UNIVERSITÀ DEGLI STUDI DI TRENTO

Dipartimento di Ingegneria e Scienza dell'Informazione

Corso di Laurea in
Informatica

Elaborato finale

robustbasePy : Package in Pyhton per l'analisi dei dati tramite metodi robusti

Supervisore

Claudio Agostinelli

Laureando

Alberto Bellumat

Anno accademico 2018/2019

Ringraziamenti

Questo elaborato finale rappresenta il culmine della mia carriera accademica all'università di Trento, che mi ha sempre fornito nuove ed intriganti esperienze e sfide durante questi tre anni. Ringrazio il professore Claudio Agostinelli, relatore di questa tesi, per la disponibilità e la cortesia dimostratami durante il periodo di stesura della tesi. Ringrazio i miei compagni di studio che nell'arco di questi tre anni mi hanno aiutato nei i progetti di classe e alle lezioni. Ringrazio anche i miei amici di Milano che hanno reso la mia permanenza a Trento piacevole. Un sentito ringraziamento a mio padre e a mia madre che con il loro supporto, sia morale che economico, sono riuscito a svolgere un percorso arduo e impegnativo fino ad oggi.

Indice

Sommario

Molto spesso i ricercatori svolgono studi su insiemi di dati eterogenei, e non ci si stupisce se all'interno di quest'ultimi siano presenti dati distinti dagli altri (ma non necessariamente inattendibili). Da questi insiemi di dati, se applicati i metodi classici della statistica, si corre il rischio di produrre modelli molto semplificati e la loro validità è, nei casi migliori, approssimativa.
Infatti, le procedure classiche della statistica sono molto sensibili alle deviazioni dal modello ipotizzato, ed in particolare alla presenza degli outliers (valori anomali). Per far fronte a questo problema, la soluzione più semplice, ma poco ortodossa, consiste nell'individuare e sostituire o eliminare ogni outliers. Questo approccio è inefficiente in presenza di elevate quantità di dati, e questo fu uno dei molteplici motivi per cui nacque la statistica robusta.
Lo sviluppo del pacchetto software `robustbasePy`, argomento di questa tesi, è stato voluto a causa della scarsa presenza di codice in Python per la computazione di metodi robusti. `robustbasePy`è stato sviluppato in Python, linguaggio che sta riscuotendo successo nell'ambiente scientifico negli ultimi anni, ma all'interno del codice sono presenti anche chiamate a funzioni scritte in C. Durante la fase di sviluppo sono state utilizzate librerie open source di Python come ad esempio sciPy, che fornisce un ecosistema di tool per la matematica, le scienze e l'ingegneria. Il pacchetto è stato scritto dal laureando, fatta ad eccezione di alcune funzioni C, già presenti in altri progetti open source. Gli unici aiuti sono stati chiesti al professore Claudio Agostinelli durante la stesura dell'elaborato finale.
Il lavoro svolto può essere diviso in tre parti:

- **Studio dei requisiti:** Inizialmente, si è cercato di capire ed analizzare i requisiti richiesti dal professore Claudio Agostinelli, oltre allo studio dei concetti base della statistica robusta.
- **Implementazione e documentazione:** Prendendo come punto di partenza le funzioni e algoritmi già presenti e definiti nel pacchetto `robustbase` in R, è iniziata lo sviluppo del package `robustbasePy`. è stato scelto di interfacciare il codice Python ad alcune funzioni C, già presenti nel pacchetto `robustbase`. Una volta completato il pacchetto software, si è passato alla stesura della documentazione, dove vengono le descritte dettagliatamente i servizi offerti da `robustbasePy`.
- **Simulazione:** Infine, si è testato il pacchetto software su un dataset. La simulazione è presente e descritta a pagina 14 dell'elaborato finale.

1 La statistica robusta

1.1 Introduzione alla statistica robusta

I metodi statistici classici sono fondamentalmente ipersensibili ai requisiti originali richiesti per l'utilizzo del modello statistico adottato per il trattamento di dati sperimentali. Quando si ha a che fare con problemi applicati le deviazioni dai requisiti iniziali del modello sono inevitabili, quindi l'uso di metodi standard in tali condizioni potrebbe rivelarsi inefficace. Inoltre spesso portano a distorsioni significative delle inferenze statistiche. Da ciò emerge la necessità di costruire nuovi metodi di elaborazione delle informazioni che siano robusti rispetto a possibili deviazioni delle caratteristiche dei dati reali.
La robustezza è una proprietà di un metodo statistico che caratterizza l'indipendenza dei risultati dall'influenza di vari outliers o spikes ("valori anomali" o "picchi"), è la resistenza all'interferenza.

Un metodo resistente agli outliers (definito anche come metodo robusto) è un metodo in grado di identificare i suddetti outliers, ridurre il loro impatto o escluderli dal campione.
Di fatto, la presenza anche solo di poche osservazioni anomale (outlier) nei campioni può influenzare notevolmente il risultato di uno studio. Gli stimatori classici, come per esempio il metodo dei minimi quadrati e il metodo della massima verosimiglianza, sono soggetti a questo genere di distorsione su distribuzioni specifiche e spesso si comportano molto male [**?**], quindi l'interpretazione di statistiche calcolate usando insiemi di dati contenenti outlier sarà probabilmente fuorviante. Per eliminare l'effetto di tale interferenza si usano vari approcci per ridurre o eliminare l'effetto delle osservazioni inaccurate (outlier).
Il compito principale dei metodi resistenti agli outlier è distinguere le osservazioni inaccurate da quelle accurate. Qui anche gli approcci più semplici - come quello soggettivo (basato sull'intuito del ricercatore) - possono portare a miglioramenti significativi, ma per motivare il rifiuto di determinati valori gli statistici utilizzano metodi basati su giustificazioni matematiche rigorose. Si tratta di una procedura non banale che definisce una delle aree della scienza statistica.
La statistica robusta nasce come risultato di questo desiderio di colmare il vuoto tra i due estremi relativi agli assunti aprioristici sulla natura dei dati da elaborare, per estrarre nella maniera più efficientemente possibile informazioni utili dai dati sperimentali. Il problema è che i dati vanno elaborati per assicurare un'elevata qualità dei risultati. Cercare di analizzare i dati usando metodi non robusti potrebbe portare a conclusioni parziali. Svolgendo studi basati su un insieme di dati estremamente eterogeneo i ricercatori necessitano di metodi statistici che corrispondano a modelli affidabili e identifichino tendenze informative, concentrandosi sul sotto-campione omogeneo dominante senza permettere sottogruppi strutturalmente diversi di distorcere i risultati. La statistica robusta è una soluzione a questo problema. [**?**]
Sostanzialmente questo termine (statistica robusta) descrive qualsiasi metodo statistico che si comporti bene con insiemi di dati presi da una vasta gamma di distribuzioni di probabilità e che risulti poco influenzato da picchi nei dati (outlier) o da piccole deviazioni rispetto agli assunti del modello per quell'insieme di dati. Il nome di statistica robusta è quello generalmente accettato per il settore della statistica che sviluppa procedure statistiche insensibili agli errori nei risultati. [**?**]
La teoria della statistica robusta introduce, oltre al concetto di modello statistico di base (ideale), un'ulteriore entità chiamata supermodello. Il concetto di supermodello serve a descrivere un modello di base in cui crediamo e che usiamo come base e per descrivere le potenziali deviazioni da questo modello di assunti di base in condizioni sperimentali reali. Lo scopo principale è sviluppare procedure statistiche che funzionino sufficientemente bene sia con il modello ideale che all'interno del supermodello.

