati con $\theta$ , sui quali viene effettuata l'inferenza tramite la distribuzione di probabilità a posteriori indicata
con p($\theta$ | D ) dove D è un set di dati e la notazione | significa "condizionato a". Questa distribuzione a posteriori è data dal teorema di Bayes come segue:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}. \tag{2.1}$$

Il termine p($\theta$ | D ) rappresenta la probabilità che i dati vengano generati. Dati i parametri è nota come distribuzione di probabilità a priori che codifica la conoscenza preliminare di questi parametri. Il termine p($\theta$ | D ) è la distribuzione di probabilità marginale dei dati. È una costante normalizzante ed è indipendente dai parametri $\theta$ .Riportiamo la descrizione del modello di regressione per il flusso di criminalità assumendo come variabile indipendente la variabile spaziale. Consideriamo l'analisi in due punti:

- Regressione lineare bayesiana con errori iid

- Regressione lineare bayesiana con dipendenza spaziale

Creiamoci prima una panoramica d'insieme dei dati che andremo a porre in analisi.

FONTI:

- NSW Bureau of Crime Statistics and Research (BOCSAR)
- Australian Bureau of Statistics (ABS)

DATI:

- violenza domestica
- Furto con scasso
- Furto di veicoli a motore

LOCALITA':

- Australia: New South Wales (NSW).

TECNICA DI ANALISI:

- regressione bayesiana semi-parametrica spaziale-demografica

VARIABILI :

- tasso di criminalità (dipendente)
- località
- variabili esplicative basate sulla posizione e su caratteristiche demografiche , ambientali , la densità dei trasporti, etc.

L'obiettivo a cui questa analisi conduce è duplice:

1. Valutare le prestazione predittive del modello teorico di regressione
2. Valutare la capacità del modello di fare inferenze su specifici fattori causali

### 2.3.1 Regressione lineare bayesiana con errori iid

Supponiamo di voler prendere il tasso di criminalità $y$, nella località i, dove le caratteristiche specifiche della località sono contenute in $x$. Un approccio consiste nell'assumere che il tasso di criminalità osservato $y_{\mathrm{i}}$, è una combinazione di un segnale, $f$ , corrotto dal rumore,$e_{\mathrm{i}}$, tale che:

$$y_i = f(\mathbf{x}_i) + e_i. \tag{2.2}$$

Il rumore statistico è generalmente costituito da errori e residui:

- Gli errori fanno riferimento a errori di misurazione ed errori di campionamento evidenziando le differenze tra i valori osservati che abbiamo effettivamente misurato e i loro "valori reali".

- Il residuo dei dati osservati è la differenza tra il valore osservato e il valore previsto. Nell'analisi di regressione , è la distanza tra il punto dati osservato e la linea di regressione.[10]

Si presume che il rumore sia conforme a una certa distribuzione, che in questo caso assumiamo essere gaussiana, quindi $e \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2)$ , dove $i.i.d$ si riferisce a campioni distribuiti in modo indipendente e identico. il segnale $f$ nella regressione lineare si assume che sia lineare, quindi $f(x_\text{i})=x_\text{i}\beta$, dove $x_\text{i}=$ ( 1 ,$x_{\text{i}1}$, ... ,$x_{\text{i}P}$),e $\beta=$ ($\beta_0$,$\beta_1$, ... ,$\beta_P$), dove $x_\text{ik}$ è l' i- esimo valore osservato della caratteristica $k$ , e $P$ è il numero delle caratteristiche.
I parametri che specificano completamente questo modello sono dati da $\theta = \beta$ ,$\sigma_\text{e}$ , i dati sono contrassegnati da $D=$ ($X$, $y$), dove $X=$ ($X_1$', ..., $X_\text{n}$')', $y$=($y_1$, ...,$y_\text{n}$)' e $n$ è il numero di località con tassi di criminalità registrati e le rispettive caratteristiche specifiche della località. Se la $\text{e}_\text{i}$ è conforme approssimativamente alle nostre ipotesi, vale a dire $e \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_e^2)$ , quindi la probabilità $p(\mathbf{y}|X, \theta)$ è la distribuzione gaussiana multivariata. Allo stesso modo, la distribuzione predittiva del tasso di criminalità inosservato in una particolare località $y^\star$ , con caratteristiche $\mathbf{x}^\star$ è dato da $p(y^\star|\mathbf{x}^\star, \mathcal{D})$ed è uguale a

$$p(y^\star|\mathbf{x}^\star, \mathcal{D}) = \int_{R^{|\theta|}} p(y^\star|\mathbf{x}^\star, \mathcal{D}, \theta)p(\theta|\mathcal{D})\text{d}\theta. \tag{2.3}$$

### 2.3.2 Regressione lineare bayesiana con dipendenza spaziale

Quando si tratta di dati spaziali, come descritto in questo documento, non è realistico presumere che il tasso di criminalità dipenda solo da quelle caratteristiche specifiche del luogo che vengono misurate, perché è probabile che le posizioni vicine tra loro nello spazio siano correlate. Quindi allentiamo l'ipotesi che gli errori in $e$ =($e_1$, ... ,$e_\text{n}$) siano indipendenti e assumiamo la seguente funzione:

$$y_i = f(\mathbf{x}_i) + h(\mathbf{u}_i) + \epsilon_i. \tag{2.4}$$

dove i termini di errore sono forniti da $e$i $\sim^{\text{iid}} N(0$ ,$\sigma^2{}_e)$ , $u_\text{i}$ è il vettore delle coordinate spaziali della posizione $i$ e $h(u)$ è una funzione non parametrica di $u$. Inoltre,assumiamo che la relazione tra le caratteristiche specifiche del luogo di criminalità in $X$, è indipendente dalla relazione tra il tasso di criminalità e le coordinate spaziali in $u$.

### 2.3.3 Regressione sui tassi di criminalità

In questa sezione , applichiamo la metodologia definita nella sua base teorica sopra per eseguire un'analisi su particolari tipi di reati, nello specifico riportiamo : aggressioni legate alla violenza domestica, furti con scasso e furti di veicoli a motore nel South Wales in Australia considerando un arco temporale che copre il periodo 2009-2013.
Le informazioni spaziali fornite su ciascun episodio di criminalità sono un identificatore di area geografica, denominato Statistical Area Level 2 (SA2). Le SA2 sono aree geografiche che presentano una distribuzione della popolazione relativamente omogenea. A questo livello di granularità, è possibile visualizzare modelli interessanti preservando la privacy degli individui. Le variabili esplicative sono selezionate dai dati demografici a livello SA2, estratte dai dati del censimento per gli anni 2001, 2006 e 2011. Queste informazioni sono disponibili pubblicamente presso l'Australian Bureau of Statistics (ABS). Le dodici caratteristiche demografiche e le loro statistiche riassuntive sono presentate nella tabella sottostante.

| Descrizione variabile | Min | Max | Significare | SD |
|---|---|---|---|---|
| Numero di maschi separati (per 100 maschi totali) | 1 | 4 | 2 | 1 |
| Percentuale di disoccupazione | 2 | 14.3 | 6 | 2.2 |
| Densità di popolazione (per km $^2$ ) | 0,02 | 14301 | 1466 | 2027 |
| Reddito settimanale totale medio della famiglia | 618 | 2610 | 1264 | 449 |
| Rimborso mensile mediano del mutuo | 300 | 3289 | 1898 | 535 |
| Affitto medio settimanale | 50 | 690 | 292 | 112 |
| Percentuale di persone senza religione | 4 | 42 | 18 | 7 |
| Età media | 22 | 59 | 39.16 | 5.31 |
| Percentuale di immigrati | 2 | 63 | 21 | 15 |
| Percentuale di persone che parlano solo inglese | 13 | 97 | 78 | 21 |
| Percentuale con solo istruzione professionale (certificato di livello 1 o 2 per tutti i livelli) | 3 | 13 | 7 | 1 |
| Numero di famiglie con genitore solo (per 100 abitanti totali) | 1 | 10 | 4 | 1 |

Figura 2.1: Caratteristiche demografiche e statistiche riassuntive nelle aree statistiche basate sui dati del censimento ABS 2011

I conteggi dei crimini per tipi di crimini specifici vengono aggregati nello spazio in SA2 e i tassi di criminalità vengono calcolati utilizzando le informazioni sulla popolazione corrispondenti (per mille persone). Riportiamo i dati osservati rispetto al tasso di criminalità previsto mostrando la diagnostica per ciascun modello e l'adattamento rispettivamente per aggressioni correlate a violenza domestica , furti con scasso e furti di veicoli a motore.

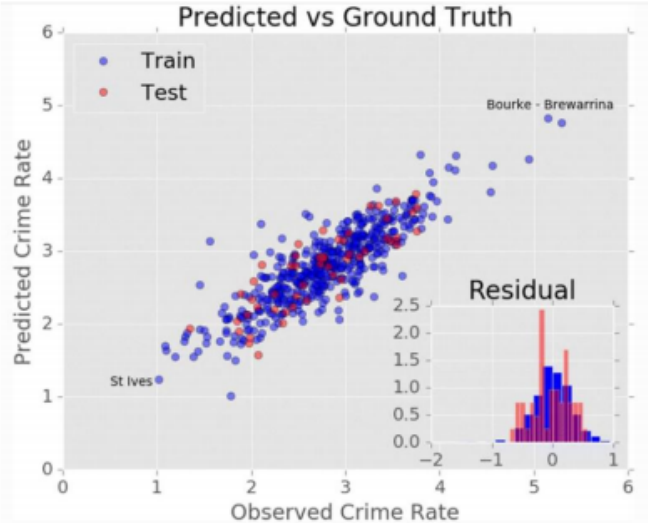



Figura 2.2: Rappresentazione degli Assalti correlati a Violenza Domestica

Il grafico sopra riportato rappresenta i dati correlati alle violenze domestiche: nell'asse delle ordinate viene indicato il valore atteso del tasso di criminalità previsto per gli attacchi correlati a violenza domestica in funzione del tasso di criminalità osservato, nell'asse orizzontale, per tutte le regioni del NSW in SA2 tra gli anni 2009 e 2013 .I punti blu e rossi mostrano le stime rispettivamente agli addestramenti e i test di prova dei dati.

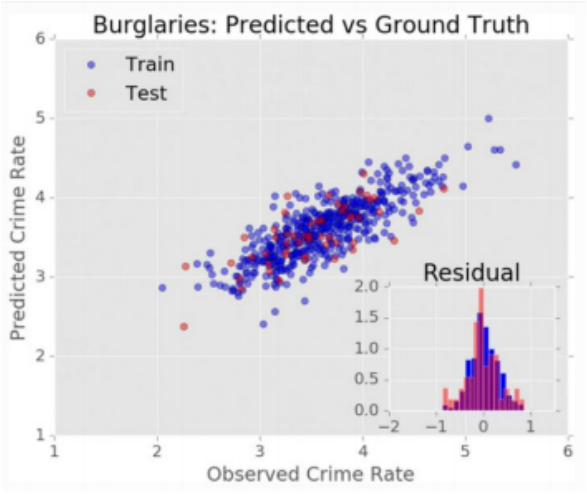



Figura 2.3: Rappresentazione grafica dei crimini relatici al furto con scasso

Il grafico rappresenta i dati correlati ai furti con scasso : nell 'asse verticale si riporta il valore atteso del tasso di criminalità previsto per i furti con scasso in funzione del tasso di criminalità osservato, riportato nell'asse orizzontale, per tutte le regioni del NSW in SA2 tra gli anni 2009 e 2013. I punti blu e rossi mostrano le stime degli addestramenti e i test di prova delle posizioni.

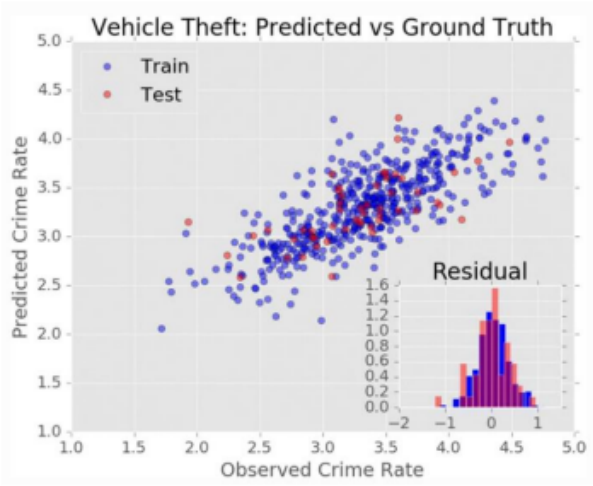



Figura 2.4: Rappresentazione grafica dei crimini relativi a furto di veicoli a motore

I dati rappresentati in questo ultimo grafico fanno riferimento a furti di veicoli a motore: nell'asse verticale viene riportato il valore atteso del tasso di criminalità previsto per il furto di veicoli a motore in funzione del tasso di criminalità osservato, nell'asse orizzontale, per tutte le regioni del NSW in SA2 tra gli anni 2009 e 2013. I punti blu e rossi mostrano le stime rispettivamente per le posizioni degli addestramenti e i test di prova dei dati.
Ogni punto nei grafici rappresenta una regione. Graficamente si può concludere che esiste una correlazione tra i tassi di criminalità previsti e osservati per tutti i tipi di crimini selezionati. Queste cifre mostrano anche che i residui sono indipendenti e seguono una distribuzione gaussiana, come ipotizzato dal modello la cui descrizione è stata definita precedentemente nella **Regressione lineare bayesiana con dipendenza spaziale**. L'indipendenza della relazione tra il tasso di criminalità e le coordinate spaziali è definito in :

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)\mathbf{u}_i h(\mathbf{u})\mathbf{xu} \tag{2.5}$$

Una stima quantitativa dell'errore è l'errore quadratico medio (RMSE) ed è calcolata sul tasso di criminalità del registro e sui conteggi dei crimini in base

alle seguenti espressioni:

$$\text{RMSE}_{\text{rate}} = \sqrt{\frac{\sum_{j=1}^{N} (y_j - y_j^\star)^2}{N}} \tag{2.6}$$

$$\text{RMSE}_{\text{counts}} = \sqrt{\frac{\sum_{j=1}^{N} P_j^2 (e^{y_j} - e^{y_j^\star})^2}{N}} \tag{2.7}$$

dove $y_{\text{i}}$ è il tasso di criminalità registrato nella posizione $j$, $\text{y}_{\text{j}}^\star$ è la stima media a posteriori del tasso di criminalità registrato nella posizione $j$, $N$ è il numero di posizioni nel set di test / addestramento e $P_{\text{j}}$ è la popolazione totale nella posizione $j$ (su una popolazione di 1000 persone). Calcoliamo l'errore nel numero di crimini per contestualizzare l'entità degli episodi di criminalità nella discussione.

| | Tasso di criminalità del registro RMSE | | Conteggio dei crimini RMSE | | % entro CI | | Correlazione Pred / Oss | |
|---|---|---|---|---|---|---|---|---|
| | Treno | Test | Treno | Test | Treno | Test | Treno | Test |
| DV | 0.314 | 0.318 | 92.1 | 88.4 | 94 | 98 | 0.86 | 0.85 |
| Furti | 0.301 | 0.350 | 191.5 | 208.0 | 90 | 90 | 0.84 | 0.72 |
| MVT | 0.345 | 0.372 | 191.7 | 191.3 | 89 | 88 | 0.78 | 0.66 |

Figura 2.5: Statistiche degli errori per i modelli per diversi tipi di criminalità

La tabella riporta il RSME sul tasso di criminalità , i conteggi dei crimini, la percentuale delle posizioni di avvenimento del crimine all'interno dell'intervallo credibile del 95% e la correlazione tra i valori previsti e osservati. Gli Intervalli credibili sono calcolati in base alla densità predittiva a posteriori. La percentuale delle previsioni all'interno dell'IC rappresenta una misura di accuratezza rispetto alla quantificazione dell'incertezza. Se le ipotesi del nostro modello sono corrette, ci aspetteremmo che il 95% del tasso di criminalità effettivo nei luoghi di test si trovi all'interno della distribuzione predittiva posteriore del 95%. Dal grafico gli indicatori principali ci permettono di dedurre che esiste un'elevata correlazione lineare tra i valori previsti e osservati per ogni crimine , il che indica che il modello sta trovando associazioni all'interno delle covariate e utilizzando tali informazioni per spiegare il tasso di criminalità. Le prime due colonne numeriche della tabella mostrano una grandezza simile all'errore per diversi tipi di crimini. Notiamo che gli indicatori di prestazione suggeriscono che le previsioni per le aggressioni correlate a violenze domestiche sono più accurate di quelle per furti con scasso e furto di veicoli a motore, ovvero RMSE inferiore, maggiore percentuale all'interno dell'intervallo credibile e maggiore correlazione tra previsioni e osservazioni.