1.1.1 Outliers

Gli *outliers* sono punti di osservazione chiaramente distinti dal resto dei dati del campione. Ciò non significa necessariamente che siano dati inattendibili. Spesso contengono i valori minimi o massimi del campione, o entrambi, e potrebbero anche essere dati reali, ma comunque vanno sempre controllati per evitare possibili errori. Queste fluttuazioni anomale causano problemi di vario tipo per i modelli statistici standard, tanto da portare allo sviluppo di metodi statistici meno sensibili - e quindi più robusti relativamente agli outlier. [**?**]
Mostriamo un esempio facendo un confronto tra i due metodi comuni utilizzati per stimare il centro di un insieme di dati: la media del campione e la mediana del campione. Per le definizioni di media del campione e mediana del campione si fa riferimento a [Basic Statistics], pagina 149 e pagina 180.
La media del campione può risultare completamente stravolta se il campione contiene anche un solo outlier; se nel campione il valore di un dato qualsiasi tende all'infinito, allora inevitabilmente, anche la media campionaria tende all'infinito. Questa proprietà è in netto contrasto con il comportamento della mediana del campione, che è solo scarsamente influenzata dal cambiamento di un singolo valore tendente all'infinito. Quindi la mediana non verrà influenzata, almeno non significativamente, dalla presenza di outliers. Si può affermare che la mediana è *robusta* rispetto agli *outliers* ma la media non lo è. Per la distribuzione normale la media è considerata lo stimatore ottimale del punto, ma a causa delle sue proprietà può fornire stime decisamente peggiori per distribuzioni vicine alla normale.
Per stimare la "tendenza centrale" di un campione con outliers lo statistico necessita di una misura

migliore, come ad esempio la mediana del campione. Ed è proprio questo obiettivo che i metodi robusti mirano a fornire.
Lavorando con dati sperimentali (misure) un ricercatore può controllare e raffinare l'insieme di dati (per esempio rimuovendo gli outlier). Ma ciò non sempre avviene per vari motivi: [**?**].

1. Sebbene esistano metodi semplici di analisi visiva ed esplorativa dei dati che possono aiutare a rendere più chiaro la presenza degli outliers, nella pratica però persino gli statistici più esperti spesso saltano lo screening dei dati.

2. Di fronte a insiemi di dati altamente eterogenei un approccio estremo (scegliere se tenere o scartare un'osservazione) può rivelarsi inefficiente. Un approccio più morbido rispetto al rifiutare totalmente le osservazioni anomale potrebbe essere dare loro un peso minore.

3. Quando i dati sono eterogenei spesso è molto difficile individuare gli outlier.

4. Spesso non è chiaro a quali ignoti fattori siano da attribuire i rilevamenti di outliers. Talvolta non è pratico determinare se gli outliers sono dati di cattiva qualità. Nel campo degli esperimenti gli outliers possono essere dovuti a fluttuazioni casuali, ma possono anche indicare fenomeni scientificamente interessanti. In ogni caso scartare gli errori gravi modificherà la teoria della distribuzione, che necessita quindi di aggiustamenti. Alcune caratteristica della distribuzione, come la varianza, verranno sottostimate se calcolate usando dati raffinati.

1.2 Misurare la robustezza

L'approccio quantitativo per la determinazione della robustezza degli stimatori si basa - come quello qualitativo - sul richiedere che a cambiamenti arbitrariamente piccoli nella distribuzione delle osservazioni corrispondano solo cambiamenti sufficientemente piccoli nelle caratteristiche di qualità degli stimatori. Per chiarire questo requisito, specifichiamo il criterio di qualità della procedura e imponiamo determinate restrizioni al suo comportamento all'interno del modello adottato per descrivere possibili cambiamenti nella distribuzione delle osservazioni.
Per studiare la robustezza di uno stimatore si possono usare due misure semplici: la **funzione di influenza** e il **punto di rottura**. L'idea di base di questo approccio è indagare il comportamento di una funzione stimatrice influenzata da outliers o da qualsiasi dato, anche arbitrario. La **sensibilità** (in inglese detta "gross error sensitivity") che misura l'effetto massimale di un'osservazione isolata sul valore stimato si basa sulla funzione di influenza. Si può visualizzare ciò che fa la sensibilità come il misurare la stabilità locale della procedura di stima. [**?**].
Il punto di rottura misura l'affidabilità globale o la sicurezza della procedura di stima determinando la più piccola parte di osservazioni anomale nel campione studiato che basta per ottenere stime non realistiche. Il punto di rottura della mediana campionaria è 50%, ovvero può tollerare fino al 50% di outliers nel campione prima di diventare grande; mentre il punto di rottura della media campionaria è 0%.

1.2.1 Modello di posizione

Il modello di regressione lineare più *semplice* (il **modello di posizione**) illustra ciò che viene misurato con due valori, la funzione di influenza e il punto di rottura:

$$Y_i = \beta_0 + E_i, \quad i = 1, \ldots, n.$$

Arthur Cushny e Alvin Peebles pubblicarono nel 1905 uno studio sull'efficacia di due farmaci, spesso citato nella statistica. L'aumento medio della durata del sonno per i due farmaci A e B fu misurato per 10 volontari. I risultati (in ore) furono [**?**]

$$1,2 \quad 2,4 \quad 1,3 \quad 1,3 \quad 0,0 \quad 1,0 \quad 1,8 \quad 0,8 \quad 4,6 \quad 1,4$$

Per molto tempo questi dati furono considerati un tipico esempio di dati distribuiti normalmente.
Lo scopo è stimare di quanto aumenti in media la durata del sonno. Per evitare una notazione non

necessaria useremo β invece di β_0. Visto che le scelte più ovvie come stimatori sono media e mediana, le useremo per stimare β: la **media aritmetica**, $\bar{y} = 1,58$, e la **mediana**, med $= 1,3$, dei dati sopra elencati. Un'altra misura della tendenza centrale può essere una media **troncata**, o una **media sfrondata al 10%**, ovvero una stima dove il 10% dei valori più grandi e il 10% dei valori più piccoli vengono esclusi per poi calcolare la media aritmetica da quelli restanti: così otteniamo $\bar{y}_{10\%} = 1,4$. Si può ottenere un effetto simile applicando una **regola di rifiuto** (ovvero un criterio per decidere quali outlier accettare o scartare) per poi calcolare una media aritmetica. I valori identificati come outlier verranno così esclusi e si calcolerà nuovamente la media aritmetica da quelli rimanenti. La regola di rifiuto è $|y_i - \bar{y}|/s > 2,18$, dove s è la deviazione standard.

Per i dati della tabella di sopra la media è $\bar{y}^* = 1,24$ applicando la regola di rifiuto.

Si può notare che gli ultimi tre stimatori riportano un aumento medio della durata del sonno decisamente inferiore.

1.2.2 Funzione di influenza empirica

Relativamente all'esempio storico di sopra, le differenze sono partite dall'osservazione che $y = 4,6$ ore. Ora è il momento di studiare come lo stimatore $\hat{\beta}(y_1, \ldots, y_n)$ cambia se cambiamo il valore y che stava causando fluttuazione. La quantità usata è la **funzione di influenza empirica** (o curva di sensibilità) [**?**]

$$SC(y; y_1, \ldots, y_{n-1}, \beta) \stackrel{\text{def}}{=} \frac{\hat{\beta}(y_1, \ldots, y_{n-1}, y) - \hat{\beta}(y_1, \ldots, y_{n-1})}{1/n}$$

(Il denominatore rende questa funzione indipendente dalla dimensione del campione). In questo esempio la funzione di influenza empirica viene concepita coma una funzione dell'osservazione aggiuntiva y. Ma la funzione di influenza empirica è influenzata anche dal campione.

La funzione di influenza empirica per i quattro stimatori dati sopra è illustrata in figura1.1 e mostra l'effetto osservato se invece dell'outlier di 4,6 ore venisse osservato un altro valore y.

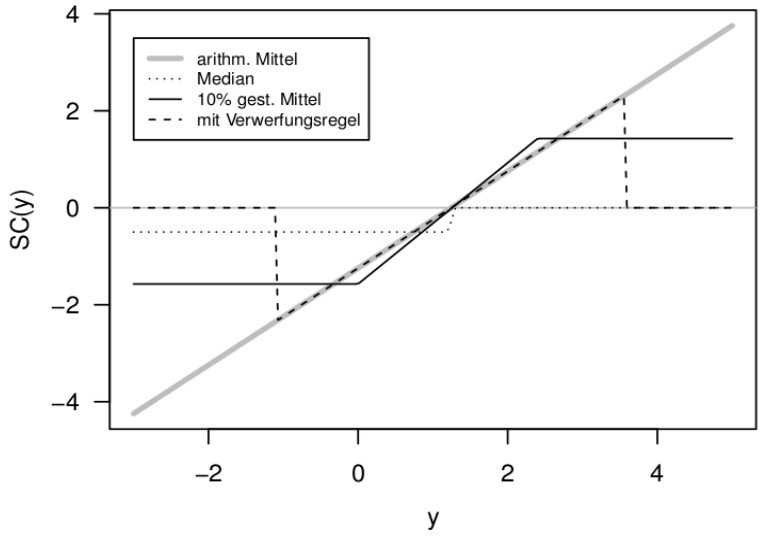

Figura 1.1: Funzione di influenza empirica per l'esempio con i dati sul sonno. L'outlier $y = 4,6$ ore varia entro $-3 < y < 5$. (preso da [**?**])

Un confronto tra le quattro stime dimostra che quella che reagisce in maniera più vistosa all'osservazione $y = 4,6$ è la media aritmetica; la mediana è la meno influenzata mentre la media sfrondata al 10% e la media aritmetica con regola di rifiuto sono nel mezzo. Nella figura di sopra la retta che rappresenta il comportamento della media aritmetica aumenta arbitrariamente considerando un outlier sempre più estremo (tendente all'infinito). Le curve di altre stime o si interrompono per valori dipendenti da problemi specifici o tornano a 0 (il che rappresenta la regola di rifiuto).

1.2.3 Funzione di influenza

La curva nel grafico precedente mostra l'effetto (ovvero l'influenza) di un'osservazione addizionale y per un dato campione. La dipendenza da un dato valore in $(y_1, \ldots, y_{n-1})$ non è il modo ideale per valutare la robustezza degli stimatori. Il prossimo passo sarà vedere cosa succede se consideriamo un campione tendente all'infinito. Un campione del genere è sostituibile con la sua distribuzione soggiacente $\mathcal{F}$ e l'osservazione aggiuntiva con un determinato punto materiale. La distribuzione usata viene generalmente assunta normale, e la funzione risultante è chiamata **funzione di influenza** IF. Per i nostri scopi va osservato che [**?**]

$$IF\left(y; \mathcal{F}, \hat{\beta}\right) \approx SC\left(y; y_1, \ldots, y_{n-1}, \hat{\beta}\right).$$

Sia **SC** la **curva di sensibilità**. La funzione di influenza è calcolabile per la maggior parte degli stimatori. Descrive l'effetto asintotico di una singola osservazione sul valore dello stimatore e permette di calcolare la dispersione asintotica.

1.2.4 Sensibilità

È intuitivamente chiaro che per uno stimatore che ha una funzione di influenza limitata IF (o SC), i valori anomali possono avere solo un'influenza limitata sul valore stimato (purché non ce ne siano troppi).
Pertanto, un valore importante per descrivere la robustezza di uno stimatore è la sensibilità dell'errore lordo

$$\gamma^* \stackrel{\text{def}}{=} \max\left(\text{over}y\right) \quad \text{di} \quad \left|IF\left(y; \hat{\beta}, \mathcal{F}\right)\right|$$

Di solito ci interessa solo la questione se la sensibilità sia limitata o meno. Questa domanda è facile da rispondere con la forma della funzione di influenza empirica (vedi Figura 1.1).
Secondo il criterio di sensibilità γ^*

- La media aritmetica $\bar{Y}$ non è un estimatore robusto per $\hat{\beta}$. Tuttavia
- la mediana, $\bar{Y}_{10\%}$ e $\bar{Y}_*$ sono estimatori robusti per il parametro di posizone β nel modello di posizione.

Fino ad ora abbiamo studiato solo l'influenza di una osservazione (insolita o estremista) sullo stimatore. Come reagisce lo stimatore a due valori anomali? Se consideriamo il caso $y_{n-1} = y_n \to \infty$, allora tiene

$$\begin{aligned} \bar{y} &\to \infty; & \text{med} &= 1.3 \ \text{(rimane costante)} \\ \bar{y}_{10\%} &\to \infty; & \bar{y}^* &\to \infty \end{aligned}$$

È importante notare che la regola di rifiuto non identifica alcuna osservazione come valori anomali. Cioè, questa procedura di stima non offre ciò che promette. Sebbene gli ultimi tre stimatori abbiano mostrato un comportamento molto simile rispetto alla funzione di influenza, in questa nuova situazione sono chiaramente distinti l'uno dall'altro. Tranne che per la mediana, si comportano come la media aritmetica e sono quindi meno robusti nei confronti di valori anomali rispetto alla mediana.
Per uno stimatore con funzione di influenza limitata IF l'influenza degli outlierx sul valore stimato sarà limitata. Ciò è vero finché gli outliers non superano la soglia di tolleranza. Ecco quindi una nuova quantità: la **sensibilità agli errori gravi**, un valore importante per descrivere la robustezza di uno stimatore [**?**]

$$\gamma^* \stackrel{\text{def}}{=} \max\left(\text{su}y\right) \quad \text{di} \quad \left|IF\left(y; \hat{\beta}, \mathcal{F}\right)\right|$$

(l'espressione matematica precisa per il massimo è l'estremo superiore). La forma della funzione di influenza empirica (Figura 1.1) aiuterà a capire se la sensibilità è limitata o no.
Seguendo il criterio di sensibilità γ^* si può concludere che nel caso di una osservazione inaccurata relativamente allo stimatore:

- la media aritmetica $\bar{y}$ non è uno stimatore robusto per il punto $\hat{\beta}$.

- per il parametro di posizione β nel modello di posizione della mediana, $\bar{y}_{10\%}$ e $\bar{y}_*$ sono stimatori robusti (resistenti agli outlier).