### 2.3.4 Inferenza sui dati demografici

Per capire come i fattori demografici selezionati contribuiscono a un tipo di crimine specifico, dobbiamo guardare alla distribuzione a posteriori sui parametri di regressione. Come descritto nella sezione " Regressione lineare bayesiana con dipendenza spaziale ", i valori di $\beta$ possono essere interpretati come un aumento percentuale del tasso di criminalità che risulterebbe da un aumento percentuale o da ogni aumento percentuale della variabile indipendente. In questo caso particolare, ciascuno $\beta$i rappresenta come un aumento/diminuzione unitario della percentuale della variabile demografica $i$ è correlato all'aumento/diminuzione percentuale del tasso di criminalità.
L'inferenza di aggressioni relative alle violenze domestiche è rappresentata dalla tabella che segue, la quale mostra la media a posteriori del coefficiente di regressione, la violenza domestica e gli intervalli credibili al 95% . Le righe vengono ordinate in base alla media posteriore discendente

| Parametro | Media posteriore | SD posteriore | Intervallo credibile al 95% |
|---|---|---|---|
| Percentuale di maschi separati | 1.40 | 0.10 | [1.20, 1.59] |
| Densità demografica | 0.96 | 0.13 | [0,68, 1,20] |
| Percentuale di disoccupazione | 0.70 | 0.14 | [0,42, 0,98] |
| Percentuale di persone che parlano solo inglese | 0.65 | 0.14 | [0,37, 0,94] |
| Percentuale di persone Cert 1 o 2 | 0,53 | 0.11 | [0,32, 0,75] |
| Numero di famiglie con un genitore unico | 0.23 | 0.14 | [- 0,04, 0,53] |
| Reddito familiare totale mediano | - 0,09 | 0.18 | [- 0,45, 0,28] |
| Percentuale di immigrati | - 0,19 | 0.15 | [- 0,45, 0,14] |
| Percentuale di persone senza religione | - 0.49 | 0,09 | [- 0,68, - 0,31] |
| Rimborso mensile mediano del mutuo | - 0.66 | 0.28 | [- 1,22, - 0,10] |
| Affitto mediano | - 0.91 | 0.24 | [- 1,40, - 0,43] |
| Età media | - 1.05 | 0.13 | [- 1,29, - 0,77] |

Figura 2.6: Dati relativi all'inferenza di aggressioni correlate a violenza domestica

L'inferenza relativa ai furti con scasso viene rappresentato nel grafico sottostante:

| Parametro | Media posteriore | SD posteriore | Intervallo credibile al 95% |
|---|---|---|---|
| Percentuale di disoccupazione | 1.19 | 0.15 | [0,89, 1,49] |
| Percentuale di persone Cert 1 o 2 | 0.98 | 0.12 | [0,73, 1,21] |
| Percentuale di maschi separati | 0.83 | 0.11 | [0,62, 1,05] |
| Densità demografica | 0.72 | 0.14 | [0.46, 1.01] |
| Percentuale di persone che parlano solo inglese | 0.45 | 0.17 | [0,13, 0,80] |
| Reddito familiare totale mediano | 0.24 | 0.20 | [- 0,17, 0,62] |
| Percentuale di persone senza religione | 0,07 | 0.11 | [- 0,14, 0,28] |
| Rimborso mensile mediano del mutuo | 0,05 | 0.29 | [- 0,54, 0,64] |
| Numero di famiglie con un genitore unico | - 0.23 | 0.16 | [- 0,55, 0,06] |
| Età media | - 0,32 | 0.14 | [- 0,63, - 0,05] |
| Affitto mediano | - 0.63 | 0.25 | [- 1,12, - 0,12] |
| Percentuale di immigrati | - 0,82 | 0.17 | [- 1,15, - 0,47] |

Figura 2.7: Dati relativi all'inferenza sui furti con scasso

Per i furti di veicoli a motore invece abbiamo :

| Parametro | Media posteriore | SD posteriore | Intervallo credibile al 95% |
|---|---|---|---|
| Densità demografica | 1.81 | 0.15 | [1.52, 2.10] |
| Percentuale di maschi separati | 1.29 | 0.11 | [1.06, 1.51] |
| Percentuale di persone Cert 1 o 2 | 0.78 | 0.13 | [0,52, 1,02] |
| Affitto mediano | 0.70 | 0.27 | [0,14, 1,22] |
| Percentuale di disoccupazione | 0.39 | 0.16 | [0,09, 0,72] |
| Percentuale di persone che parlano solo inglese | 0.21 | 0.17 | [- 0,10, 0,56] |
| Percentuale di persone senza religione | 0.19 | 0.11 | [- 0,01, 0,42] |
| Rimborso mensile mediano del mutuo | 0.16 | 0.32 | [- 0,50, 0,79] |
| Numero di famiglie con un genitore unico | - 0.42 | 0.16 | [- 0,73, - 0,10] |
| Reddito familiare totale mediano | - 0.47 | 0.21 | [- 0,87, - 0,05] |
| Età media | - 0,98 | 0.15 | [- 1,27, - 0,69] |
| Percentuale di immigrati | - 1.20 | 0.18 | [- 1,54, - 0,84] |

Figura 2.8: Dati relativi all'inferenzasul furto di veicoli a motore

Una rappresentazione grafica di questi valori è indicata come boxplot:

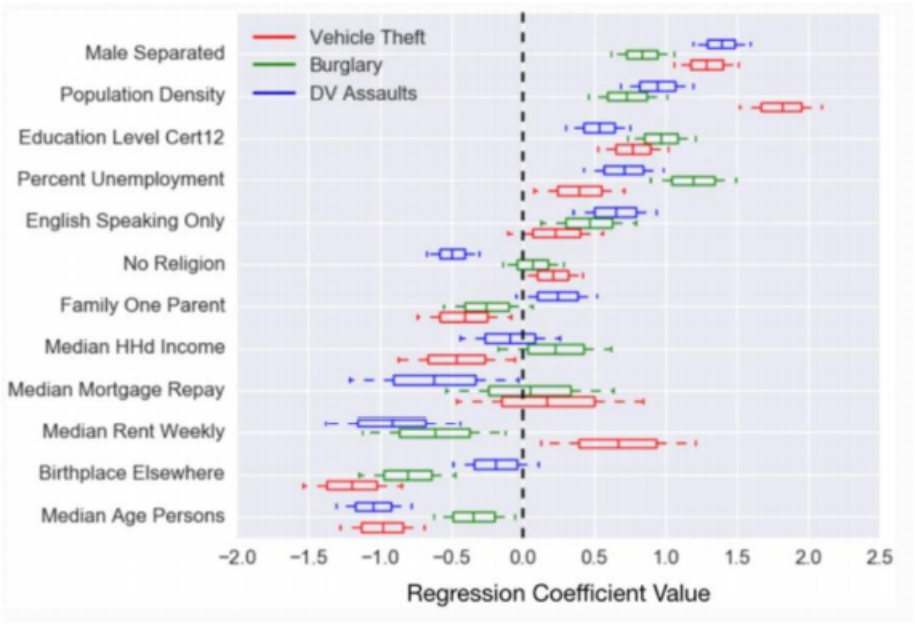



Figura 2.9: Box plot dei coefficienti di regressione demografica per tre diversi tipi di reato: violenza domestica, furti con scasso e furti di veicoli a motore

Abbiamo analizzato i Box-plot dei campioni sulla base del coefficiente di regressione tratte dalle loro distribuzioni a posteriori per i tre diversi tipi di reato. Il riquadro è definito da un intervallo compreso tra -1<x<+1 relativo alla deviazione standard per una distribuzione gaussiana e la mediana viene rappresentata come linea verticale all'interno della scatola grafica.[33] La linea orizzontale tratteggiata indica gli intervalli di confidenza del 96%, cioè 2 e 98 percentili.
Esploriamo ulteriormente la natura, variabile nel tempo, della dipendenza tra il tasso di criminalità,le caratteristiche demografiche e l'ubicazione spaziale conducendo tre studi trasversali che aggregano la criminalità su tre periodi. Ogni periodo si estende su 5 anni ed è incentrato sui dati del censimento 2001, 2006 e 2011. I risultati sono riportati nel Boxplot rappresentato di seguito:

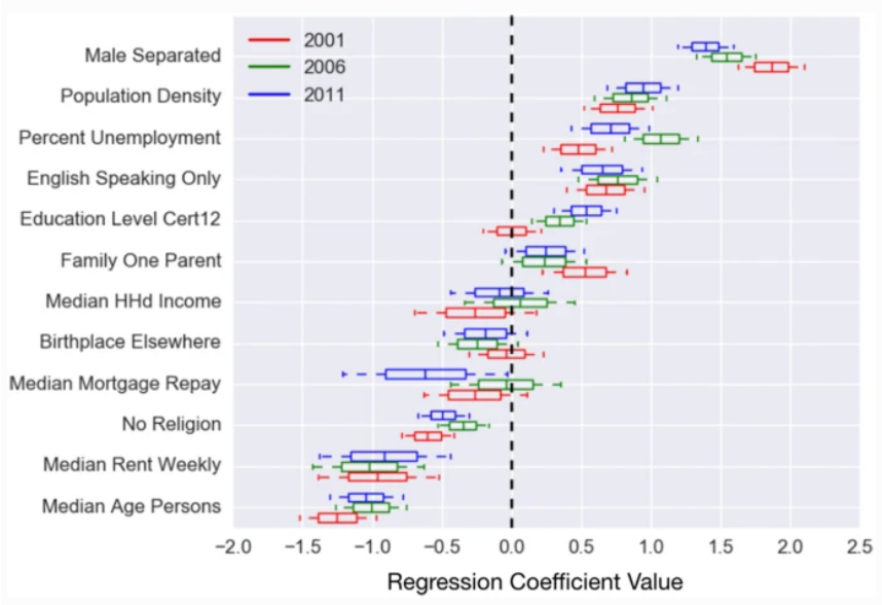



Figura 2.10: Box plot dei coefficienti di regressione in più periodi di tempo per fattori demografici e attacchi correlati a violenza domestica

Un'altro risultato interessante viene mostrato, dove i tassi di criminalità effettivi , mostrati nella sezione superiore, sono confrontati con quelli previsti. La capacità del modello di catturare le dipendenze spaziali e fornire stime accurate dei veri livelli di criminalità, basate solo su informazioni demografiche e spaziali, è sorprendente.

| | Tasso di criminalità registrato RMSE | | Conteggio dei crimini RMSE | |
|---|---|---|---|---|
| | Treno | Test | Treno | Test |
| Spatial-dem. reg. | 0.314 | 0.318 | 92.1 | 88.4 |
| Ingenuo | 0.544 | 0.552 | 162.4 | 158.1 |

Figura 2.11: Dati a Confronto dell'RMSE

Con i risulati ottenuti in questa sezione, abbiamo collegato alcuni dei risultati alla teoria criminologica esistente.
Le tabelle relative alle singole inferenze mostrano le statistiche di riepilogo per ogni variabile e per ogni tipo di crimine. Una media posteriore positiva è collegata ad un aumento del tasso di criminalità per il particolare tipo di crimine. Tuttavia, è necessario porre attenzione sull'intervallo credibile. Se l'intervallo credibile contiene zero, allora c'è una probabilità non trascurabile che questo parametro sia zero, il che implica che non c'è relazione tra quella

specifica variabile demografica e il tasso di criminalità. Un intervallo credibile più breve rappresenta anche una minore incertezza intorno al valore del coefficiente di regressione specifico, aumentando l'affidabilità della relazione tra quella specifica covariata e il crimine.
Abbiamo raggruppato le variabili in tre categorie.

1. La prima categoria è costituita dalle covariate che sono positivamente correlate a un aumento della criminalità: percentuale di maschi separati, densità di popolazione, istruzione e disoccupazione. Questo è simile ai risultati riportati negli studi di Nivette [31], che ha scoperto che la proporzione di maschi e densità di popolazione era correlata positivamente alla criminalità.

2. La seconda categoria è composta da quelle covariate che hanno una relazione negativa con tutti e tre i tipi di criminalità, essendo Age e Immigrants.

3. La terza categoria racchiude le covariate per le quali l'impatto varia a seconda del tipo di criminalità: religione, famiglia monoparentale, reddito, mutuo e affitto

### 2.3.5 Risultati

L'osservazione principale è che alcuni fattori demografici come affitto, mutuo e religione hanno impatti diversi su determinati tipi di criminalità. Di particolare rilievo è il fatto che le aree con un'elevata percentuale di persone che si dichiarano religiose hanno meno probabilità di subire furti o rapine, ma più probabilità di essere vittime di violenza domestica. Allo stesso modo, le aree con pagamenti elevati di mutui / affitti hanno meno probabilità di subire violenza domestica, ma è più probabile che subiscano furti o rapine. Tuttavia, vivere in un'area con un'elevata popolazione di immigrati è associato a tassi di criminalità inferiori in tutti e tre i tipi di criminalità; minore furto, minore furto con scasso e minore violenza domestica. Una delle questioni aperte, oggetto di ricerche future, è se l'immigrazione stessa riduca il numero effettivo di crimini commessi in queste aree. Si può vedere che gli stessi fattori demografici contribuiscono in modo simile alle aggressioni correlate alla violenza domestica in tutti gli anni con la maggiore variazione nel tempo nelle aree dell'istruzione e della disoccupazione. Questi risultati suggeriscono che il tempo di modellazione esplicito potrebbe mostrare ulteriori variazioni se i dati fossero disponibili con una risoluzione temporale più fine.

### 2.3.6 Errori di previsioni

L'incertezza di previsione per i tipi di criminalità non correlati a violenza domestica è dovuta al fatto che crimini come furti di veicoli a motore e furti con scasso non sono necessariamente commessi da criminali che vivono nella stessa area, mentre la maggior parte delle aggressioni di violenza domestica si verificano nella residenza delle persone di interesse.Poiché il nostro attuale modello demografico riflette solo i dati degli individui che vivono in quella particolare area, la popolazione transitoria non viene attualmente presa in considerazione e quindi porta a maggiori incertezze nelle nostre previsioni. Pertanto, l'inclusione di variabili che stimano le caratteristiche legate a aspetti ambientali e di spostamenti migliorerà la qualità delle previsioni di reati commessi lontano dall'indirizzo di residenza di un individuo.Gli errori di previsione e gli schemi catturati dal modello, rappresentati dalla componente di regressione parametrica, dipenderanno fortemente dal sottoinsieme selezionato di variabili esplicative.

### 2.3.7 Conclusione dell'analisi

E' stato presentato un modello completamente probabilistico in grado di prevedere con precisione i tassi di criminalità e fornire incertezze che circondano tali previsioni, fornendo allo stesso tempo inferenza sui possibili fattori specifici della posizione associati al crimine. I risultati convalidano i principi criminologici teorici esistenti riguardanti i fattori associati ad attività criminale alta e bassa, compresi i limiti sul grado della relazione. I risultati mostrano anche come questo modello possa essere utilizzato per comprendere diversi tipi di criminalità e quali siano i limiti a seconda delle caratteristiche specifiche del luogo utilizzate per descrivere quel particolare fenomeno.[28] Le variabili utilizzate a supporto dell'analisi sono state: il tasso di criminalità come variabile dipendente e il valore delle caratteristiche specifiche della posizione diverse da zero come variabili indipendenti dal momento che la relazione tra questi due insiemi di variabili è approssimativamente lineare e i residui approssimativamente distribuiti normalmente. Notare che i termini di errore spazialmente indipendenti sono stati sostituiti da $e_{\mathrm{i}}$. Abbiamo aggregato i dati sulla criminalità in 5 anni, intorno ai dati del censimento del 2011, per ottenere inferenze statistiche significative per il processo decisionale a lungo termine. Gli intervalli credibili differiscono dagli intervalli di confidenza in quanto gli intervalli credibili sono associati a distribuzioni posteriori, mentre gli intervalli di confidenza spesso presumono che la distribuzione delle stime di campionamento sia gaussiana.