L'ipotesi seguente è che il campione contenga due outlier. Se consideriamo il caso $y_{n-1} = y_n \to \infty$, ne segue che

$$\begin{aligned} \bar{y} &\to \infty; & \text{med} &= 1,3 \ \ (\ \text{rimanecostante}) \\ \bar{y}_{10\%} &\to \infty; & \bar{y}^* &\to \infty \end{aligned}$$

Da ciò si può conclude, come detto in precedenza, che la regola di rifiuto non identifica alcuna osservazione come outlier. Inoltre, nonostante mostrino un comportamento simile relativamente alla funzione di influenza, questi stimatori si comportano diversamente in presenza di due outliers.

1.2.5 Punto di Rottura

La robustezza si può caratterizzare con una misura semplice ovvero il **punto di rottura** $\epsilon_n^*\left(\hat{\beta}; \underline{y}\right)$. Il punto di rottura è il rapporto massimo di osservazioni nel campione che non impediscono allo stimatore di fornire stime affidabili. Sia $\mathcal{X}_m$ l'insieme di tutti i dati reali $\underline{y}^* = \{y_1^*, \ldots, y_n^*\}$ di dimensione n che hanno $(n-m)$ elementi in comune con $\underline{y} = \{y_1, y_2, \ldots, y_n\}$. Quindi

$$\epsilon_n^*\left(\hat{\beta}; \underline{y}\right) = \frac{m^*}{n},$$

dove

$$m^* = \max\left\{m \geq 0 : \left|\hat{\beta}\left(\underline{y}^*\right) - \left(\underline{y}\right)\right| < \infty \ \text{ per tutti } \ \underline{y}^* \in \mathcal{X}_m\right\}$$

Se si modificano abbastanza osservazioni nessuno stimatore comune può fornire stime affidabili. Quindi lo stimatore è in condizione di *rottura* ("breakdown" in inglese). Per le procedura di stima comune il punto di rottura viene raggiunto quando $m > n/2$ (se non prima) e quindi il punto di rottura massimo $\epsilon_n^*\left(\hat{\beta}; \underline{y}\right)$ è minore di $1/2$ [**?**].
Si può quindi concludere che

- La sensibilità agli errori gravi $\gamma^*\left(\hat{\beta}, \mathcal{X}\right)$, misura l'effetto massimo di un piccolo disturbo nei dati.
- mentre il punto di rottura $\epsilon_n^*\left(\hat{\beta}; y_1, \ldots, y_n\right)$ misura la dimensione minima che porta al fallimento dello stimatore.

1.3 Regressione robusta

La **stima a scala** e la **stima di posizione** sono due metodi molto noti nella statistica robusta che si propongono di superare i procedimenti della statistica tradizionale in presenza di outliers. Con la stima di posizione la media sfrondata si comporta relativamente bene rispetto alla media, ma ci sono stime ancora più robuste. Nel caso di stima di posizione la stima usata solitamente è la deviazione standard. La presenza di outliers nel campione la rende inaffidabile ed eccessivamente ampia. La **regressione robusta** è un tipo di analisi di regressione che aiuta a superare alcune delle carenze dei metodi tradizionali, parametrici e non. [**?**].
La regressione robusta è un metodo di regressione usato quando sono presenti outlier che influenzano il modello o quando la distribuzione non è normale. È uno strumento importante per lavorare con dati influenzati dagli outlier e assicurarsi che il modello risultante sia robusto contro gli outliers [**?**].
Un metodo di stima usato con la regressione robusta è la **stima M**. Nell'ambito della stima robusta sono stati proposti anche stimatori R e L. Grazie ai loro punti di rottura generalmente elevati e alla loro efficienza gli *stimatori M* [**?**] sono i più usati nel campo della statistica robusta.

1.3.1 Stima M

Gli stimatori M hanno un ruolo fondamentale non solo nei modelli a locazione ma anche in quelli a regressione. [**?**]. Gli stimatori M sono stime ottenute dalle soluzioni di equazioni solitamente provenienti dalla ricerca di estremi. Questa classe include i metodi di massima verosimiglianza e dei minimi quadrati.

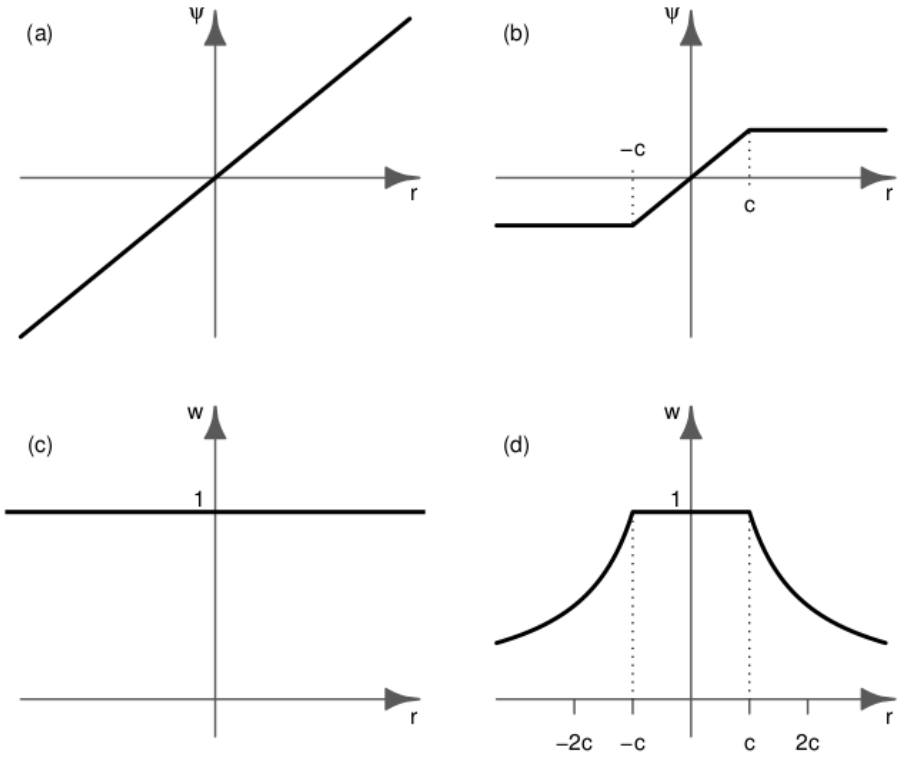

Figura 1.2: ψ- e funzione di peso per la stima ordinaria dei minimi quadrati (a sinistra), che non è robusta, e per un robusto stimatore M (a destra). La seconda funzione ψ è anche conosciuta come funzione ψ di Huber.