# Capitolo 3

# Implementazione e Analisi a confronto

In questa sezione verranno presentate delle analisi svolte personalmente su particolari datasets relativi all' ambito criminologico. Lo strumento a supporto di questa analisi è il software R Studio.[25] La ricerca dei dati è stata orientata al rilevamento di statistiche sul territorio italiano e americano. La fonte da cui sono stati recuperati i primi dati è Open Data, inoltre i datasets presentati sono stati filtrati per riportare risultati relativi alla regione Emilia Romagna, in particolare si propone:

- Rischio di criminalità
- Tasso di omicidi per la regione
- Furti denunciati ogni mille abitanti

L'analisi che verrà proposta nei 3 casi segue l'analisi di Regressione Lineare e polinomiale, verranno proposti grafici e codice in R a supporto della descrizione. Per quanto riguarda la seconda parte dei dati , si tratta di statistiche relative ai differenti tipologie di crimine avvenuti a Baltimore nell'anno 2017. La fonte di ricerca è stata data.world. Il dataset è stato ripulito, ridimensionato e converito a livello dei valori delle variabili, ovvero sono state sostituite significativamente con valori numerici.[10]

## 3.1 Preparazione dei Dati

La preparazione dei dati è stata differente per i datasets che coinvolgono la regressione univariata e quella multivariata. Nel primo caso i datasets proposti sono stati recuperati presso Open Data, focalizzando la ricerca su

crimini avvenuti sul territorio italiano. I dati raccolti presentano tre Features di cui sono le medesime ovver Anno, Regione, la terza caratteristica indica il valore della Tipologia di dato che si sta considerando. La pulizia dei dati fatta, è composta delle seguenti fasi:

- scarimento del dataset
- importazione del datasets in R Studio
- eliminazione delle righe non utili

Riportiamo l'esempio di pulizia dei dati per il primo dataset che consideriamo. Il filtraggio dei dati è avvenuto manualmente, utilizzando il comando in R per l'eliminazione di più righe ,di cui riporto la sintassi generica come segue:

```
>newdata <- mydata[-c(1, 2, 3),]
```

Riportiamo l'esempio di pulizia del primo dataset di cui si propone l'analisi, data la medesima procedura applicata per le successive.

| X | X2..Rischio.di.criminalità | X.1 | X.2 | X.3 | X.4 | X.5 | X.6 | X.7 | X.8 | X.9 | X.10 | X.11 | X.12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Famiglie che avvertono molto o abbastanza disagio ai rischi... | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| 2 | Regione | 2002.00 | 2003.00 | 2004.00 | 2005.00 | 2006.00 | 2007.00 | 2008.00 | 2009.00 | 2010.00 | 2011.00 | 2012.00 | 2013.00 |
| 3 | Piemonte | 33,67 | 32.69 | 27.68 | 27.68 | 30.86 | 33.10 | 37.73 | 37.47 | 30.27 | 26.66 | 26.70 | 27.40 | 30.30 |
| 4 | Valle d'Aosta | 13,21 | 11.32 | 9.43 | 9.43 | 12.73 | 16.36 | 16.07 | 18.97 | 12.50 | 15.79 | 10.80 | 13.10 | 12.50 |
| 5 | Lombardia | 34,24 | 32.36 | 31.55 | 31.55 | 31.26 | 32.41 | 41.39 | 42.37 | 35.15 | 33.42 | 32.20 | 29.00 | 36.90 |
| 6 | Trentino-Alto Adige | 15,54 | 15.15 | 12.50 | 12.50 | 11.43 | 9.49 | 10.91 | 11.08 | 9.18 | 9.03 | 8.30 | 8.60 | 9.60 |
| 7 | Veneto | 32,84 | 32.20 | 27.61 | 27.61 | 37.65 | 33.41 | 29.18 | 39.64 | 29.33 | 23.99 | 25.20 | 29.40 | 31.70 |
| 8 | Friuli-Venezia Giulia | 18,13 | 19.46 | 16.27 | 16.27 | 19.34 | 17.47 | 17.67 | 21.41 | 15.53 | 14.26 | 13.80 | 13.60 | 17.20 |
| 9 | Liguria | 29,44 | 29.32 | 24.19 | 24.19 | 24.93 | 24.63 | 26.06 | 28.19 | 26.31 | 22.02 | 19.90 | 22.00 | 23.90 |
| 10 | Emilia-Romagna | 28,18 | 25.97 | 24.30 | 24.30 | 24.25 | 28.62 | 31.02 | 35.94 | 26.48 | 25.23 | 25.60 | 25.40 | 32.40 |
| 11 | Toscana | 25,4 | 26.05 | 17.71 | 17.71 | 25.52 | 28.48 | 33.76 | 31.09 | 25.42 | 22.35 | 18.90 | 24.80 | 26.40 |
| 12 | Umbria | 23,75 | 24.92 | 23.44 | 23.44 | 35.03 | 31.36 | 27.83 | 38.94 | 28.09 | 21.88 | 21.90 | 32.70 | 36.80 |
| 13 | Marche | 13,2 | 11.76 | 10.35 | 10.35 | 13.89 | 22.35 | 25.20 | 26.49 | 15.95 | 15.50 | 17.60 | 19.90 | 27.50 |
| 14 | Lazio | 40,66 | 39.31 | 38.78 | 38.78 | 31.60 | 40.66 | 46.28 | 47.09 | 39.40 | 37.75 | 34.80 | 32.90 | 40.80 |
| 15 | Abruzzo | 10,99 | 11.37 | 15.58 | 15.58 | 13.06 | 17.07 | 23.85 | 28.66 | 22.50 | 22.10 | 17.60 | 16.90 | 25.40 |
| 16 | Molise | 8,26 | 10.48 | 13.01 | 13.01 | 11.76 | 6.56 | 12.00 | 16.47 | 11.38 | 10.24 | 13.80 | 13.80 | 9.40 |
| 17 | Campania | 50,13 | 44.71 | 48.30 | 48.30 | 52.60 | 51.35 | 53.91 | 53.64 | 48.89 | 40.16 | 45.40 | 38.70 | 36.10 |
| 18 | Puglia | 31,91 | 27.93 | 26.44 | 26.44 | 29.73 | 34.23 | 35.53 | 36.49 | 26.03 | 25.47 | 24.40 | 25.70 | 33.30 |
| 19 | Basilicata | 11,63 | 8.49 | 7.14 | 7.14 | 13.76 | 11.21 | 9.72 | 11.76 | 7.02 | 5.15 | 8.20 | 14.60 | 14.10 |
| 20 | Calabria | 17,81 | 13.25 | 12.81 | 12.81 | 15.70 | 26.56 | 22.64 | 30.45 | 20.21 | 22.82 | 14.70 | 17.80 | 21.60 |
| 21 | Sicilia | 25,7 | 25.86 | 23.41 | 23.41 | 22.95 | 24.95 | 27.70 | 27.48 | 25.16 | 24.09 | 22.70 | 20.70 | 27.30 |
| 22 | Sardegna | 17,53 | 14.33 | 17.01 | 17.01 | 17.25 | 15.49 | 18.65 | 20.00 | 13.63 | 12.16 | 16.20 | 14.30 | 13.30 |
| 23 | ITALIA | 30,62 | 29.15 | 27.40 | 27.40 | 29.16 | 31.32 | 34.60 | 36.78 | 29.70 | 27.14 | 26.60 | 26.35 | 30.95 |
| 24 | Nord | 31,1 | 29.85 | 27.07 | 27.07 | 29.39 | 29.96 | 33.74 | 36.92 | 29.43 | 26.81 | 26.47 | 26.25 | 31.38 |
| 25 | Nord-ovest | 33,33 | 31.93 | 29.41 | 29.41 | 30.27 | 31.59 | 36.36 | 39.18 | 32.59 | 30.13 | 29.16 | 27.65 | 33.42 |
| 26 | Nord-est | 27,8 | 26.82 | 23.64 | 23.64 | 28.09 | 27.39 | 27.07 | 33.67 | 24.91 | 22.07 | 22.39 | 24.23 | 28.43 |
| 27 | Centro | 31,2 | 30.46 | 27.37 | 27.37 | 27.71 | 33.69 | 36.04 | 36.88 | 31.17 | 28.90 | 26.66 | 28.69 | 34.35 |
| 28 | Centro-Nord | 31,13 | 30.03 | 27.16 | 27.16 | 28.91 | 31.06 | 34.99 | 37.50 | 29.95 | 27.43 | 26.52 | 26.98 | 32.27 |
| 29 | Mezzogiorno | 30,19 | 27.34 | 27.90 | 27.90 | 29.70 | 31.87 | 33.79 | 35.25 | 29.16 | 26.54 | 26.75 | 25.01 | 28.15 |
| 30 | Fonte: elaborazione su dati Istat | | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |

Showing 1 to 30 of 30 entries, 14 total columns

Figura 3.1: Dataset completo

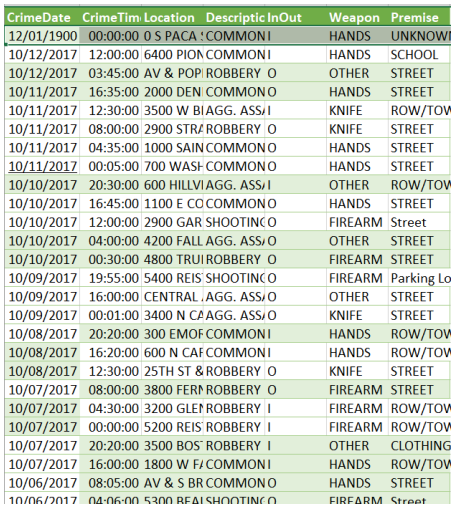
| | Regione | Anno | Rischio |
|---|---|---|---|
| 1 | Emilia-Romagna | 2001 | 28.18 |
| 2 | Emilia-Romagna | 2002 | 25.97 |
| 3 | Emilia-Romagna | 2003 | 24.30 |
| 4 | Emilia-Romagna | 2004 | 24.30 |
| 5 | Emilia-Romagna | 2005 | 24.25 |
| 6 | Emilia-Romagna | 2006 | 28.62 |
| 7 | Emilia-Romagna | 2007 | 31.04 |
| 8 | Emilia-Romagna | 2008 | 35.94 |
| 9 | Emilia-Romagna | 2009 | 26.48 |
| 10 | Emilia-Romagna | 2010 | 25.23 |
| 11 | Emilia-Romagna | 2011 | 25.60 |
| 12 | Emilia-Romagna | 2012 | 25.40 |
| 13 | Emilia-Romagna | 2013 | 32.40 |

Figura 3.2: Dataset pulito

A tal proposito , sono state eliminate le righe relative alle regione diverse dall'Emilia Romagna. Si tratta di una scelta che vuole mettere in evidenza i dati variabili per la specifica regione. Il metodo di analisi proposto segue la regressione lineare e in seguito polinomiale , che rappresentano strumenti appresi durante il corso di studi.

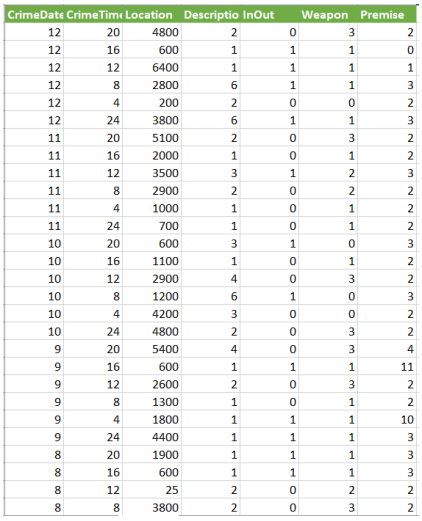
Per quanto riguarda l'analisi di regressione multivariata, si propone un dataset con più features relativo ai crimini avvenuti a Baltimore. La ricerca di

un dataset che rispecchiasse le caratteristiche cercate è stata difficile, tuttavia la ricerca è avvenuta presso la banca dati data.world . Il dataset inizialmente presentava più di 28000 mila righe e più colonne. É stato ridimensionato in questo modo: è stato scelto un giorno il 10 di ogni mese dell'anno 2017, e sono state considerate 6 indicazioni dell'orario del crimine, con un'alternanza di 4 ore. Sono state mantenute le informazioni relative a: descrizione, luogo del crimine, strumenti di crimine, il codice postale della locazione e infine 'indicazione se il crimine sia stato svolto in un ambiente interno o esterno. Tuttavia per permettere l'analisi che lavora con valori numerici ho costruito una leggenda e associato a ciascun valore un numero. Riportiamo prima il dataset pulito e ridimensionato e affianco la traduzione numerica dei dati.



| CrimeDate | CrimeTim | Location | Descriptio | InOut | Weapon | Premise |
|---|---|---|---|---|---|---|
| 12/01/1900 | 00:00:00 | 0 S PACA S | COMMON | I | HANDS | UNKNOWI |
| 10/12/2017 | 12:00:00 | 6400 PIOI | COMMON | I | HANDS | SCHOOL |
| 10/12/2017 | 03:45:00 | AV & POP | ROBBERY | O | OTHER | STREET |
| 10/11/2017 | 16:35:00 | 2000 DEN | COMMON | O | HANDS | STREET |
| 10/11/2017 | 12:30:00 | 3500 W BI | AGG. ASS | I | KNIFE | ROW/TOW |
| 10/11/2017 | 08:00:00 | 2900 STRA | ROBBERY | O | KNIFE | STREET |
| 10/11/2017 | 04:35:00 | 1000 SAIN | COMMON | O | HANDS | STREET |
| 10/11/2017 | 00:05:00 | 700 WASH | COMMON | O | HANDS | STREET |
| 10/10/2017 | 20:30:00 | 600 HILLVI | AGG. ASS | I | OTHER | ROW/TOW |
| 10/10/2017 | 16:45:00 | 1100 E CO | COMMON | O | HANDS | STREET |
| 10/10/2017 | 12:00:00 | 2900 GAR | SHOOTIN | O | FIREARM | Street |
| 10/10/2017 | 04:00:00 | 4200 FALL | AGG. ASS | O | OTHER | STREET |
| 10/10/2017 | 00:30:00 | 4800 TRUI | ROBBERY | O | FIREARM | STREET |
| 10/09/2017 | 19:55:00 | 5400 REIS | SHOOTIN | O | FIREARM | Parking Lo |
| 10/09/2017 | 16:00:00 | CENTRAL | AGG. ASS | O | OTHER | STREET |
| 10/09/2017 | 00:01:00 | 3400 N CA | AGG. ASS | O | KNIFE | STREET |
| 10/08/2017 | 20:20:00 | 300 EMOF | COMMON | I | HANDS | ROW/TOW |
| 10/08/2017 | 16:20:00 | 600 N CAF | COMMON | I | HANDS | ROW/TOW |
| 10/08/2017 | 12:30:00 | 25TH ST & | ROBBERY | O | KNIFE | STREET |
| 10/07/2017 | 08:00:00 | 3800 FERI | ROBBERY | O | FIREARM | STREET |
| 10/07/2017 | 04:30:00 | 3200 GLEI | ROBBERY | I | FIREARM | ROW/TOW |
| 10/07/2017 | 00:00:00 | 5200 REIS | ROBBERY | I | FIREARM | ROW/TOW |
| 10/07/2017 | 20:20:00 | 3500 BOS | ROBBERY | I | OTHER | CLOTHING |
| 10/07/2017 | 16:00:00 | 1800 W FA | COMMON | I | HANDS | ROW/TOW |
| 10/06/2017 | 08:05:00 | AV & S BR | COMMON | O | HANDS | STREET |
| 10/06/2017 | 04:06:00 | 5300 BEAI | SHOOTIN | O | FIREARM | Street |

Figura 3.3: Dati relativi ai crimini avvenuti a Baltimore il 10 di ogni mese dell'anno 2017



| CrimeDate | CrimeTim | Location | Descriptio | InOut | Weapon | Premise |
|---|---|---|---|---|---|---|
| 12 | 20 | 4800 | 2 | 0 | 3 | 2 |
| 12 | 16 | 600 | 1 | 1 | 1 | 0 |
| 12 | 12 | 6400 | 1 | 1 | 1 | 1 |
| 12 | 8 | 2800 | 6 | 1 | 1 | 3 |
| 12 | 4 | 200 | 2 | 0 | 0 | 2 |
| 12 | 24 | 3800 | 6 | 1 | 1 | 3 |
| 11 | 20 | 5100 | 2 | 0 | 3 | 2 |
| 11 | 16 | 2000 | 1 | 0 | 1 | 2 |
| 11 | 12 | 3500 | 3 | 1 | 2 | 3 |
| 11 | 8 | 2900 | 2 | 0 | 2 | 2 |
| 11 | 4 | 1000 | 1 | 0 | 1 | 2 |
| 11 | 24 | 700 | 1 | 0 | 1 | 2 |
| 10 | 20 | 600 | 3 | 1 | 0 | 3 |
| 10 | 16 | 1100 | 1 | 0 | 1 | 2 |
| 10 | 12 | 2900 | 4 | 0 | 3 | 2 |
| 10 | 8 | 1200 | 6 | 1 | 0 | 3 |
| 10 | 4 | 4200 | 3 | 0 | 0 | 2 |
| 10 | 24 | 4800 | 2 | 0 | 3 | 2 |
| 9 | 20 | 5400 | 4 | 0 | 3 | 4 |
| 9 | 16 | 600 | 1 | 1 | 1 | 11 |
| 9 | 12 | 2600 | 2 | 0 | 3 | 2 |
| 9 | 8 | 1300 | 1 | 0 | 1 | 2 |
| 9 | 4 | 1800 | 1 | 1 | 1 | 10 |
| 9 | 24 | 4400 | 1 | 1 | 1 | 3 |
| 8 | 20 | 1900 | 1 | 1 | 1 | 3 |
| 8 | 16 | 600 | 1 | 1 | 1 | 3 |
| 8 | 12 | 25 | 2 | 0 | 2 | 2 |
| 8 | 8 | 3800 | 2 | 0 | 3 | 2 |