Lo stimatore M è un analogo della media ponderata e i pesi hanno lo scopo di impedire che gli outliers influenzino lo stimatore.

$$\hat{\beta}_M = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{1.1}$$

1.3.2 Funzione di influenza degli stimatori M

Per stabilire pesi (o funzioni di peso) appropriati vanno consultate le funzioni di influenza IF corrispondenti. La funzione di influenza di $\hat{\beta}_M$ è uguale a

$$IF\left(y, \hat{\beta}, \mathcal{N}\right) = const \ \cdot w \cdot \tilde{r} = const \cdot \psi\left(\tilde{r}\right),$$

dove $\tilde{r}\ (y-\beta)/\sigma$ e $\psi\left(\tilde{r}\right) \stackrel{\text{def}}{=} w \cdot \tilde{r}$ [**?**]. Una caratteristica importante e degna di nota degli stimatori M è la sua funzione di influenza $IF\left(y, \hat{\beta}, \mathcal{N}\right)$ è proporzionale alla funzione, il che stabilisce il tipo specifico di stimatore M. Molte caratteristiche numeriche degli stimatori sono stabilite dalla funzione di influenza. Di conseguenza questa caratteristica permette di costruire stimatori M con date proprietà di robustezza mediante opportuna scelta di funzione ψ. Anche il metodo dei minimi quadrati è uno stimatore, ma non è robusto. Le figure 1.2(a) e 1.2(c) mostrano che se $\psi(u) = u$, tutti i pesi saranno uguali a 1. Ciò significa che la media ponderata di sopra diventerà una media ordinaria. Il secondo corrisponde al classico stimatore funzione ψ limitata che ricorda lo stimatore M della funzione ψ di Huber mostrato nella figura 1.2(b). La funzione di peso corrispondente si può determinare mediante $w_i = \psi\left(\tilde{r}_i/\tilde{r}_i\right)$ (vedi figura 1.2(d)). Esistono altre funzioni ψ come: biquadrato di Tukey e Hampel che sono usate per il calcolo di stimatori M. È possibile vedere dettagli di queste funzioni in [**?**]

1.3.3 Stimatore M

Lo stimatore M è definito da un'equazione implicita,

$$\sum_{i=1}^{n} \psi\left(\frac{r_i\left(\hat{\beta}\right)}{\sigma}\right) = 0 \quad r_i\left(\hat{\beta}\right) = x_i - \beta,$$

, dove σ è il parametro di scala. L'equazione corrisponde alla normale equazione degli stimatori dei minimi quadrati. La funzione ψ svolge un ruolo importante nella descrizione delle proprietà degli stimatori M.

1.3.4 Calcolo degli stimatori M

I pesi nell'equazione 1.1 dipendono dal parametro ignoto β (e σ). Ciò rende impossibile calcolare esplicitamente la media ponderata. Questa rappresentazione degli stimatori M permette di costruire un semplice algoritmo iterativo per calcolare lo stimatore M.

1. La mediana è una stima iniziale, e stima β, e poi stima σ (vedere sotto).
2. i pesi w_i calcolati come $w_i = \psi\left(\bar{r}_i/\bar{r}_i\right)$
3. Il prossimo passo è calcolare una nuova stima di β usando l'equazione 1.1
4. Ripetere i passi 2 e 3 finché l'algoritmo non converge.

1.3.5 Stima del parametro di scala

I punti in cui i dati perdono influenza vanno determinati per fare una stima M applicabile a questo caso, tranne che per media aritmetica o mediana (stima L_1). Per quanto riguarda la funzione ψ di Huber, questo il punto in cui la funzione ha un punto critico. Solo considerando la scala dei residuali si può farlo in modo ragionevole. Quindi, nella definizione di stimatore M (equazione 1.1) divisione σ (parametro di scala) è già applicato, come nella regola di rifiuto:

$$\left|\left(x_i - \bar{X}\right)/S\right| > 2,18 \quad (n = 10).$$

Come detto in precedenza, la deviazione standard come stima di σ non è abbastanza resistente per questa analisi.
Il parametro di scala si può stimare con il seguente stimatore robusto:

$$s_{MAD} = \text{med}_i\left(\left|x_i - \text{med}_k\left(x_k\right)\right|\right)/0,6745$$

(**deviazione assoluta mediana**). Si sono ottenuti risultati consistenti con la "correzione" 1/0,6745 . Lo stimatore s_{MAD} si può ora usare nello stimatore M al posto di σ, anche se potrebbe essere noto da altre fonti.
In presenza di uno stimatore di scala adeguato, i punti di non differenziabilità della funzione ψ vanno comunque determinati esplicitamente mediante una cosiddetta costante di tuning. Di norma viene definita in modo che lo stimatore corrispondente abbia un'efficienza relativa del 95% per la distribuzione del modello. La costante di tuning per la ψ-funzione di Tuning è di $c = 1,345$, in base alle considerazioni precedenti.

2 La funzione `lmrob`

2.1 Introduzione

La regressione lineare è un approccio per modellare la relazione tra una risposta scalare o variabile dipendente Y e una o più variabili esplicative o indipendenti indicate con X.
Nella regressione lineare, i dati sono modellati utilizzando le funzioni di predittore lineare e i parametri

del modello sconosciuto sono stimati dai dati.
Un modello di regressione lineare che coinvolge una variabile indipendente può essere espresso come

$$\begin{array}{rcl} Y_i & = & X_i\beta + \epsilon_i, \\ \Rightarrow \quad \epsilon_i(\beta) & = & Y_i - X_i\beta, \end{array} \tag{2.1}$$

Y_i è la variabile di risposta sull'osservazione i-esima, β la stima dei coefficienti o dei parametri, X_i è il valore della variabile indipendente sull'osservazione i-esima, e ϵ_i è una variabile casuale distribuita normalmente. L'errore ϵ_i $N(0, \sigma^2)$ non è correlato reciprocamente.
Il metodo di regressione più comunemente utilizzato è il metodo dei minimi quadrati ordinari (OLS). La stima OLS è ottenuta come soluzione del problema

$$\hat{\beta} = arg\min_{\beta} \sum_{i=1}^{n} \epsilon_i^2(\beta) \tag{2.2}$$

Prendendo le derivate parziali rispetto a β e impostandole uguali a zero si ottengono le normali equazioni e si ottiene il modello di regressione stimato

$$\hat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i, \tag{2.3}$$

Le stime OLS per la regressione lineare sono ottimali quando tutte le ipotesi di regressione sono valide. Quando alcune di queste ipotesi non sono valide, la regressione dei minimi quadrati può avere un rendimento scarso.
Un'analisi di regressione robusta fornisce un'alternativa al modello di regressione dei minimi quadrati quando le ipotesi fondamentali non sono soddisfatte dalla natura dei dati. Quando l'analista valuta i suoi modelli di regressione statistica e mette alla prova le sue ipotesi, egli trova spesso che le ipotesi sono sostanzialmente violate. A volte l'analista può trasformare le sue variabili per conformarsi a tali ipotesi. Spesso, tuttavia, una trasformazione non eliminerà o attenuerà la leva dei valori anomali dell'influenza che distorcono la predizione e distorcono il significato delle stime dei parametri. In queste circostanze, una regressione robusta e resistente all'influsso di valori anomali (outliners) può essere l'unica risorsa ragionevole.
I metodi ben noti di stima robusta sono la stima M, la stima S e la stima MM. La stima M è un'estensione del metodo di massima verosimiglianza ed è una stima robusta, mentre la stima S e la stima MM sono lo sviluppo del metodo di stima M. Usando questi metodi è possibile eliminare alcuni dei dati, che in alcuni casi non potrebbero sempre essere fatti anche se tali dati sono importanti. In questo tesi viene presentata la funzione `lmrob` scritta in linguaggio python che offre i metodi di stima sopra menzionati per determinare il modello di regressione ottimale.

2.2 La teoria dietro la funzione `lmrob`

`lmrob` è una funzione programmata in python per determinare i modelli di regressione robusta e dovrebbe essere usata quando la distribuzione dei residui non è normale o vi è la presenza di outliers che influenzano il modello. `lmrob` è utilizzata per analizzare i dati che sono influenzati dagli outliers in modo che i modelli risultanti siano resistenti a quest'ultimi. In questa sezione spiegheremo i metodi robusti di stima forniti dalla funzione.

2.2.1 Stima M

Uno dei metodi robusti di stima della regressione che fornisce la funzione `lmrob` è la stima M.
Gli stimatori M tentano di minimizzare la somma di una funzione scelta $rho(.)$, la quale agisce sui residui.
Gli stimatori M sono dati da

$$\hat{\beta} = arg\min_{\beta} \sum_{i=1}^{n} \rho(\epsilon_i(\beta)) \tag{2.4}$$

La M sta per "massima verosimiglianza" poiché $\rho(.)$ è correlato alla funzione di verosimiglianza per un'adeguata distribuzione residua presunta.

Si noti che, se si assume la normalità, $\rho(z) = \frac{1}{2}z^2$ risulta nella stima ordinaria dei minimi quadrati. Alcuni stimatori M sono influenzati dalla scala dei residui, quindi viene utilizzata una versione invariante della scala dello stimatore M:

$$\hat{\beta} = arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i(\beta)}{\tau}\right) \tag{2.5}$$

dove τè una misura della scala. Una stima di τ è data da

$$\hat{\tau} = \frac{\text{med}_i|r_i - \tilde{r}|}{0.6745} \tag{2.6}$$

dove $\tilde{r}$ è la mediana dei residui. La minimizzazione di quanto sopra si ottiene principalmente in due passaggi:

1. Set $\frac{\partial \rho}{\partial \beta_j} = 0$ per ogni $j = 0, 1, \ldots, p-1$ risultante in un insieme di equazioni non lineari p

$$\sum_{i=1}^{n} X_{i,j}\psi\left(\frac{\epsilon_i}{\tau}\right) = 0 \tag{2.7}$$

   dove $\psi(.) = \rho'(.)$. $\psi(.)$ è chiamata la funzione di influenza.

2. Un metodo numerico chiamato iterativamente **reweighted least squares (IRLS)** viene utilizzato per stimare iterativamente la stima dei minimi quadrati pesati fino a quando non viene soddisfatto un criterio di arresto. In particolare, per le iterazioni $t = 0, 1, \ldots$

$$\hat{\beta}^{(t+1)} = \left(\mathbf{X}^{\mathrm{T}}\left[\mathbf{W}^{-1}\right]^{(t)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\left[\mathbf{W}^{-1}\right]^{(t)}\mathbf{Y} \tag{2.8}$$

   dove $\left[\mathbf{W}^{-1}\right]^{(t)} = \text{diag}\left(w_1^{(t)}, \ldots, w_n^{(t)}\right)$ tale che

$$w_i^{(t)} = \begin{cases} \frac{\psi\left[\left(Y_i - X_i\beta^{(t)}\right)/\hat{\tau^{(t)}}\right]}{\left(Y_i X_i\beta^{(t)}\right)/\hat{\tau^{(t)}}} & \text{if } \; Y_i \neq X_i\beta^{(t)}, \\ 1, & \text{if } \; Y_i = X_i\beta^{(t)} \end{cases} \tag{2.9}$$

`lmrob` ha diverse funzioni scelte nella stima M. Qui mostreremo solo i più comuni: la funzione biquadrata di Tukey definita come

$$\begin{aligned} \rho(z) &= \begin{cases} \frac{c^2}{3}\left\{1 - \left[1 - \left(\frac{z}{c}\right)^2\right]^3\right\}, & \text{if } \; |z| < c \\ 2c, & \text{if } \; |z| \geq c \end{cases} \\ \psi(z) &= \begin{cases} z\left[1 - \left(\frac{z}{c}\right)^2\right]^2, & \text{if } \; |z| < c \\ 0, & \text{if } \; |z| \geq c \end{cases} \\ w(z) &= \begin{cases} \left[1 - \left(\frac{z}{c}\right)^2\right]^2, & \text{if } \; |z| < c \\ 0, & \text{if } \; |z| \geq c \end{cases} \end{aligned} \tag{2.10}$$

dove $c \approx 4.685$.

2.2.2 Stima S

Altro metodo robusto di stima della regressione che la funzione `lmrob` fornisce è la stima S. Questo metodo si basa sulla scala residua della stima M. La debolezza della stima M è la mancanza di considerazione sulla distribuzione dei dati e non una funzione dei dati complessivi, poiché utilizza solo la mediana come valore ponderato. Questo metodo utilizza la deviazione standard residua per superare l'inefficienza della mediana.
Lo stimatore S è definito da

$$\hat{\beta} = arg \min_{\beta} \hat{\sigma}_s (r_i) \tag{2.11}$$

con determinazione dello stimatore di scala robusto minimo $\hat{\sigma}_s$ e soddisfacente

$$arg \min_{\beta} \sum_{i=1}^{n} \rho \left(\frac{\epsilon_i(\beta)}{\hat{\sigma}_s} \right)$$

dove

$$\hat{\sigma}_s = \sqrt{\frac{1}{nK} \sum_{i=1}^{n} w_i r_i^2}$$

$K = 0.199$, $w_i = w_\sigma(z) = \frac{\rho(z)}{z^2}$ e la stima iniziale è

$$\hat{\sigma}_s = \frac{\text{med}_i |r_i - \tilde{r}|}{0.6745}$$

Tra le funzioni scelte nella stima S che `lmrob` fornisce, di nuovo il più comune è la funzione bi-quadrata di Tukey che è definita come

$$\begin{aligned} \psi(z) &= \begin{cases} z \left[1 - \left(\frac{z}{c} \right)^2 \right]^2, & \text{if } |z| < c \\ 0, & \text{if } |z| \geq c \end{cases} \\ w(z) &= \begin{cases} \left[1 - \left(\frac{z}{c} \right)^2 \right]^2, & \text{if } |z| < c \\ 0, & \text{if } |z| \geq c \end{cases} \end{aligned}$$

$z = \frac{r_i}{\sigma_s}$ e $c \approx 1.547$.

2.2.3 Stima MM

La procedura di stima MM consiste nel stimare il parametro di regressione usando la stima S che minimizza la scala del residuo dalla stima M e quindi procede con la stima M. La stima MM ha lo scopo di ottenere stime con un valore di scomposizione elevato e più efficienti. Il valore del punto di rottura è una misura comune della proporzione di valori anomali che possono essere affrontati prima che queste osservazioni influenzino il modello.
Lo stimatore MM è la soluzione di

$$\sum_{i=1}^{n} \rho_1 \left(\frac{r_i}{s_{MM}} \right) X_{ij} = 0$$

dove s_{MM} è la deviazione standard ottenuta dal residuo della stima S e ρ è una funzione bi-quadrata di Tukey definita per l'equazione 2.10.

2.3 `lmrob`: implementazione ed esempi di utilizzo

Nella sezione precedente abbiamo spiegato la teoria statistica relativa alla funzione `lmrob` insieme a ciò che la funzione cerca di risolvere.
In questa sezione esamineremo in dettaglio l'implementazione della funzione (gli argomenti che la funzione accetta e la sua descrizione) e forniremo anche alcuni esempi di utilizzo.

2.3.1 lmrob

```
lmrob( formula="", data={}, subset=None, weights=array([], dtype=float64),
       na_action="drop", method="", model=True, return_x=True,
       return_y=True, singular_ok=True, contrasts={},
       offset=array([], dtype=float64), control=None, init={}, **kwargs )
```

Gli argomenti principali, quelli con cui la funzione esegue i metodi di stima su un insieme di dati, sono: formula, data, method e control. Gli altri argomenti sono opzionali e configurano solo gli output della funzione; la loro descrizione dettagliata può essere vista nella documentazione della funzione (si veda []).
I quattro argomenti menzionati sono necessari ed importanti affinché la funzione `lmrob` possa funzionare .