Figura 3.4: Dati convertiti in valori numerici riguardanti i crimini avvenuti a Baltimore

In particolare la conversione è stata consentita ponendo le seguenti associazioni. Innanzitutto dato che si fa riferimento allo stesso giorno in mesi differenti, si tiene solo l'informazione del mese, per l'orario invece si tiene conto dell'unità dell'ora tralasciando minuti e secondi. Per le altre caratteristiche abbiamo che:

| DESCRIZIONE: | |
|---|---|
| Common Assault | 1 |
| Robbery | 2 |
| Aggression | 3 |
| Shooting | 4 |
| Homicide | 5 |
| Bulgary | 6 |
| Larceny | 7 |

Tabella 3.1: Leggenda dei valori di descrizione del crimine

| **IN/OUT** : | |
|---|---|
| In | 1 |
| Out | 0 |

Tabella 3.2: Leggenda dei valori di in/out del crimine

| WEAPON: | |
|---|---|
| others | 0 hands |
| 1 | |
| knife | 2 |
| firearm | 3 |

Tabella 3.3: Leggenda dei valori degli strumenti del crimine

| PREMISE: | |
|---|---|
| Unknow | 0 |
| School | 1 |
| street | 2 |
| Row/Town | 3 |
| Parking | 4 |
| Clothing/shops | 5 |
| Yard | 6 |
| apartament | 7 |
| convenienc | 8 |
| restaurant | 9 |
| hospital | 10 |
| Bus/rail | 11 |
| Park | 12 |
| vacant bui | 13 |

Tabella 3.4: Leggenda dei valori relativi al luogo di avvenimento del crimine

In questo modo risulterà possibile condurre l'analisi multivariata e ottenere dei risultati tenendo a mente tale leggenda di conversione.

## 3.2 Regressione Univariata

La Regressione univariata, o più comunemente denominata Regressione lineare è una metodologia che viene usata per predire il valore di una variabile dipendente Y con una o più variabili predittive in ingresso X. Lo scopo della regressione lineare è modellare una variabile continua Y come una funzione matematica di una o più variabili X , in modo da poter utilizzare questo modello di regressione per prevedere Y quando si conosce solo X. Questa equazione matematica può essere generalizzata come segue:

$Y=\beta_0+\beta_1 X + \epsilon_\mathrm{i}$.

Dove $\beta_0$ è l'intercetta, $\beta_1$ è la pendenza e vengono generalmente chiamati coefficienti di regressione . L'ultimo valore $\epsilon$ è il termine di errore. L'analisi proposta parte con un analisi grafica , per la costruzione di un semplice modello di regressione di previsione della variabile dipendente stabilendo una relazione lineare statisticamente significativa con la variabile indipendente( nelle nostre analisi Anno e riferimento spaziale della regione). I grafici a supporto di tale analisi sono :

- **Grafico a dispersione**: visualizza la relazione lineare tra le variabili
- **Box plot**: individua eventuali osservazioni anomale o outliers nella variabile indipendente
- **Grafico densità**: visualizza la distribuzione della variabile indipendente. Idealmente preferibile una distribuzione vicina alla normale(curva a campana).

**Correlazione**

Il grafico a dispersione offre un'ottima rappresentazione visiva della relazione lineare delle varaibili, tuttavia per quantificare in modo oggettivo e preciso la forza della relazione che c'è tra le due variabili andiamo a calcolare l'indice di **Correlazione**.
La correlazione è una misura statistica che suggerisce il livello di dipendenza lineare tra due variabili e ne indica la forza e la direzione. Tale indice può assumere valori compresi tra -1 e +1. Se si avvicina più a 1 avremo una correlazione positiva, se invervamente si avvicina a -1 si avrà una correlazione negativa, invece un valore più vicino a 0 suggerisce una relazione debole tra le variabili, infine si ha bassa correlazione quando -0,2<x <0,2 e suggerisce probabilmente che gran parte della variazione della variabile dipendente non è spiegata da quella indipendente, nel qual caso dovremmo probabilmente cercare variabili esplicative migliori. Sulla base di ciò si possono distinguere tre tipi di relazione:

- relazione positiva : all'aumentare dei valori di una variabile, aumentano in media anche i valori dell'altra variabile.
- relazione negativa : all'aumentare dei valori di una variabile, i valori dell'altra variabile in media diminuiscono.
- relazione nulla : all'aumentare dei valori di una variabile, i valori dell'altra variabile non risulteranno in media né aumentare né diminuire.

Il passo successivo consiste nella costruzione del modello lineare. La regressione lineare, che rappresenta la relazione più semplice e frequente tra due variabili quantitative, può essere positiva o negativa : tale relazione è indicata dal segno del coefficiente $\beta$.
La formula per calcolare il coefficiente angolare e l'intercetta è la seguente :

$$b = \frac{Codevianza(XY)}{Devianza(X)} = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

$$a = \bar{Y} - b\bar{X}$$

Per costruire la retta che descrive la distribuzione dei punti, ci si può riferire a diversi principi. Il più comune è il metodo dei minimi quadrati , ed è il metodo utilizzato dal software statistico R. Il modello lineare utilizza la funzione lm(), funzione che accetta due argomenti principali :

- Formula
- Dati

I dati sono tipicamente un dataframe e la formula è un oggetto di una classe formula. Ma la convenzione più comune è scrivere la formula direttamente al posto dell'argomento.
Il primo problema che si pone è quello di decidere quale sia la variabile dipendente Y e quale la variabile indipendente X. In generale definiamo :

- **variabile indipendente** : variabile affetta da errore durante la misura (o affetta da errore casuale).
- **variabile dipendente** : variabile affetta da errore, e di cui si vuole stimare una relazione.

Riportiamo un esempio molto semplice per comprendere il concetto che poi verrà applicato alle singole analisi. Supponiamo di voler ricavare una relazione lineare tra il peso (kg) e l'altezza (cm) di 10 individui.
**Altezza**: 175, 168, 170, 171, 169, 165, 165, 160, 180, 186.
**Peso**: 80, 68, 72, 75, 70, 65, 62, 60, 85, 90.
Assumiamo che la variabile Peso sia la variabile indipendente (X), e la variabile Altezza quella dipendente (Y). Quindi il nostro problema è quello di cercare una relazione lineare che ci permetta di calcolare l'altezza, essendo noto il peso di un individuo. La formula più semplice è quella di una generica retta del tipo $Y = a + b\ X$. In R si calcolano i due parametri procedendo in questo modo:

```
> Altezza = c( 175, 168, 170, 171, 169, 165, 165, 160, 180, 186)
> Peso=c(80, 68, 72, 75, 70, 65, 62, 60, 85, 90)
> modelloLineare= lm(formula=Altezza~Peso, x=TRUE, y=TRUE)
> modelloLineare

Call:
lm(formula = Altezza ~ Peso, x = TRUE, y = TRUE)
```

```
Coefficients:
(Intercept)         Peso
   115.2002       0.7662
```

Ora che abbiamo costruito il modello lineare, abbiamo anche stabilito la relazione tra le variabili sotto forma di una formula matematica per l'altezza in funzione del peso. L'output della funzione è rappresentato dai due parametri $a$ e $b$: $a$=115.2002 (l'intercetta), $b$=0.7662 (il coefficiente angolare). Il semplice calcolo della retta non è però sufficiente. Occorre valutare la significatività della retta, ossia se il coefficiente angolare $b$ si discosta da zero in modo significativo. A questo proposito si visualizzano le statistiche di riepilogo da cui poi trarre considerazioni importanti.

```
> summary(modelloLineare)

Call:
lm(formula = Altezza ~ Peso, x = TRUE, y = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6622 -0.9683 -0.1622  0.5679  2.2979

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.20021    3.48450   33.06 7.64e-10 ***
Peso          0.76616    0.04754   16.12 2.21e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.405 on 8 degrees of freedom
Multiple R-squared:  0.9701,Adjusted R-squared:  0.9664
F-statistic: 259.7 on 1 and 8 DF,  p-value: 2.206e-07
```

Le statistiche di riepilogo sopra ci dicono una serie di cose. Innanzitutto osserviamo che sono stati forniti anche in questo caso i valori dell'**intercetta** e del **coefficiente angolare**. I valori p sono molto importanti perché, possiamo considerare un modello lineare statisticamente significativo solo quando entrambi questi valori p sono inferiori al livello di significatività statistica predeterminato, che è idealmente 0,05. Questo è interpretato visivamente

dalle stelle significative alla fine della riga. Maggiore è il numero di stelle accanto al valore p della variabile, più significativa è la variabile.
**Ipotesi nulla e alternativa**:
Quando c'è un valore p, c'è un'ipotesi nulla e un'ipotesi alternativa ad esso associata. Nella regressione lineare, l'ipotesi nulla è quando i coefficienti associati alle variabili sono uguali a zero. L'ipotesi alternativa è che i coefficienti non siano uguali a zero (cioè esiste una relazione tra la variabile indipendente in questione e la variabile dipendente).

$$\begin{cases} H_0 : \beta = 0 & \text{Coefficiente non significativo} \\ H_1 : \beta \neq 0 & \text{Coefficiente significativo} \end{cases}$$

Dal risultato ottenuto bisogna considerare importanti risultati, uno di questi è **Multiple R-squared o** $R^2$che è il coefficiente di determinazione, e descrive la **bontà** del modello, indica la proporzione di variazione nella variabile dipendente che è stata spiegata da questo modello, in altre parole quanto il modello trovato spiega i dati campionari. In questo caso **"Multiple R-squared"**è pari a "0.9701" ciò permette di dedurre che il modello è adeguato a descrivere il 97% dei dati ricavati.
Per quanto concerne il valore t possiamo affermare che un valore t più grande indica che è meno probabile che il coefficiente non sia uguale a zero. Quindi, più alto è il valore t, meglio è. Il test t di Student relativo al coefficiente angolare ha in questo caso valore 16.12 (quello relativo alla variabile peso); il risultato della statistica F di Fisher è invece 259.7. In entrambi i casi questi valori sono ampiamente superiori ai valori Pr(>| t |), difatti i relativi p-value sono di molto inferiori a 0.05, e pertanto viene rifiutata l'ipotesi nulla (b = 0): quindi la regressione è significativa (il valore del coefficiente angolare così calcolato è statisticamente differente da zero). Infine è necessatrio ricordare che sia gli errori standard che la statistica F sono misure di bontà di adattamento.

### 3.2.1 Analisi di regressione lineare a confronto

Come anticipato, in questa sezione andremo a visualizzare i risultati delle analisi di regressione lineare relativi ai tre differenti datasets, analisi eseguiti con R Studio.

**RISCHIO DI CRIMINALITA'**

Il primo dataset che prendiamo in considerazione è relativo al Rischio di criminalità.

Il dataset è stato pulito e circoscritto alla regione dell'Emilia Romagna.
Da console possiamo visualizzare i dati con il seguente comando:

```
> data <- read.csv2("/Users/utente/Desktop/TESI/Rischiocriminalità.csv")
> View(data)
```

Il dataframe visualizzato è il seguente:

| | Regione | Anno | Rischio |
|---|---|---|---|
| 1 | Emilia-Romagna | 2001 | 28.18 |
| 2 | Emilia-Romagna | 2002 | 25.97 |
| 3 | Emilia-Romagna | 2003 | 24.30 |
| 4 | Emilia-Romagna | 2004 | 24.30 |
| 5 | Emilia-Romagna | 2005 | 24.25 |
| 6 | Emilia-Romagna | 2006 | 28.62 |
| 7 | Emilia-Romagna | 2007 | 31.04 |
| 8 | Emilia-Romagna | 2008 | 35.94 |
| 9 | Emilia-Romagna | 2009 | 26.48 |
| 10 | Emilia-Romagna | 2010 | 25.23 |
| 11 | Emilia-Romagna | 2011 | 25.60 |
| 12 | Emilia-Romagna | 2012 | 25.40 |
| 13 | Emilia-Romagna | 2013 | 32.40 |

Figura 3.5: Dati relativi al rischio di Criminalità in Emilia Romagna nel periodo 2001-2013

Presentiamo in maniera ordinata il procedimento di analisi seguito.
Vengono prese in considerazione le seguenti variabili:

- Anno :variabile indipendente
- Rischio: variabile dipendente

Partiamo con un'analisi grafica per costruire un semplice modello per prevedere il rischio di criminalità stabilendo una relazione lineare statisticamente significativa con l'anno in cui si è verificata tale criminalità, con riferimento spaziale alla regione dell'Emilia Romagna.

Andiamo a costruire il grafico a dispersione per indiviaduare la relazione tra la variabile dipendente (Rischio) e la variabile indipendente (Anno). Il codice in R che ne permette la visualizzazione grafica è :

```
> scatter.smooth(x=data$Anno, y=data$Rischio, main="Rischio ~ Anno")
```

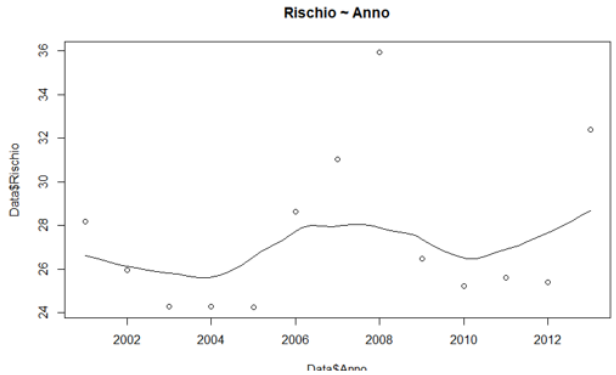



Figura 3.6: Rappresentazione del grafico a dispersione per le variabili Rischio-Anno

Dal grafico a dispersione, la relazione tra le variabili non sembra essere completamente positiva, questo sicuramente è dovuto alla presenza di pochi valori in gioco. Tuttavia l'esattezza della relazione verrà determinata in seguito attraverso il calcolo dell'indice di **Correlazione**.

Proseguiamo con la verifica della presenza di valori anomali. L'utilizzo del boxplot permette di rappresentare sullo stesso grafico cinque tra le misure di posizione più utilizzate in statistica. La sintesi a cinque numeri di una variabile quantitativa costituisce infatti la struttura del boxplot. Al suo interno sono indicati il valore minimo, il primo quartile , la mediana , il terzo quartile ed il valore massimo di una variabile. In genere, qualsiasi punto che si trova al di fuori dell'intervallo interquartile di 1,5 * ( 1,5 * I Q R ) è considerato un valore anomalo, dove l'IQR è calcolato come la distanza tra i valori del 25 percentile e del 75 percentile per quella variabile. Graficamente per i dati che possediamo otteniamo i seguenti grafici :

```
> par(mfrow=c(1, 2))
> boxplot(data$Anno, main="Anno", sub=paste("Outlier rows: ",
boxplot.stats(data$Anno)$out))
> boxplot(data$Rischio, main="Rischio", sub=paste("Outlier rows: ",
boxplot.stats(data$Rischio)$out))
```

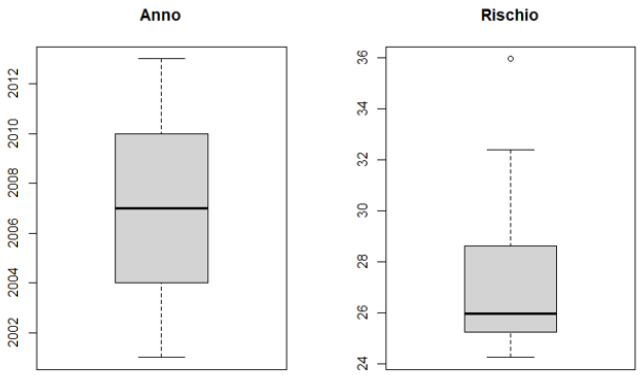



Figura 3.7: Rappresentazione del boxplot dei valori anomali-outliers

Dal boxplot Anno possiamo notare una distribuzione simmetrica, il primo ed il terzo quartile sono alla stessa distanza dalla mediana. In altre parole, la linea della mediana si trova esattamente a metà della scatola. Si nota come il 50% dei dati relativo all'anno ricada nell'intervallo compreso tra 2004 (primo quartile, corrispondente al 25esimo percentile) e 2010 (terzo quartile, corrispondente al 75esimo percentile).Inoltre, non sono presenti valori anomali e di conseguenza gli estremi dei baffi corrispondono con il valore minimo (2001) e massimo (2013).
Dal boxplot Rischio invece possiamo notare la presenza di un outlier, rappresentano come punto in corrispondenza del valore 36 , ciò significa che il rischio di criminalità ha toccato solo una volta quel valore, mentre il 50% delle percentuali di rischio considerati ricoprono un intervallo compreso tra 25(primo quartile) e 27(terzo quartile) con la mediana più bassa a indicare un accentramento dei dati verso un rischio pari a 26. La distribuzione in questo caso è asimmetrica a destra.

Visualizziamo ora il grafico di densità, che permette di determinare se la variabile di risposta, ovvero l'Anno, sia vicino alla normalità. Il grafico che otteniamo è il seguente:

```
>library(e1071)
>par(mfrow=c(1, 2))
>plot(density(data$Anno), main="Density Plot:Anno" ,ylab="Frequency",
sub=paste("Skewness:", round(e1071::skewness(data$Anno),2)))
>polygon(density(data$Anno),col="red")

>library(e1071)
>par(mfrow=c(1, 2))
>plot(density(data$Rischio), main="Density Plot:Rischio", ylab="Frequency",
```

```
sub=paste("Skewness:",round(e1071::skewness(data$Rischio),2)))
>polygon(density(data$Rischio),col="red")
```

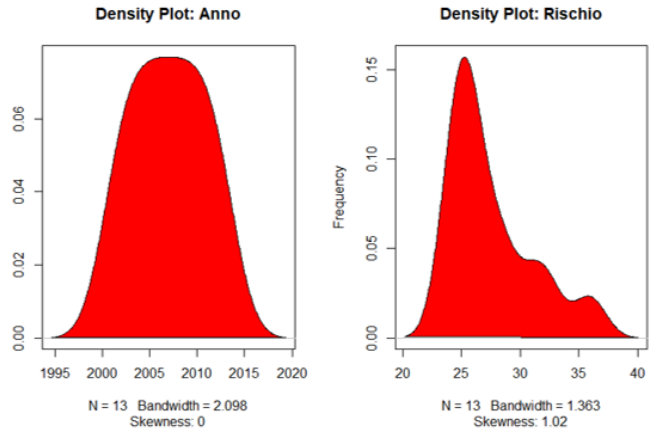



Figura 3.8: Rappresentazione del grafico di densità per ciascuno variabile

Il grafico di densità relativo all'anno presenta una forma a campana e dunque una distribuzione vicina alla normale, come si presumeva. Viene parallelamente rappresentato il grafico di densità relativo al rischio per conferma di una distribuzione variabile e non normale.
Calcoliamo ora con esattezza il rapporto tra le variabii. Il livello di dipendenza lineare tra le due variabili è possibile ottenerlo calcolando la correlazione come segue:

```
>cor(data$Anno, data$Rischio)
[1] 0.2498161
```

Per definizione sappiamo che la correlazione può assumere valori compresi tra -1 e +1. Nel nostro caso la correlazione fra anno e rischio di criminalità è x $>$ 0,2, dunque risulta essere abbastanza positiva ed esplicativa, ovvero la variabile dipendente è sufficientemente spiegata dalla variabile indipendente.

Costruiamo ora il modello lineare utilizzando la funzione lm() e gli argomenti che accetta come parametri nella seguente modalità:

```
>linearMod <- lm(Rischio ~ Anno, data=data)
>print(linearMod)

Call:
lm(formula = Rischio ~ Anno, data=data)
```

```
Coefficients
(Intercept)            Anno
-438.2843            0.2321
```

Ora che abbiamo definito il modello lineare abbiamo anche stabilito la relazione tra la variabile indipendente e dipendente sotto forma di una formula matematica per il Rischio in funzione dell'Anno. L'output risultante qui sopra presenta i Coefficienti in due componenti:

- Intercetta: -438.2843
- Anno: 0.2321

Questi sono chiamati coefficienti $\beta$ , risalendo alla formula teorica per cui si ha:
$Y = \beta_0 + \beta_1 X + \epsilon_i$.

In altre parole abbiamo : Rischio= Intercept+($\beta$*Anno). La retta di equazione lineare sarà y(x)=-438.2843+0.2321*x.

In R eseguiamo il seguente codice:

```
> coefregr <- lm(Rischio~Anno, data=data)
> coef(coefregr)
(Intercept)           Anno
-438.2842857    0.2320879

>plot(Rischio~Anno, data=data, pch=16)
>abline(coef(coefregr))
```

A livello grafico la retta si presenta così :

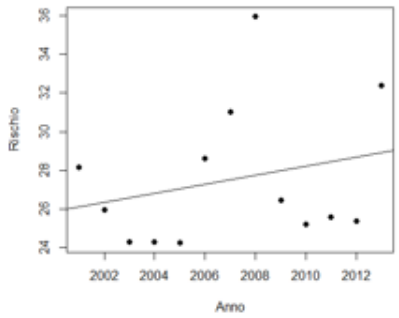



Figura 3.9: Rappresentazione della retta di equazione lineare che rapporta le due variabili

Dal grafico, possiamo dedurre come il rischio di criminalità abbia tassi differenti sopra e sotto la retta. I punti sotto rappresentano un basso rischio, i punti al di sopra della retta presentano tassi elevati , il picco viene raggiunto nell'anno 2008 con un rischio pari al 36. Dal grafico siamo in grado, inoltre di confermare il risultato della correlazione, che risulta essere positiva .
Con il semplice comando summary() in R , abbiamo la diagnostica di regressione lineare che ci dà le informazioni relative a minimo, massimo, media, range interquantile. In più per i coefficienti di rischio e anno ricaviamo corrispondentemente le informazioni relative all'errore standard, t-value.Otteniamo anche il valore di $R^2$,il p-value e il valore F-statistic.

```
> linearMod <- lm(Rischio~Anno, data=EmiliaStatistica)
> summary(linearMod)

Call:
lm(formula = Rischio ~ Anno, data = EmiliaStatistica)

Residuals:
   Min     1Q Median     3Q    Max
-3.277 -2.802 -1.500  2.056  8.192

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -438.2843   544.3658  -0.805    0.438
Anno           0.2321     0.2712   0.856    0.410

Residual standard error: 3.659 on 11 degrees of freedom
Multiple R-squared:  0.06241,Adjusted R-squared:  -0.02283
F-statistic: 0.7322 on 1 and 11 DF,  p-value: 0.4104
```

Dai risultati ottenuti siamo in grado di determinare se la regressione sia significativa o meno. Il coefficiente di determinazione è pari a 0.06241 , ciò significa che il modello è adeguato a descrivere solamente lo 0.07% dei dati ricavati. Il valore t del test Student relativo alla variabile Anno ha valore 0.856, mentre il valore della statistica F di Fisher è 0.7322. In entrambi i casi questi valori risultano essere superiori al valore p, che tuttavia supera il valore ideale di 0.05. Questo implica il non rifiuto dell'ipotesi nulla , cioè che il coefficiente $\beta$ del predittore sia zero , e dunque che il modello non sia statisticamente significativo. Sulla basi di queste informazioni possiamo costruire anche i seguenti grafici:

- Istogramma

- Q-Q plot

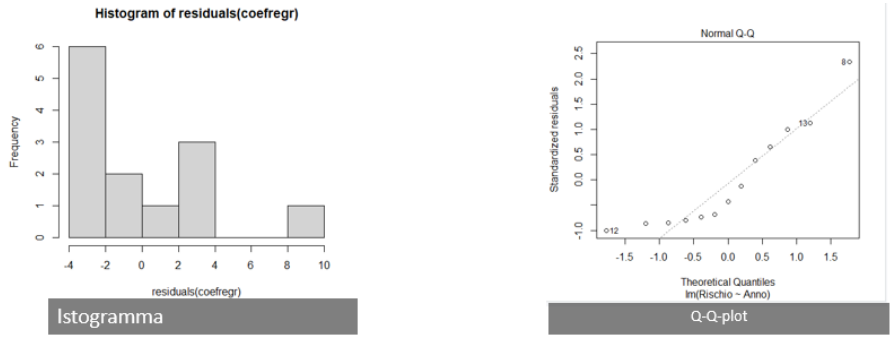



Figura 3.10: Rappresentazione grafica dell'Istogramma e Q-Qplot

Il modello di regressione lineare inoltre ci permette anche di predire valori futuri. Per fare questo, basta calcolare, nell'equazione della retta, il valore di y in corrispondenza della x desiderata.

```
> fitted(linearMod)[1:5]
       1        2        3        4        5
26.12363 26.35571 26.58780 26.81989 27.05198

> predict(linearMod, newdata = data.frame(Anno=c(7,9,2)))
        1         2         3
-436.6597 -436.1955 -437.8201
plot(linearMod)
```

È interessante analizzare il grafico dei residui rispetto ai valori predetti, come ulteriore informazione sulla bontà del modello.

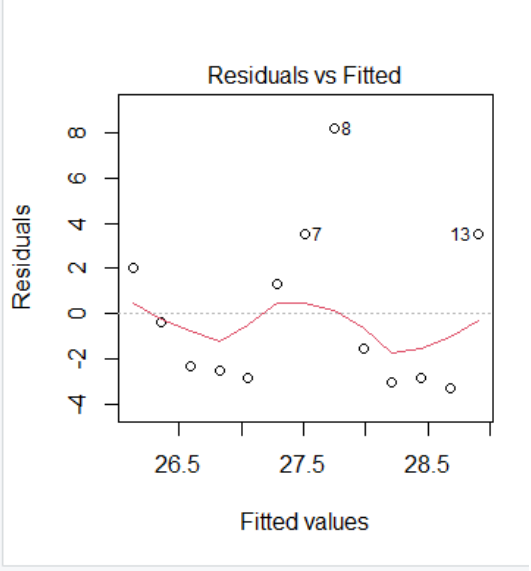



Figura 3.11: Rappresentazione grafica del grafico dei residui

## TASSO DI FURTI E OMICIDI A CONFRONTO

Allo stesso modo con cui è stata condotta la analisi sopra, riportiamo quella relativa agli altri due dataset a confronto. Tuttavia il codice R non verrà ripetuto in quanto ha la stessa sintassi di quella proposta, verranno mostrati , invece, i grafici , e discussi i risultati.
I datasets proposti fanno riferimento a statistiche dei furti ogni 1000 abitanti e al tasso di omicidio, sempre all'interno della regione Emilia Romagna. Il peridio di tempo per entrambi i set di dati copre il periodo 1995-2013. Visualizziamo i datasets:

| | Anno | Regione | Furti |
|---|---|---|---|
| 1 | 1995 | Emilia-Romagna | 25.18 |
| 2 | 1996 | Emilia-Romagna | 25.98 |
| 3 | 1997 | Emilia-Romagna | 29.09 |
| 4 | 1998 | Emilia-Romagna | 31.22 |
| 5 | 1999 | Emilia-Romagna | 29.73 |
| 6 | 2000 | Emilia-Romagna | 30.70 |
| 7 | 2001 | Emilia-Romagna | 29.13 |
| 8 | 2002 | Emilia-Romagna | 29.31 |
| 9 | 2003 | Emilia-Romagna | 30.37 |
| 10 | 2004 | Emilia-Romagna | 35.13 |
| 11 | 2005 | Emilia-Romagna | 36.26 |
| 12 | 2006 | Emilia-Romagna | 36.61 |
| 13 | 2007 | Emilia-Romagna | 38.53 |
| 14 | 2008 | Emilia-Romagna | 31.26 |
| 15 | 2009 | Emilia-Romagna | 29.80 |
| 16 | 2010 | Emilia-Romagna | 28.40 |
| 17 | 2011 | Emilia-Romagna | 33.14 |
| 18 | 2012 | Emilia-Romagna | 34.37 |
| 19 | 2013 | Emilia-Romagna | 34.99 |

Figura 3.12: Dati relativi ai furti in Emilia-Romagna

| | Anno | Regione | tasso |
|---|---|---|---|
| 1 | 1995 | Emilia-Romagna | 0.77 |
| 2 | 1996 | Emilia-Romagna | 0.77 |
| 3 | 1997 | Emilia-Romagna | 0.95 |
| 4 | 1998 | Emilia-Romagna | 0.92 |
| 5 | 1999 | Emilia-Romagna | 0.79 |
| 6 | 2000 | Emilia-Romagna | 0.78 |
| 7 | 2001 | Emilia-Romagna | 0.86 |
| 8 | 2002 | Emilia-Romagna | 0.85 |
| 9 | 2003 | Emilia-Romagna | 1.07 |
| 10 | 2004 | Emilia-Romagna | 0.76 |
| 11 | 2005 | Emilia-Romagna | 0.66 |
| 12 | 2006 | Emilia-Romagna | 0.70 |
| 13 | 2007 | Emilia-Romagna | 0.62 |
| 14 | 2008 | Emilia-Romagna | 0.71 |
| 15 | 2009 | Emilia-Romagna | 0.82 |
| 16 | 2010 | Emilia-Romagna | 0.51 |
| 17 | 2011 | Emilia-Romagna | 0.85 |
| 18 | 2012 | Emilia-Romagna | 0.85 |
| 19 | 2013 | Emilia-Romagna | 0.59 |

Figura 3.13: Dati relativi agli omicidi in Emilia-Romagna

Definiamo le variabili per entrambi i casi:
Per il dataset relativo ai furti le variabili sono:

- Anno: variabile indipendente
- Furti: variabile dipendente

Per il dataset realtivo al tasso di Omicidi le variabili sono:

- Anno: variabile indipendente
- tasso: variabile dipendente

Proseguiamo con l'analisi grafica e dunque la visualizzazione nell'ordine dei seguenti grafici:

1. grafico a dispersione

2. boxplot dei valori anomali

3. grafico di densità

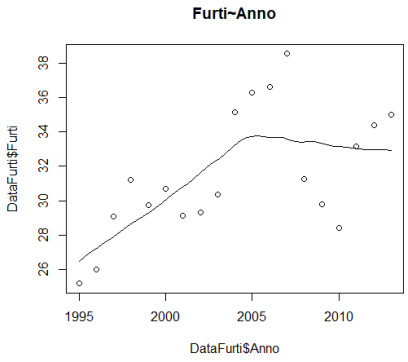



Figura 3.14: Rappresentazione del grafico di dispersione per le variabili Furti-Anno

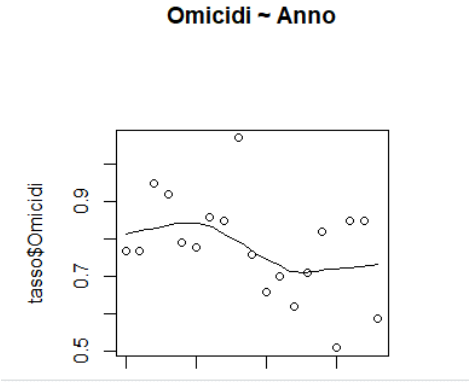



Figura 3.15: Rappresentazione del grafico di dispersione per le variabili tasso-Anno

Dai grafici a dispersione è possibile rappresentare la relazione delle variabili in gioco. Il grafico a dispersione o scatterplot rappresenta il metodo più utilizzato in statistica descrittiva per valutare la relazione tra due variabili quantitative. In questo tipo di grafico le due variabili sono riportate su uno spazio cartesiano. I valori di una variabile sono indicati sull'asse orizzontale delle x, mentre i valori dell'altra variabile sono rappresentati sull'asse verticale delle y. Ogni unità statistica è rappresenta da un punto posizionato sul grafico in base alle sue coordinate. Quindi questo grafico sarà costituito da tanti punti quante sono le unità statistiche oggetto di studio. É possibile notare come la relazione delle variabili realtive ai furti risultano essere positivamente correlate, all'aumetare di una, aumenta anche l'altra. Contrariamente la relazione relativa ai dati di omicidio sembra non essere esplicativa, ovvero all'aumentare della variabile Anno ,la variabile tasso non risulta in media né aumentare né diminuire.In quest'ultimo caso si dice che la relazione tra le variabili è nulla.
Proseguiamo con la verifica di valori anomali e la rispettiva rappresentazione:

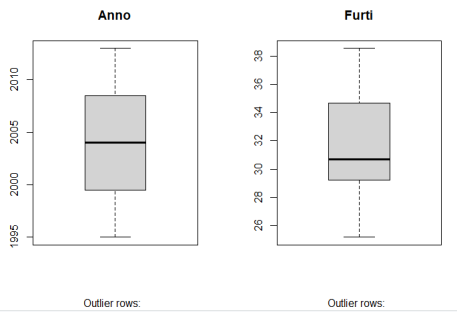



Figura 3.16: Rappresentazione del boxplot di valori anomali nel dataset dei furti

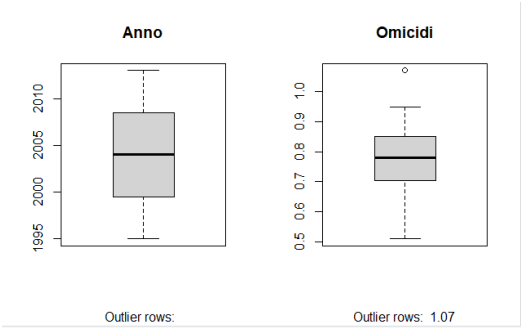



Figura 3.17: Rappresentazione del boxplot di valori anomali nel dataset del tasso di omicidi

Dai Boxplot rappresentati si può notare come la distribuzione dell'Anno sia simmetrica per entrambi i casi, e la mediana si trova esattamente tra il primo e il terzo quartile. Si nota come il 50% dei dati relativo all'anno ricada nell'intervallo compreso tra 1995 (primo quartile, corrispondente al 25esimo percentile) e 2010 (terzo quartile, corrispondente al 75esimo percentile).Inoltre, non sono presenti valori anomali e di conseguenza gli estremi dei baffi corrispondo con il valore minimo (1995) e massimo (2010). Tale osservazione può essere riferita anche in confronto con il Boxplot Anno del dataset relatico al Rischio di criminalità, in particolare alla Figura 3.5. Focalizzandoci invece sul boxplot dei furti notiamo come la distribuzione dei dati sia asimmetrica a destra, ovvero la mediana risulta più vicino al primo quartile. La presenza di valori anomali si riscontra solo per il dataset degli omicidi , col la presenza di un outlier ad altezza 1.07. In generale il 50% dei dati si dilata tra valori compresi tra 0.7 e poco meno di 0.9.

Infine visualizziamo il grafico di densità per le due analisi:

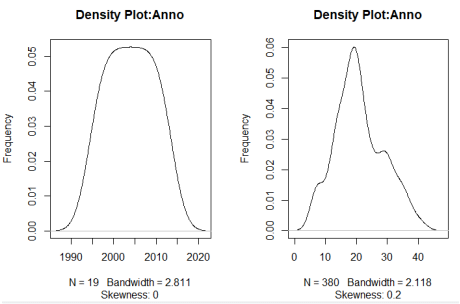



Figura 3.18: Rappresentazione del grafico di densità per i furti

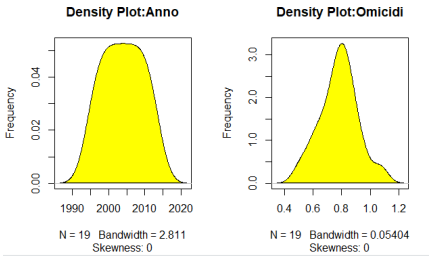



Figura 3.19: Rappresentazione del grafico di densità per gli omicidi

Il grafico di densità viene utilizzato per analizzare la distribuzione della variabile indipendente "Anno" in entrambe le analisi tale distribuzione risulta essere normale, questo lo dimostra la forma a campana del grafico.

La relazione tra le variabili come anticipato è meglio quantificata dalla correlazione . Di seguito calcoliamo le relazioni per entrambe le analisi e le visualizziamo corrispondentemente come segue:

```
> cor(dataset2$Anno, dataset2$Furti)
[1] 0.5912484
```

```
 > cor(tassoOmicidio$Anno, tassoOmicidio$tasso)
[1] -0.4245134
```

É evidente come la correlazione rifertita al rapporto tra le variabili dell'analisi dei furti sia positiva, contrariamente il valore della correlazione riferita all'analisi del tasso di omicidio risulta essere negativo e adirittura minore del del -0.2 per cui la correlazione delle variabili si può considerare di bassa esplicabilità. Costruiamo ora il modello lineare riportando i risultati nel seguente ordine:

**DATASET FURTI**

```
> lmFurti <- lm(Furti~Anno, data=dataset2)
> print(lmFurti)

Call:
lm(formula = Furti ~ Anno, data = dataset2)

Coefficients:
(Intercept)         Anno
  -735.3272       0.3827
```

La retta di regressione la otteniamo nel seguente modo:

```
> coefficienti <- lm(Furti~Anno, data = dataset2)
> coef(coefficienti)
 (Intercept)         Anno
-735.3271579    0.3826667
```

Graficamente:

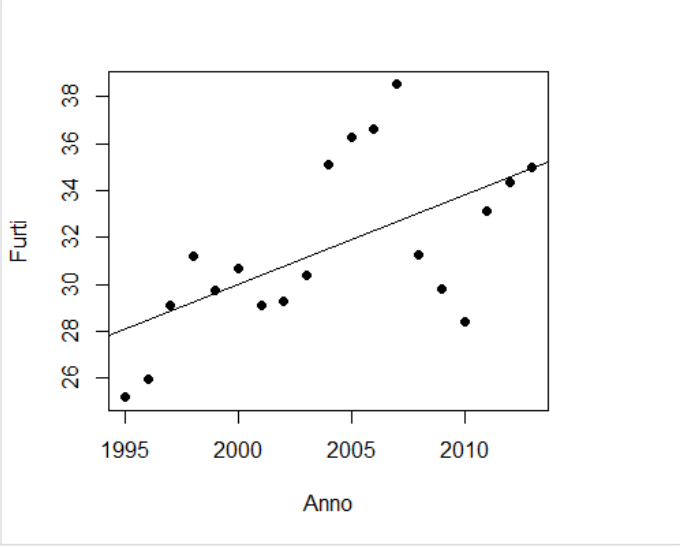



Figura 3.20: Rappresentazione della retta di regressione lineare dei furti

Dalla retta, possiamo dedurre come il tasso di furti sia differente sopra e sotto la retta. I punti sotto rappresentano un basso indice di furto, i punti al di sopra della retta presentano tassi elevati , il picco viene raggiunto nell'anno 2007 con un valore pari circa a 38. Dall'andamento dei punti sul grafico e l'inclinazione della retta si può confermare che ci sia una buona correlazione delle variabili.
La diagnostica significativa dei valori che otteniamo dalla modellazione lineare è la seguente:

```
> summary(lmFurti)

Call:
lm(formula = Furti ~ Anno, data = dataset2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4328 -1.6345 -0.2282  1.3365  5.8452

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -735.3272   253.7022  -2.898  0.01000 **
Anno           0.3827     0.1266   3.023  0.00767 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 3.022 on 17 degrees of freedom
Multiple R-squared:  0.3496,Adjusted R-squared:  0.3113
F-statistic: 9.137 on 1 and 17 DF,  p-value: 0.007674
```

Analizziamo gli indicatori più significativi: il coefficiente di determinazione è pari a 0.3496 , ciò significa che il modello è adeguato a descrivere quasi il 35% dei dati ricavati. Il valore t del test Student relativo alla variabile Anno ha valore 3.023, mentre il valore della statistica F di Fisher è 9.137. In entrambi i casi questi valori risultano essere superiori al valore p corrispondente, che risulta inferiore al valore ideale di 0.05. Questo implica il rifiuto dell'ipotesi nulla.Possiamo affermare che tale modello sia statisticamente significativo.

**DATASET TASSI DI OMICIDIO**:

Riportiamo i passi della costruzione del modello lineare:

```
> lineare<- lm(tasso ~ Anno, data=tassoOmicidio)
> print(lineare)

Call:
lm(formula = tasso ~ Anno, data = tassoOmicidio)

Coefficients:
(Intercept)         Anno
      20.82        -0.01
```

La retta di regressione lineare è data da :

```
> coefregr <- lm(tasso~Anno, data=tassoOmicidio)
> coef(coefregr)
(Intercept)        Anno
   20.82053    -0.01000
> plot(tasso~Anno, data=tassoOmicidio, pch=16)
> abline(coef(coefregr))
```

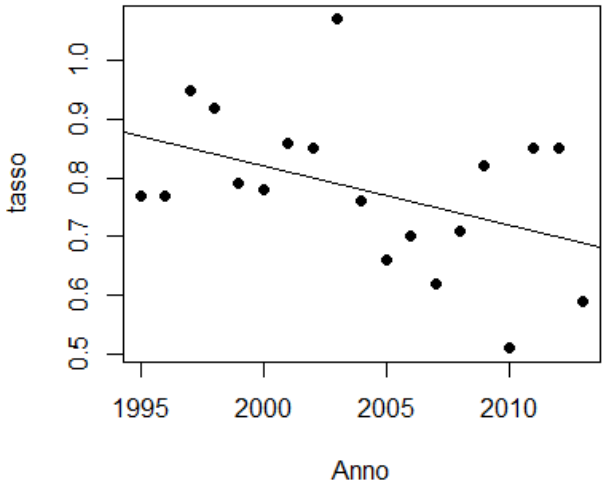



Figura 3.21: Rappresentazione della retta di regressione lineare del tasso di omicidio

Il grafico della retta di regressione lineare presenta un'inclinazione decrescente. Ciò conferma il risultato del coefficiente di correlazione che risulta essere negativo.
La diagnostica significativa dei valori che otteniamo dalla modellazione lineare è la seguente:

```
> summary(lineare)

Call:
lm(formula = tasso ~ Anno, data = tassoOmicidio)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21053 -0.09553 -0.03053  0.08447  0.27947

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.820526  10.366551   2.008   0.0608 .
Anno        -0.010000   0.005173  -1.933   0.0701 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.1235 on 17 degrees of freedom
Multiple R-squared:  0.1802,Adjusted R-squared:  0.132
F-statistic: 3.737 on 1 and 17 DF,  p-value: 0.07005
```

La diagnostica permette di definire se la regressione sia sigificativa, caratteristica non valida in questo specifico caso poichè abbiamo il coefficiente di determinazione pari a 0.1802, il valore t uguale a -1.933 che è minore del valore p. Il p-value stesso è maggiore , anche se di poco, del valore ideale di 0.05. Questo implica l'accettazione dell'ipotesi nulla e l'affermazione di un modello non sufficientemente significativo.
Da queste informazioni possiamo costruire i seguenti grafici:

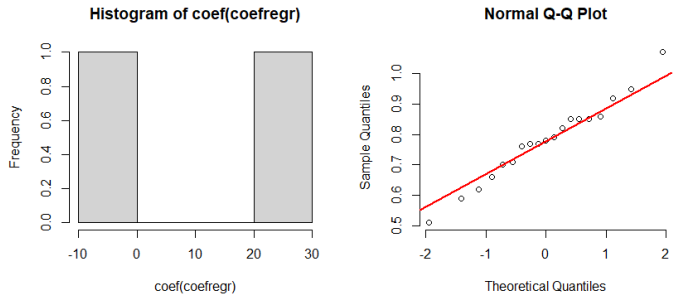



Figura 3.22: Rappresentazione dell'istogramma e del Q-Q Plot relativo al modello lineare del tasso di Omicidi

### 3.2.2 Regressione Polinomiale lineare

Accanto alla regressione lineare vi è quella polinomiale, che utilizza lo stesso metodo della regressione lineare, ma assume che la funzione che meglio descrive l'andamento dei dati non sia una retta, ma un polinomio. La formula matematica è definita come segue:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_n x^n$$

.

A che grado del polinomio fermarci? Dipende dal grado di precisione che cerchiamo. Maggiore è il grado del polinomio, maggiore sarà la precisione del modello, ma maggiori saranno le difficoltà di calcolo; inoltre bisogna verificare la significatività dei coefficienti che vengono trovati.
In R per fittare un modello di regressione polinomiale (non ortogonale) esistono due metodi, tra loro identici. Supponiamo di cercare i valori dei coefficienti beta per un polinomio di 1° grado, quindi di 2° grado, quindi di 3°

grado relativamente a tutte e tre le analisi descritte sopra.
Partiamo dal primo dataset concernente i dati relativi al rischio di criminalità e seguiamo i seguenti passaggi.
Calcoliamo il polinomio di grado 2 e la sua diagnostica:

```
> fit2<- lm(EmiliaStatistica$Rischio ~ EmiliaStatistica$Anno +
I(EmiliaStatistica$Anno^2))
> summary(fit2)

Call:
lm(formula = EmiliaStatistica$Rischio ~ EmiliaStatistica$Anno +
I(EmiliaStatistica$Anno^2))

Residuals:
   Min     1Q Median     3Q    Max
-3.105 -2.813 -1.656  2.399  7.989

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -6.313e+04  3.449e+05  -0.183    0.858
EmiliaStatistica$Anno        6.271e+01  3.437e+02   0.182    0.859
I(EmiliaStatistica$Anno^2) -1.556e-02  8.563e-02  -0.182    0.859

Residual standard error: 3.831 on 10 degrees of freedom
Multiple R-squared:  0.0655,Adjusted R-squared:  -0.1214
F-statistic: 0.3504 on 2 and 10 DF,  p-value: 0.7127
```

L'output di summary(fit2) è assolutamente identico a quello di primo grado. Abbiamo ottenuto i valori di $\beta 0$ (-6.313e+04), $\beta 1$ (6.271e+01) e $\beta 2$ (-1.556e-02). L'equazione del polinomio di 2° grado del nostro modello è quindi:

$$f(x) = -6.313e + 04 + 6.271e + 01x - 1.556e - 02^2$$

Il valore del Coefficiente di determinazione che misura la bontà del modello è pari a 0.0655, ciò significa che circa solo l'0.07 % dei dati viene spiegato dalle variabili esplicative. Il valore F- statistic(0.3504) risulta essere maggiore del valore F-tabulato( 0.7127) ciò Se calcoliamo il polinomio di grado 3 otteniamo:

```
> fit3<- lm(EmiliaStatistica$Rischio ~ EmiliaStatistica$Anno +
I(EmiliaStatistica$Anno^2) + I(EmiliaStatistica$Anno^3))
> summary(fit3)
```

```
Call:
lm(formula = EmiliaStatistica$Rischio ~ EmiliaStatistica$Anno +
I(EmiliaStatistica$Anno^2) + I(EmiliaStatistica$Anno^3))

Residuals:
   Min     1Q Median     3Q    Max
-3.105 -2.813 -1.656  2.399  7.989

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -6.313e+04  3.449e+05  -0.183    0.858
EmiliaStatistica$Anno       6.271e+01  3.437e+02   0.182    0.859
I(EmiliaStatistica$Anno^2) -1.556e-02  8.563e-02  -0.182    0.859
I(EmiliaStatistica$Anno^3)         NA         NA      NA       NA

Residual standard error: 3.831 on 10 degrees of freedom
Multiple R-squared:  0.0655,Adjusted R-squared:  -0.1214
F-statistic: 0.3504 on 2 and 10 DF,  p-value: 0.7127
```

In quest'ultimo caso tuttavia i coefficienti non sono significativi, quindi il modello migliore è il polinomio di grado 2.
Rappresentiamo ora graficamente i risultati ottenuti , in particolare visualizzeremo solamente le linee e non i punti, per comodità grafica:

```
> plot(EmiliaStatistica$Anno, EmiliaStatistica$Rischio, type="I", lwd=3)
```

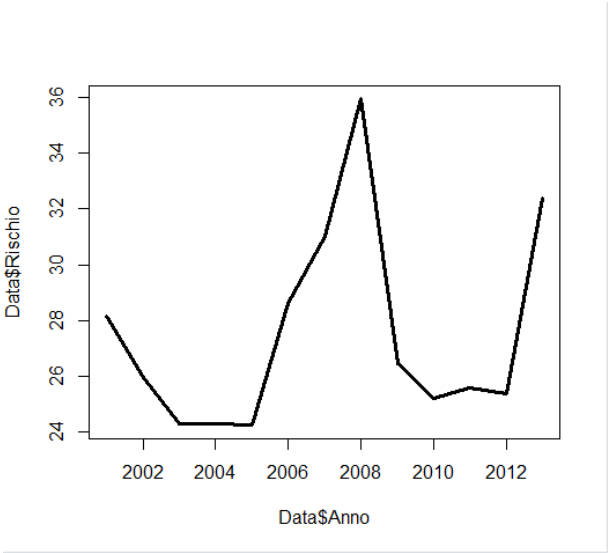



Figura 3.23: Rappresentazione grafica del modello polinomiale di grado 3 relativo al rischio di criminalità

Ora aggiungiamo a questo grafico l'andamento del polinomio di 2° grado in questo modo:

```
>points(EmiliaStatistica$Anno, predict(fit2), type="l", col="red", lwd=2)
```

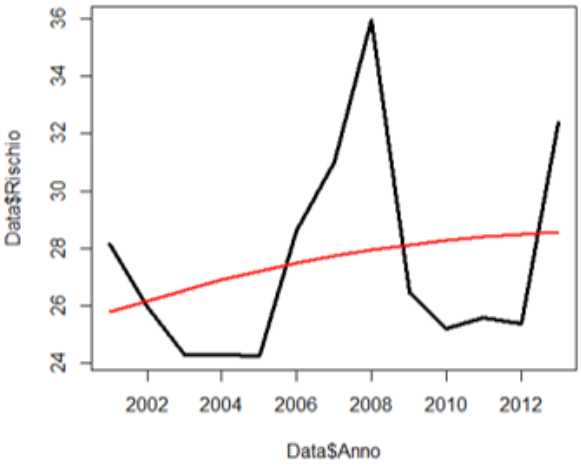



Figura 3.24: Rappresentazione grafica del modello polinomiale di grado 2 e 3 relativo al rischio di criminalità

Visualizziamo ora i risultati delle analisi polinomiali degli ultimi due datasets.