- **formula:** [str] Una descrizione simbolica del modello da inserire.
- **data:** [dict] Un dataframe, contenente le variabili presenti nel modello.
- **method:** [str] Stringa che specifica il metodi robusti di stima da applicare sui dati.
- **control:** [object] Istanze di una classe chiamata LmrobControl. Questo oggetto specifica il tipo di funzione $\rho(.)$ (e $\psi(.)$) da utilizzare negli attributi control.chi e/o control.psi. Di default usa la funzione bi-quadrata di Tukey.

Tra i parametri che la funzione restituisce si hanno: coefficienti, residui e scala. Gli altri parametri forniscono ulteriori informazioni sul particolare modello e possono essere visualizzati più dettagliatamente nella documentazione della funzione (si veda []).

- **coefficients:** [array_like] Una descrizione simbolica del modello da inserire. Con essi viene costruito il modello denotato dall'equazione **??**.
- **residuals:** [array_like] Residui associati allo stimatore.
- **scale:** La scala residua utilizzata nello stimatore M.

2.3.2 Esempio di utilizzo

Per mostrare un esempio generale dell'uso della funzione `lmrob`, presentiamo sotto le seguenti righe di codice python.

```
from lmrob import *

x = [...] # data
y = [...] # data

data = {
  "y" : y,
  "x" : x
}

# Imposta la formula
formula = "y ~ x"

# Imposta il metodo di stima
method = "S"

# Imposta il tipo di funzione da usare nel metodo di stima
control = LmrobControl(psi="bisquare", method=method)

# Esegue la fzunione lmrob
m0 = lmrob(formula, data=data, method=method, control=control)
```

In questo esempio, le variabili x e y utilizzate sono matrici di vettori contenenti dati (per la precisione, numeri reali).
Con la seguente riga di codice `formula = "y ~ x"` viene stabilita la relazione delle variabili del modello: Y come variabile dipendente e X come variabile indipendente.
Nel dizionario `data` vengono incapsulato i dati che vogliamo modellare.
La riga di codice `method = "S"` stabilisce che il modello di regressione sarà determinato con il metodo di stima S.
Il tipo di funzione per lo stimatore M è stato stabilito con `control = LmrobControl (psi ="bisquare ", metodo =metodo)`.
Alla fine, il risultato della chiamata alla funzione viene incapsulato nella variabile m0.

2.4 Simulazione

Per dimostrare l'utilità della funzione `lmrob`, esaminiamo un esempio con una grande percentuale di outliers.
Useremo come esempio l'indagine statistica belga (a cura del ministero dell'Economia), in cui vengono forniti i numeri totali di chiamate internazionali effettuate dal 1950 al 1973.
I dati sui quali si svolgerà la simulazione sono stati ottenuti dal seguente riferimento e sono elencati nella Tabella 2.1.
Dal 1964 al 1969 contiene una forte contaminazione. Nella tabella 2.1, abbiamo contrassegnato i valori spuri con (*).

Tabella 2.1: Caso di studio.

X	Y	r_{LOS}	r_S	r_{MM}
50	0.44	1.245	2.03100e-01	1.77400e-01
51	0.47	0.7709	1.22900e-01	9.73000e-02
52	0.47	0.2668	1.27000e-02	-1.28000e-02
53	0.59	-0.1173	2.25000e-02	-2.90000e-03
54	0.66	-0.5514	1.77000e-02	-4.30000e-02
55	0.73	-0.9855	5.79000e-02	-8.31000e-02
56	0.81	-1.4096	8.81000e-02	-1.13200e-01
57	0.88	-1.8437	1.28300e-01	-1.53300e-01
58	1.06	-2.1678	5.85000e-02	-8.34000e-02
59	1.20	-2.5319	2.87000e-02	-5.35000e-02
60	1.35	-2.886	1.11000e-02	-1.36000e-02
61	1.49	-3.2501	4.09000e-02	1.63000e-02
62	1.61	-3.6342	5.07000e-02	2.62000e-02
63	2.12 (*)	-3.6283	4.50500e-01	4.26100e-01
64	11.9 (*)	5.6476	1.01203e+01	1.00960e+01
65	12.4 (*)	5.6435	1.05101e+01	1.04859e+01
66	14.2 (*)	6.9394	1.21999e+01	1.21758e+01
67	15.9 (*)	8.1353	1.37897e+01	1.37657e+01
68	18.2 (*)	9.9312	1.59795e+01	1.59556e+01
69	21.2 (*)	12.4271	1.88693e+01	1.88455e+01
70	4.30 (*)	-4.977	1.85910e+00	1.83540e+00
71	2.40	-7.3811	1.51100e-01	-1.74700e-01
72	2.70	-7.5852	3.87000e-02	1.52000e-02
73	2.90	-7.8893	1.28500e-01	1.05100e-01

Applicando il metodo OLS, otteniamo $Y = 0.5041X - 26.01$ che corrisponde alla linea tratteggiata nella figura 2.1.
Questa linea tratteggiata è stata largamente influenzata dai valori tra 1964-1969. I residui OLS

vengono etichettati nella tabella 2.1 come r_{OLS}.
Ora applichiamo la stima S e la stima MM della funzione `lmrob` con la classe bi-quadrata.

```
x = [50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73] #
                                                    data
y = [0.44,0.47,0.47,0.59,0.66,0.73,0.81,0.88,1.06,1.20,1.35,1.49,1.61,2.12,11.9,12.
                                                    4,14.2,15.9,18.2,21.2,4.30,2.40,2.70,2.90]
                                                     # data

data = {
  "y" : y,
  "x" : x
}

# Imposta la formula
formula = "y ~ x"

# Imposta il metodo di stima
method = "S"

method_2= "MM"

# Imposta il tipo di funzione da usare nel metodo di stima
control = LmrobControl(psi="bisquare", method=method)

# Esegue la fzunione lmrob
m0 = lmrob(formula, data=data, method=method, control=control)

m1 = lmrob(formula, data= data, method=method_2, control=control)
```

Questo produce $Y = 0.1102X - 5.2731$ per la stima S (tracciata come una linea continua nella figura 2.1) e $Y = 0.1101X - 5.2424$ per la stima MM (tracciata come una linea continua), entrambi trascuarno i valori anomali e forniscono una dei modelli di regressione ottimali. I residui della stima S e stima MM sono rispettivamente etichettate come r_S e r_{MM} nella Tabella 2.1.