Regressione polinomiale relativa ai furti:

```
 fit2b<- lm(DataFurti$Furti ~ DataFurti$Anno + I(DataFurti$Anno^2))
> summary(fit2b)

Call:
lm(formula = DataFurti$Furti ~ DataFurti$Anno + I(DataFurti$Anno^2))

Residuals:
   Min     1Q Median     3Q    Max
-5.156 -2.174 -0.124  2.233  4.877

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.859e+05  9.694e+04  -1.917   0.0732 .
DataFurti$Anno       1.851e+02  9.675e+01   1.914   0.0737 .
I(DataFurti$Anno^2) -4.610e-02  2.414e-02  -1.910   0.0743 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.811 on 16 degrees of freedom
Multiple R-squared:  0.4703,Adjusted R-squared:  0.4041
F-statistic: 7.103 on 2 and 16 DF,  p-value: 0.006196
```

Abbiamo ottenuto i valori di $\beta 0$ (-1.859e+05), $\beta 1$ ( 1.851e+02 ) e $\beta 2$ (-4.610e-02). L'equazione del polinomio di 2° grado del nostro modello è quindi:

$$f(x) = -1.859e + 05 + 1.851e + 02x - 4.610e - 02^2$$

Se calcoliamo il polinomio di grado 3 otteniamo:

```
> fit3b<- lm(DataFurti$Furti ~ DataFurti$Anno +
I(DataFurti$Anno^2)+I(DataFurti$Anno^3))
> summary(fit3b)

Call:
lm(formula = DataFurti$Furti ~ DataFurti$Anno + I(DataFurti$Anno^2) +
I(DataFurti$Anno^3))

Residuals:
   Min     1Q Median     3Q    Max
```

```
-5.156 -2.174 -0.124  2.233  4.877

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.859e+05  9.694e+04  -1.917   0.0732 .
DataFurti$Anno       1.851e+02  9.675e+01   1.914   0.0737 .
I(DataFurti$Anno^2) -4.610e-02  2.414e-02  -1.910   0.0743 .
I(DataFurti$Anno^3)         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.811 on 16 degrees of freedom
Multiple R-squared:  0.4703,Adjusted R-squared:  0.4041
F-statistic: 7.103 on 2 and 16 DF,  p-value: 0.006196
```

Il modello di grado 2 risulta essere migliore, qui il coefficiente $\beta 3$ non è determinato.
Graficamente seguiamo lo stesso procedimento dell'analisi sopra e otteniamo:

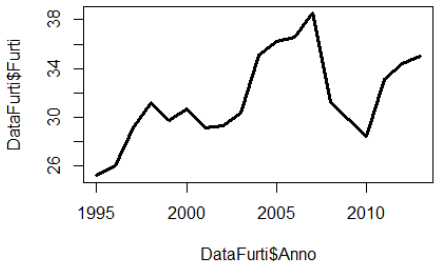



Figura 3.25: Rappresentazione grafica del modello polinomiale di grado 3 relativo ai furti

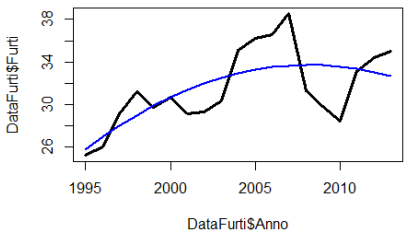



Figura 3.26: Rappresentazione grafica del modello polinomiale di grado 2 e 3 relativo ai furti

Regressione polinomiale relativa ai tassi di Omicidio:

```
> fit2c<- lm(tassoOmicidio$tasso ~ tassoOmicidio$Anno + I(tassoOmicidio$Anno^
> summary(fit2c)

Call:
lm(formula = tassoOmicidio$tasso ~ tassoOmicidio$Anno + I(tassoOmicidio$Anno^

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.20786 -0.07660 -0.03676  0.08470  0.26656

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.767e+03  4.367e+03  -0.405    0.691
tassoOmicidio$Anno       1.774e+00  4.358e+00   0.407    0.689
I(tassoOmicidio$Anno^2) -4.452e-04  1.087e-03  -0.409    0.688

Residual standard error: 0.1266 on 16 degrees of freedom
Multiple R-squared:  0.1887,Adjusted R-squared:  0.0873
F-statistic: 1.861 on 2 and 16 DF,  p-value: 0.1877
```

Abbiamo ottenuto i valori di $\beta 0$ (-1.767e+03), $\beta 1$ ( 1.774e+00 ) e $\beta 2$ (-4.452e-04). L'equazione del polinomio di 2° grado del nostro modello è quindi:

$$f(x) = -1.767e + 03 + 1.774e + 00x - 4.452e - 04^2$$

Se calcoliamo il polinomio di grado 3 otteniamo:

```
> fit3c<- lm(tassoOmicidio$tasso ~ tassoOmicidio$Anno +
I(tassoOmicidio$Anno^2)+ I(tassoOmicidio$Anno^3) )
> summary(fit3c)

Call:
lm(formula = tassoOmicidio$tasso ~ tassoOmicidio$Anno +
I(tassoOmicidio$Anno^2) +
    I(tassoOmicidio$Anno^3))

Residuals:
     Min       1Q   Median       3Q      Max
-0.20786 -0.07660 -0.03676  0.08470  0.26656

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.767e+03  4.367e+03  -0.405    0.691
tassoOmicidio$Anno       1.774e+00  4.358e+00   0.407    0.689
I(tassoOmicidio$Anno^2) -4.452e-04  1.087e-03  -0.409    0.688
I(tassoOmicidio$Anno^3)         NA         NA      NA       NA

Residual standard error: 0.1266 on 16 degrees of freedom
Multiple R-squared:  0.1887,Adjusted R-squared:  0.0873
```

```
F-statistic: 1.861 on 2 and 16 DF,  p-value: 0.1877
```

Graficamente otteniamo:

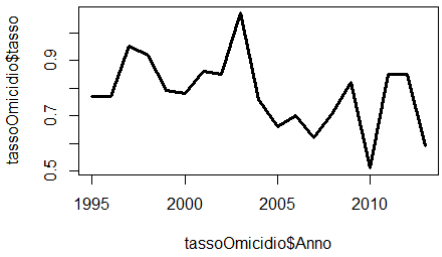



Figura 3.27: Rappresentazione grafica del modello polinomiale di grado 3 relativo ai tasso di omicidio

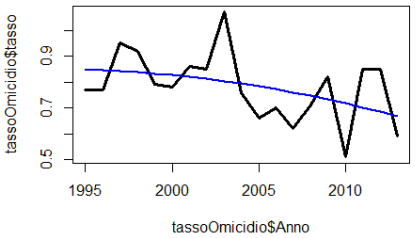



Figura 3.28: Rappresentazione grafica del modello polinomiale di grado 2 e 3 relativo ai tassi di omicidio

### 3.2.3 Risultati a confronto

Le tre analisi condotte , presentano solo un modello di regressione significativo , ovvero quello relativo ai tassi di furti. Questo risultato è supportato in primis dai grafici: il grafico a dispersione che anticipa la relazione positiva fra le variabili in analisi (Anno e Furti) , il boxplot con una distribuzione leggermente assimetrica e l'assenza di valori anomali o outliers, il grafico di densità che rispecchia la distribuzione normale della variabile indipendente. Gli indicatori numerici della diagnostica di regressione hanno confermato tale previsione. É stata svolta inoltre l'analisi polinomiale poichè maggiore è il grado del polinomio, maggiore sarà la precisione del modello. Da questa è possibile confermare il risultato appena descritto ovvero della significatività del modello relativo al dataset dei furti poichè la diagnostica anche nel caso polinomiale conferma che il modello sia statisticamente significativo.

## 3.3 Regressione Multivariata

Nella regressione lineare semplice, abbiamo immaginato che una certa variabile Y dipendesse dall'andamento di un'altra variabile (X), in maniera lineare con andamento crescente o decrescente. Abbiamo quindi visto come realizzare e disegnare la retta che pone in relazione le due variabili, e come valutare la bontà del modello. Nella realtà (scientifica, economica, psicometrica, etc.),

quasi mai un evento dipende solamente dall'andamento di un certo fattore. Tutti gli eventi (anche i più comuni) sono influenzati da numerosissimi elementi. Risulta pertanto molto più utile formulare un modello che tenga conto di tutte queste influenze. Ciò si ottiene con lo studio della regressione lineare multipla (o multivariata). In generale si indica con Y la variabile dipendente, e con X seguito da un numero in pedice le variabili indipendenti che si suppone abbiano un effetto. Le X vengono chiamate predittori e la formula generale del modello che cerchiamo è:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

dove $\beta_0$ è l'intercetta e $\beta_1, \beta_2$ sono i regressori, i quali rappresentano il coefficiente angolare della retta che otterremmo al variare del predittore corrispondente, qualora tutti gli altri predittori fossero costanti. La rappresentazione grafica dipende dal numero di predittori che si vogliono considerare. Con un solo predittore, si ottiene (come abbiamo visto) una retta; con 2 predittori si ottiene un piano nello spazio a 3 dimensioni; con 3 o più predittori non è possibile la rappresentazione grafica, in quanto occorrerebbe rappresentare uno spazio a più di 3 dimensioni.

Per l'analisi della regressione multivariata viene utilizzato un dataset che fa riferimento a diverse tipologie di crimine avvenute a Baltimore, città dello stato del Maryland negli stati Uniti d'America. I dati sono stati circoscritti con la scelta dell'anno 2017 e di un giorno in particolare il 10 di ogni mese e 6 orari differenti di una stessa giornata con un distacco di 4 ore.

Il dataset è stato adattato a valori numerici significativi , la cui conversione l'abbiamo vista nella fase di preparazione. Visualizziamo ora in R i dati utili all'analisi:

```
> library(readxl)
> analisi <- read_excel("C:\\Users\\utente\\Desktop\\TESI\\analisiMultipla.xlsx")
> View(analisi)
```

| | CrimeDate | CrimeTime | Location | Description | InOut | Weapon | Premise |
|---|---|---|---|---|---|---|---|
| 1 | 12 | 20 | 4800 | 2 | 0 | 3 | 2 |
| 2 | 12 | 16 | 600 | 1 | 1 | 1 | 0 |
| 3 | 12 | 12 | 6400 | 1 | 1 | 1 | 1 |
| 4 | 12 | 8 | 2800 | 6 | 1 | 1 | 3 |
| 5 | 12 | 4 | 200 | 2 | 0 | 0 | 2 |
| 6 | 12 | 24 | 3800 | 6 | 1 | 1 | 3 |
| 7 | 11 | 20 | 5100 | 2 | 0 | 3 | 2 |
| 8 | 11 | 16 | 2000 | 1 | 0 | 1 | 2 |
| 9 | 11 | 12 | 3500 | 3 | 1 | 2 | 3 |
| 10 | 11 | 8 | 2900 | 2 | 0 | 2 | 2 |
| 11 | 11 | 4 | 1000 | 1 | 0 | 1 | 2 |
| 12 | 11 | 24 | 700 | 1 | 0 | 1 | 2 |
| 13 | 10 | 20 | 600 | 3 | 1 | 0 | 3 |
| 14 | 10 | 16 | 1100 | 1 | 0 | 1 | 2 |
| 15 | 10 | 12 | 2900 | 4 | 0 | 3 | 2 |
| 16 | 10 | 8 | 1200 | 6 | 1 | 0 | 3 |
| 17 | 10 | 4 | 4200 | 3 | 0 | 0 | 2 |
| 18 | 10 | 24 | 4800 | 2 | 0 | 3 | 2 |
| 19 | 9 | 20 | 5400 | 4 | 0 | 3 | 4 |
| 20 | 9 | 16 | 600 | 1 | 1 | 1 | 11 |
| 21 | 9 | 12 | 2600 | 2 | 0 | 3 | 2 |
| 22 | 9 | 8 | 1300 | 1 | 0 | 1 | 2 |
| 23 | 9 | 4 | 1800 | 1 | 1 | 1 | 10 |
| 24 | 9 | 24 | 4400 | 1 | 1 | 1 | 3 |
| 25 | 8 | 20 | 1900 | 1 | 1 | 1 | 3 |
| 26 | 8 | 16 | 600 | 1 | 1 | 1 | 3 |

Figura 3.29: Dati selezionati e convertiti, relativi ai crimini a Baltimore nell'anno 2017

Ciò che si può notare dal dataset che vi è una prevelanza di crimini avvenuti per strada ( street=3) ,come arma del crimine più frequente sono state utilizzate le mani (Hands=1). I crimini più frequenti sono stati gli Assalti (common Assault=1) seguiti da furti (Robbery=2) e agressioni ( aggression = 3). In più è interessante notare che sono avventi più in luoghi interni (In =1) , tuttavia poca è la differenza tra i crimini successi all'esterno(Out=0). Date queste premesse, procediamo con l'analisi per confermare e prevedere il fenomeno criminologico a Baltimore.
Per comodità trasformiamo la tabella in un dataframe:

```
> datiAnalisi <- data.frame(analisi)
```

Disponiamo di 7 variabili e 72 osservazioni.Vogliamo indagare come la descrizione del crimine dipenda o meno dal fattore del luogo e dello strumento utilizzato. Le colonne che ci interessano, per ora, sono datiAnalisi$Description, datiAnalisi$Weapon, datiAnalisi$Premise. Cerchiamo un modello di regressione multipla a 3 predittori. La tecnica per stimare i regressori è detta ordinary least square OLS, che ripercorre gli stessi principi della tecnica dei minimi quadrati della regressione semplice. In R abbiamo:

```
> model <- lm(datiAnalisi$Description ~ datiAnalisi$Weapon +
datiAnalisi$Premise)
> summary(model)

Call:
lm(formula = datiAnalisi$Description ~ datiAnalisi$Weapon +
datiAnalisi$Premise)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1847 -1.1606 -0.6902  0.8945  5.1242

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.92394    0.43002   4.474 2.95e-05 ***
datiAnalisi$Weapon   0.26076    0.19025   1.371    0.175
datiAnalisi$Premise -0.00802    0.07479  -0.107    0.915
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.61 on 69 degrees of freedom
Multiple R-squared:  0.02791,Adjusted R-squared:  -0.0002664
F-statistic: 0.9905 on 2 and 69 DF,  p-value: 0.3766
```

Dalla diagnostica soprastante otteniamo il valore dell'intercetta e dei regressori:

- $\beta_0 = 1.92394$
- $\beta_1 = 0.26076$
- $\beta_2 = -0.00802$

$Y = 0.59655 - 1.40176 \cdot X_1 + 0.00303 \cdot X_2$.
Pertanto il modello di regressione multipla stimato è :

$Y = 1.92394 + 0.26076 \cdot X_1 - 0.00802 \cdot X_2$

Osserviamo il segno negativo del predittore relatvo a Premise, in questo contesto possiamo dedurre che più aumenta l'accadere del crimine in un determinato luogo, diminuisce la probabilità che il crimine sia della stessa tipologia . Il segno del predittore relativo a Weapon invece ha un effetto positivo , ciò implica che l'utilizzo frequente di un certo strumento, porti a aumentare la probabilità che quella determinata tipologia d crimine si verifichi con lo strumento usato più frequentemente.

Per rappresentare graficamente il modello creato installiamo il pacchetto scatterplot3d in r e speciafichiamo le variabili nel seguente ordine (X1, X2, Y):

```
> library(scatterplot3d)
> sc <- scatterplot3d(datiAnalisi$Weapon, datiAnalisi$Premise,
datiAnalisi$Description,  pch=16)
```

Otteniamo il seguente grafico:

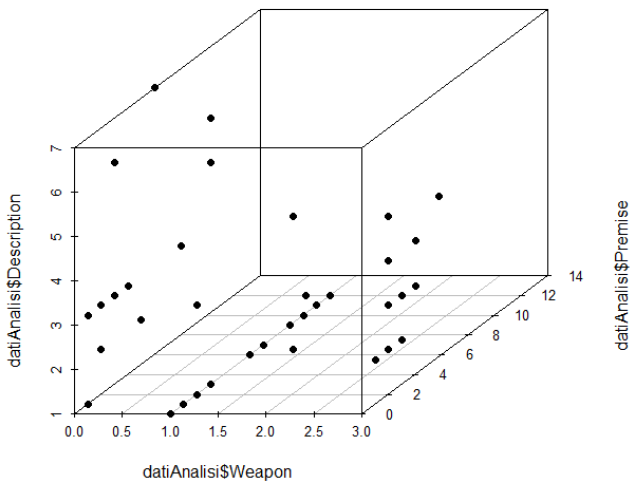



Figura 3.30: Grafico tridimensionale di rappresentazione delle variabili funzionali alla regressione multivariata

Possiamo anche intersecare il piano che abbiamo trovato, in questo modo:

```
> sc$plane(model, col="red")
```

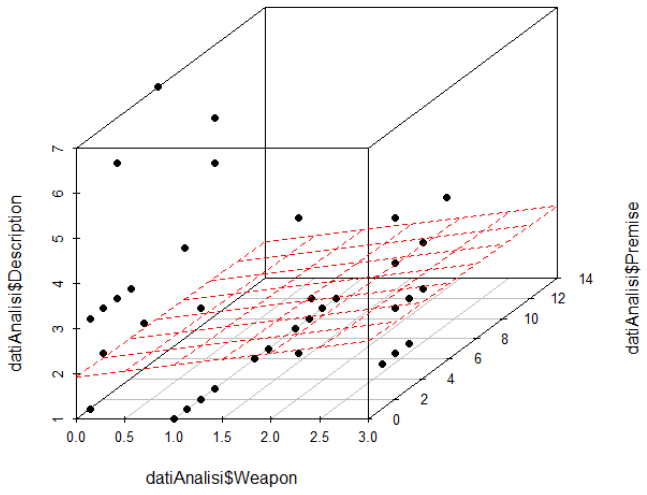



Figura 3.31: Grafico tridimensionale della regressione multivariata con intersezione del piano

Analizziano con attenzione le informazioni ottenute dalla diagnostica di regressione. Il coefficiente di determinazione ($R^2$ consente di valutare la bontà del modello, ossia la proporzione di variabilità della Y spiegata dalle variabili esplicative considerate. In questo caso R-squared è pari a 1.61 (con 69 gradi li libertà). Il valore R-squared adjusted è una correzione del coefficiente di determinazione che tiene conto del numero di predittori utilizzati (secondo alcuni statistici è più preciso quest'ultimo valore per definire la bontà del modello ). Inoltre è opportuno effettuare una verifica dell'intercetta e dei regressori, ossia un t-test con ipotesi nulla H0: $\beta = 0$. Ossia si deve verificare se il valore campionario possa essere esteso all'intera popolazione. Nel nostro caso i p-value associati alle variabili sono:

- p-value $\beta_0$= 2.95e-05
- p-value $\beta_1$= 0.175
- p-value $\beta_2$ = 0.915

Tra i tre valori risulta significativo il p-value di $\beta_0$ poichè risulta essere inferiore del valore ottimale di 0.05.
La statistica F ci indica se il modello è da scartare nella sua interezza, oppure se può essere ritenuto valido. Il valore di F-statistic è pari a = 0.9905 ed è maggiore dell'F-tabulato(cioè p-value: 0.3766),pertanto rifiutiamo l'ipotesi

nulla che il modello sia da scartare nella sua interezza.
Possiamo calcolare in questo modo gli intervalli di confidenza dell'intercetta e dei 2 regressori:

```
> confint(model)
                           2.5 %    97.5 %
(Intercept)           1.0660729 2.7817985
datiAnalisi$Weapon   -0.1187715 0.6403004
datiAnalisi$Premise -0.1572201 0.1411806
```

Se vogliamo invece calcolare l'intervallo di confidenza di un valore predetto (ossia se vogliamo prevedere il valore che assumerà Y, dati i predittori), utilizziamo la funzione predict():

```
> model <- lm(Description~ Weapon + Premise, data=datiAnalisi)
>
>
> predict(model, data.frame(Weapon=0.280, Premise=30),
interval="confidence")
       fit       lwr      upr
1 1.756357   -2.214489     5.727
```

Se volessimo considerare anche la variabile anche la variabile InOut , conduciamo l'analisi nella medesima modalità ma ora con 3 predittori:

```
> model <- lm(Description~ Weapon + Premise + InOut, data=datiAnalisi)
>
>
> summary(model)

Call:
lm(formula = Description ~ Weapon + Premise + InOut, data = datiAnalisi)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2353 -1.1346 -0.7155  0.9249  5.1140

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.86388    0.51433   3.624 0.000555 ***
Weapon       0.27822    0.20787   1.338 0.185208
Premise     -0.01184    0.07735  -0.153 0.878810
InOut        0.09317    0.43056   0.216 0.829335
```

```
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.621 on 68 degrees of freedom
Multiple R-squared:  0.02858,Adjusted R-squared:  -0.01428
F-statistic: 0.6668 on 3 and 68 DF,  p-value: 0.5753
```

l regressore $\beta_3$ associato al predittore InOut risulta essere non significato(p-value $<0.05$), e di segno positivo. É possibile dedurre come in precedenza che la tipologia del crimine sia determinata dal luogo esterno/interno in cui il fenomeno accade. Il modello a 3 predittori viene complessivamente accettato (la statistica F risulta significativa pari a 0.6668 ), e si assiste ad un miglioramento della goodness-of-fit (è aumentato il valore dell'R quadro aggiustato).

Sarebbe interessante a questo punto verificare quale delle tre variabili considerate maggiormente influisce sulla variabile Description. Per far questo possiamo confrontare i valori dei regressori $\beta$, ma solo dopo averli standardizzati (altrimenti non è possibile il confronto). In R adoperiamo il pacchetto QuantPsyc al cui interno è presente la funzione lm.beta():

```
> library(QuantPsyc)

> lm.beta(model)
     Weapon     Premise       InOut
 0.17563247 -0.01901118  0.02912957
```

Osserviamo che la variabile esplicativa che maggiormente influisce sulla descrizione del crimine è Weapon, ovvero lo strumento che viene utilizzato per commetterlo.

### 3.3.1 Risultati in sintesi

Con l'analisi di regressione multivariata abbiamo considerato un dataset che si discosta dal territorio nazionale italiano data la difficoltà di reperire dataset complessi. Tuttavia con il dataset recuperato, relativo a una zona territoriale colpita costantemente da fenomeni criminologi abbiamo potuto condurre un'analisi e trarne delle conclusioni. L'analisi di regressione multipla in questo campo ha permesso di verificare come il commettere una determinata tipologia di crimine sia influenza da fattori che abbiamo considerato come variabili, predittori. La variabile di maggiore influenza in questo contesto risulta essere lo strumento. Al contrario la località dell'accaduto non porta a standardizzare un tipo di crimine rispetto ad un altro.

# Capitolo 4

# Conclusioni

L'obiettivo di questa tesi è stato in primo luogo approfondire il campo del machine learning , determinando associazioni interessanti come nel caso della relazione con la statistica. Questo legame ha permesso l'entrata all'interno del mondo della crimonologia, dove grandi moli di dati sono all'ordine del giorno. L'applicazione del machine learning in una realtà apparentemente lontana ha permesso di ampliare ancora il panorama di estendibilità di tale disciplina. La tesi propone inizialmente un excursus dettagliato della scienza dell'apprendimento automatico, andando a coglierne le diverse procedure, tipologie e applicazioni. Solo dopo aver studiato l'esatto funzionamento di queste metodologie, con particolare attenzione all'apprendimento supervisionato è stato proposto prima un caso di studio , poi analisi specifiche, che applicano una metodologia propria di questa magrocategoria del machine learning. La metodologia di cui si fa riferimento è la regressione. I dati che vengono presi in considerazione nel caso di studio proposto sono relativi ad aggressioni legate alla violenza domestica , furto con scasso e furto di veicoli a motore, nello stato del New South Wales in Australia.Questa analisi viene condotta con il metodo della regressione bayesiana, da qui i risultati visibili , affermano il potere predittivo del metodo.Ciò che tale studio offre è quello di forninre una metodologia quantitativa basata sull'evidenza che mette in relazione il crimine con le informazioni ambientali e demografiche utilizzando algoritmi di apprendimento automatico nella gestioni dei dati in questione. In sintesi è stato possibile proporre un modello completamente probabilistico, in grado di quantificare l'incertezza nelle previsioni. Sempre la regressione a fare da protagonista di questa tesi: l'ultimo capitolo,difatti, propone delle analisi su dataset recuperati autonomamente. Qui l'analisi si dirama in regressione univariata e multivariata. Tali analisi permettono di trarre dei risultati e affermare la significatività del modello che si utilizza. La scelta di proporre delle analisi condotte con gli strumenti di statistica

posseduti, vuole mettere in evidenza l'applicabilità del modello teoricamente descritto e la possibilità di inferenza su dati in campo criminologico.

# Ringraziamenti

Siamo arrivati al momento dei ringraziamenti, un momento importante e pieno di emozioni in cui voglio aprire il mio cuore e dedicare delle parole alle persone che hanno fatto parte di questo meraviglioso percorso.Ringrazio immensamente mia madre e mio padre, che da quando ero piccola mi hanno sempre insegnato a seguire il cuore, a dar valore alle piccole cose ma non meno importanti, a tenere duro e affrontare sempre a testa alta ogni situazione. La mia laurea è il coronamento degli insegnamenti che ho ricevuto da voi e a voi dedicata. Allo stesso modo ringrazio i miei fratelli, per avermi sempre supportato, fatto arrabbiare ma sempre presenti in ogni situazione. Un grazie speciale a Kety, un'amica che mi ha visto crescere, sono otto anni che ti conosco e non c'è stato mai un litigio tra noi, sei stata un punto di riferimento per me, ricordandomi sempre di essere me stessa e non mollare mai. Sono qui,non ho mollato. Sinceri ringraziamenti anche a Priscilla e Alba, amiche presenti quotidianamente nella mia vita.Vi ringrazio perchè siete state capaci di capirmi e di sostenermi nei momenti difficili .Se ho avuto il coraggio di mettermi in gioco e di capire che, in fondo, gli ostacoli esistono per essere superati, il merito è soprattutto vostro, delle lunghe chiaccherate e discorsi motivazionali. Di persone vere nella vita ne ho conosciute poche e voi lo siete con la V maiuscola. Un posto importante nel conseguimento di questo risultato lo hanno avuto anche tutti i miei amici e colleghi . Grazie per aver condiviso con me questo percorso. Vi ringrazio per i pomeriggi a studiare insieme, a organizzare la sessione che si avvicinava: vi ringrazio per avermi fatto sentire a casa in una città per me nuova. Perchè è vero a volte casa è dove si sta bene. Vi voglio bene. A proposito di distanza , vorrei ringraziare la mia nuova coinquilina, che da quando è arrivata ha portato tanta positività, ti ringrazio per non avermi fatto mancare mai nulla, dolci compresi. Infine vorrei porre un sincero ringraziamento alla professoressa Elena Loli Piccolomini, relatrice di questa tesi di laurea, non solo per il supporto che mi ha fornito per la stesura di questa tesi, ma anche per la fiducia nell'applicare le conoscenze da lei trasmessemi. La ringrazio per i suoi consigli, per la sua disponibilità e pazienza durante il periodo di stesura. Grazie a lei ho avuto modo di acquisire ulteriori nozioni e conoscenze che mi saranno utili nella vita e nel lavoro.

# Bibliografia


[1] *Publications of the Astronomical Society of the Pacific*, 125.

[2] Artificial Intelligence, Machine Learning e Deep Learning: la storia e le differenze | Economyup.

[3] Deep Learning (apprendimento approfondito) - Cos'è e come funziona?

[4] L'utilità della mappatura degli hotspot per prevedere modelli spaziali di criminalità. *Security journal*, 21.

[5] Machine learning (apprendimento automatico) - Cos'è e come funziona?

[6] Machine Learning vs Statistics.

[7] Mapping crime: Understanding hotspots.

[8] *Predictive policing: The role of crime forecasting in law enforcement operations.*

[9] Previsione degli incidenti criminali utilizzando un modello di densità basato su pattern di punti. *International journal of forecasting*, 19.

[10] *Statistica e calcolo con R.*

[11] Statistical Methods for Machine Learning.

[12] Machine Learning Versus Statistics: When to use each, August 2017.

[13] Machine Learning & Deep Learning, November 2018.

[14] Algoritmi anti-crimine: tutte le tecnologie in campo, October 2019.

[15] Machine Learning: Tecniche e algoritmi predittivi avanzati, December 2019.

[16] KJ Bowers and SD Johnson. The handbook of security. 2014.

[17] L. Breiman. Statistical modeling: The two cultures. *Quality Engineering*, 48:81–82, 2001.

[18] Jason Brownlee. What is the Difference Between Test and Validation Datasets?, July 2017.

[19] Domenico Dodaro. Alan turing: uno spirito transumanista.

[20] John E Eck, Spencer Chainey, James G Cameron, Michael Leitner, and Ronald E Wilson. Mapping Crime: Understanding Hot Spots. page 79.

[21] Maurizio Di Paolo Emilio. Intelligenza artificiale, deep learning e machine learning: quali sono le differenze?, February 2018.

[22] Jacopo Franchi. L'Algoritmo Definitivo e il futuro del machine learning, May 2016.

[23] Matthew Garvin. *An Introduction to Statistical Learning Springer Texts in Statistics An Introduction to Statistical Learning*.

[24] Jacopo Kahl. I tre principali tipi di Machine Learning, February 2020.

[25] Brett Lantz. *Machine learning with R: learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Packt Publ, Birmingham, 2013. OCLC: 891534550.

[26] Di Il 29 Gennaio 2018 | ˜ 3 Minuto Letto. Machine Learning Vs. Statistical Learning, January 2018.

[27] Roman Marchant, Sebastian Haan, Garner Clancey, and Sally Cripps. Applying machine learning to criminology: semi-parametric spatial-demographic Bayesian regression. *Security Informatics*, 7(1):1, June 2018.

[28] Roman Marchant, Sebastian Haan, Garner Clancey, and Sally Cripps. Applying machine learning to criminology: semi-parametric spatial-demographic bayesian regression. *Security Informatics*, 7, 12 2018.

[29] Tom Mitchell. Introduction to machine learning. *Machine Learning*, 7:2–5, 1997.

[30] Maria Giuseppina Muratore. La misurazione del fenomeno della criminalità attraverso le indagini di vittimizzazione. page 12.

[31] Amy E Nivette. Cross-national predictors of crime: A meta-analysis. *Homicide Studies*, 15(2):103–131, 2011.

[32] Pierre Perruchet and Sebastien Pacton. Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, 10(5):233–238, May 2006.

[33] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[34] Matthew Stewart Researcher, PhD. The Actual Difference Between Statistics and Machine Learning, July 2020.

[35] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959.

[36] SecurityOpenLab. il machine learning può prevenire il crimine?

[37] Tarang Shah. About Train, Validation and Test Sets in Machine Learning, July 2020.

[38] Expert System Team. What is Machine Learning? A definition, May 2020.

[39] Jerome Friedman Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Number 2nd ed. Springer, 2009.

[40] Alan M Turing. Si può dire che i calcolatori automatici pensano? *Sistemi intelligenti*, 10(1):27–40, 1998.

[41] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[42] LC Yan, B Yoshua, and H Geoffrey. Deep learning. *nature*, 521(7553):436–444, 2015.