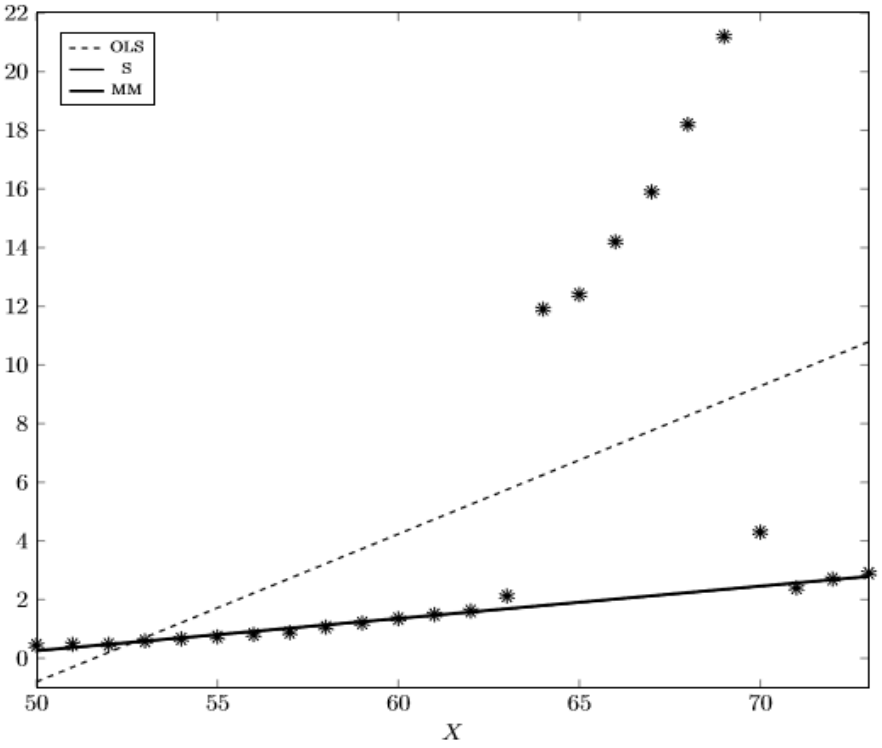

Figura 2.1: Confronto di determinati modelli di regressione

3 robustbasetPy Package

robustbasePy è un package scritto in Python per l'analisi dei dati attraverso vari metodi robusti, programmato per un ambiente Linux. Il codice è disponibile in forma open source al seguente repository github : https://github.com/dipperalbel/python-project . La documentazione è disponibile al seguente link : https://dipperalbel.github.io/python-project .

3.1 Installazione

Per installare il package, basta eseguire la seguente linea di commando all'interno di una shell :

```
$ pip install robustbasePy
```

3.2 Le funzioni del robustbasePy package

Oltre ad lmrob, il package robustbasePy offre molte altre funzioni, ed in questa sezione descriveremo le più importanti.

3.2.1 lmrob

La funzione lmrob calcola un stimatore di regressione di tipo MM come descritto in Yohai (1987) e Koller e Stahel (2011).
Di default usa una funzione di punteggio scendente bi-quadrato e restituisce uno stimatore altamente affidabile e altamente efficiente (con il 50% di punto di rottura ed il 95% efficienza asintotica per errori normali).
Il calcolo viene eseguito da una chiamata a lmrob.fit().
Da osservare che i problemi di convergenza possono ancora apparire come avvertenze,

```
S refinements did not converge (to refine_tol=1e-07) in 200 (= k_max) steps
```

e spesso si può semplicemente porre rimedio aumentando (cioè indebolendo) refine_tol o aumentando il numero consentito di iterazioni k_max in LmrobControl (Si veda il riferimento per più dettagli).

3.2.2 lmrob_fit

La funzione lmrob_fit calcola stimatori di regressione di tipo MM.
Uno stimatore S viene utilizzato come valore iniziale e da questo viene utilizzato uno stimatore M con scala fissa e funzione ψ riduttrice. Opzionalmente viene calcolato un passo D (stima della scala adattativa di progetto) e un secondo passo M. Restituisce un oggetto dizionario.
Questa funzione è la funzione di base per la stima di tipo MM, chiamata da lmrob e tipicamente non utilizzabile da sola (si veda il riferimento [6] per maggiori dettagli).

3.2.3 lmrob_S

Questa funzione calcola uno stimatore S per la regressione lineare, utilizzando l'algoritmo S veloce e restituisce un oggetto dizionario.
Questa funzione è utilizzata da lmrob_fit e in genere non deve essere utilizzata da sola (poiché uno stimatore S ha un'efficienza troppo bassa "da solo").
Per impostazione predefinita, l'algoritmo di sottocampionamento utilizza una scomposizione LU personalizzata che garantisce un sottocampione non singolare (se ciò è possibile). Ciò rende l'algoritmo Fast-S anche fattibile per dati categoriali e misti continui e categoriali.
Si può ripristinare il vecchio schema di sottocampionamento impostando il sottocampionamento dei parametri in controllo su "semplice".

3.2.4 `lmrob_M_S`

La funzione `lmrob_M_S` calcola uno stimatore M-S per la regressione lineare usando l'algoritmo "M-S". Questa funzione è usata da `lmrob` e non è pensata per essere usata da sola (perché uno stimatore M-S ha un'efficienza troppo bassa "da solo").
Uno stimatore M-S è una combinazione di uno stimatore S per le variabili continue e uno stimatore L_1 (cioè un stimatore M con $\psi(t) = \text{sign}(t)$) per le variabili categoriali. Lo stimatore S è stimato utilizzando un algoritmo di sottocampionamento.
Se il modello include interazioni tra variabili categoriali (fattore) e continue, l'algoritmo di sottocampionamento potrebbe non riuscire. In questo caso, si può scegliere di assegnare l'interazione al lato categoriale delle variabili piuttosto che al lato continuo (si veda il riferimento [6] per maggiori dettagli).

3.2.5 `lmrob__D__fit`

La funzione `lmrob__D__fit` calcola una stima della wcala adattativa di progetto per una data stima MM. Questo dovrebbe far parte di una catena di stime come SMD o SMDM.
Restituisce un oggetto dizionario.
Inoltre, questa funzione è utilizzata da `lmrob_fit` e in genere non deve essere utilizzata da sola.

3.2.6 `lmrob__M__fit`

La funzione `lmrob__M__fit` esegue le iterazioni RWLS per trovare uno stimatore M della regressione. Quando è iniziato da una beta-iniziale stimata in S, ciò si traduce in uno stimatore MM.
Restituisce un dizionario Questa funzione è utilizzata da `lmrob_fit` e in genere non deve essere utilizzata da sola.

3.2.7 `lmrob_lar`

Stimatori di regressione di tipo MM calcolati: uno stimatore S viene utilizzato come valore iniziale e da questo viene utilizzato uno stimatore M con scala fissa e funzione psi riduttrice. Opzionalmente viene calcolato un passo D (stima della scala adattativa di progetto) e un secondo passo M. Restituisce un oggetto dizionario.
Questa funzione è la funzione di base per la stima di tipo MM, chiamata da `lmrob` e che in genere non deve essere utilizzata da sola. Se indicato, init deve essere un elenco di stime iniziali contenenti almeno i coefficienti iniziali e scala come coefficiente e scala. Altrimenti chiama `lmrob.S(·)` e lo usa come stimatore iniziale.

3.2.8 `LmrobControl`

La classe `LmrobControl` sintonizza i parametri per `lmrob`, lo stimatore di regressione di tipo MM e gli stimatori S, M e D associati. Usando `setting="KS2011"` imposta i valori predefiniti come suggerito da Koller e Stahel (2011) e analogamente per `"KS2014"`.
Restituisce un oggetto dict con oltre venti componenti.
Le funzioni `.M*.default` e le liste `.M*.defaults` contengono parametri di sintonizzazione predefiniti per tutte le funzioni ψ predefinite.

3.2.9 `nlrob`

La funzione `nlrob` si adatta a un modello di regressione non lineare mediante metodi robusti.
Di default, da uno stimatore M, utilizzando i minimi quadrati pesati in iterazione (denominati IRLS o anche IWLS).

4 Conclusioni

Nella tesi abbiamo descritto come l'analisi robusta sia nece