UNIVERSITÀ
DI PARMA

Dipartimento di Scienze Economiche e Aziendali

Corso di Laurea triennale in Economia e Management

Analisi robusta multivariata, un caso di studio: Real estate dataset.

Relatore:
Chiar.mo Prof.
Marco Riani

Laureando:
Luca Ratti

Anno Accademico 2021-2022

“Tra vent’anni non sarete delusi delle cose che avete fatto, ma da quelle che non avete fatto. Allora levate l’ancora, abbandonate i porti sicuri, catturate il vento nelle vostre vele. Esplorate. Sognate. Scoprite.”

Mark Twain

INDICE

ELENCO DELLE FIGURE

INTRODUZIONE E SCOPO DELLA TESI

Lo scopo del caso di studio pone obiettivo l'analisi previsionale del prezzo dell'unità di superfice di alcune abitazioni della città di Taipei (distretto di Xindian) date le loro caratteristiche, nonché lo studio dei relativi valori anomali, *outliers*.

L'analisi avviene attraverso il modello di regressione, prima semplice e poi robusta delle variabili esplicative considerate nel dataset utilizzando il *metodo dei minimi quadrati* (OLS). L'idea principale della regressione lineare multipla è quella di colmare la mancanza di informazioni che determina una distorsione nella corretta identificazione della variabile *y*. In altri termini, la regressione con più regressori (variabili indipendenti) consente, ovviamente se i dati sono disponibili, di misurare l'effetto di una specifica variabile x_i sulla variabile *y*, tenendo costanti le altre variabili indipendenti.

L'analisi degli *outliers* si basa sullo studio dei *residui*, ovvero le differenze tra i valori osservati nel dataset e i valori stimati calcolati con l'equazione di regressione, al fine di verificare se gli stessi hanno la possibilità di influenzare il modello di regressione e, di conseguenza in questo caso specifico, la variabile *y*, ovvero il prezzo dell'abitazione per unità.

Capitolo 1
IL DATASET

1.1 **Introduzione**

Il dataset comprende 414 osservazioni riguardo delle abitazioni della città di Taipei (distretto di Xindian) e considera le seguenti variabili: la data di acquisto dell'immobile espressa in mese ed anno, il numero di anni dalla costruzione dell'immobile, la distanza dalla più vicina fermata della metro, il numero di minimarket vicini all'abitazione, latitudine, longitudine ed il prezzo per unità di superficie.

Il dataset è disponibile sul sito di Kaggle.com.

1.1.1 *Importazione*

L'importazione del dataset avviene utilizzando l'apposita funzione di Matlab *readtable* la quale permette di caricare da un file excel una tabella usata come tabella di partenza per svolgere le analisi. La funzione si esplica nel seguente metodo:

```
clc

close all

tab=readtable("Real_estate.csv","VariableNamingRule","preserve");
```

L'opzione "VariableNamingRule" collegata all'impostazione "preserve" permette di mantenere inalterate le nominazioni delle variabili.

1.1.2 *Selezione delle variabili*

Tramite le seguenti impostazioni è possibile rinominare le variabili ed eliminare la prima colonna ("No") in quanto rappresenta solamente il numero della riga. Vengono altresì create due variabili (y e X) rispettivamente contenenti i valori del prezzo delle

case per unità di superficie nella prima variabile e l'eta dell'immobile, la distanza dalla metro più vicina ed il numero di discount più vicini.

```
tab=renamevars(tab,["X1 transaction date" "X2 house age" "X3 distance to the
nearest MRT station" ...
    "X4 number of convenience stores" "X5 latitude" "X6 longitude" "Y house
price of unit area"], ...
    ["Transaction date" "House age" "Distance to the nearest MRT
station" ...
    "Number of convenience store" "Latitude" "Longitude" "House price of
unit area"]);
tab=removevars(tab,"No");

y=tab{:,"House price of unit area"};
X=tab{:,["House age" "Distance to the nearest MRT station" "Number of
convenience store"]};
```

Utilizzando le apposite funzioni head e *tail*, vengono mostrate, come segue, le prime e le ultime 8 righe del set.

Transaction date	House age	Distance to the nearest MRT station	Number of convenience store	Latitude	Longitude	House price of unit area
2012.9	32	84.879	10	24.983	121.54	37.9
2012.9	19.5	306.59	9	24.98	121.54	42.2
2013.6	13.3	561.98	5	24.987	121.54	47.3
2013.5	13.3	561.98	5	24.987	121.54	54.8
2012.8	5	390.57	5	24.979	121.54	43.1
2012.7	7.1	2175	3	24.963	121.51	32.1
2012.7	34.5	623.47	7	24.979	121.54	40.3
2013.4	20.3	287.6	6	24.98	121.54	46.7

Figure 1-1: Prime 8 righe del dataset

Transaction date	House age	Distance to the nearest MRT station	Number of convenience store	Latitude	Longitude	House price of unit area
2012.9	32	84.879	10	24.983	121.54	37.9
2012.9	19.5	306.59	9	24.98	121.54	42.2
2013.6	13.3	561.98	5	24.987	121.54	47.3
2013.5	13.3	561.98	5	24.987	121.54	54.8
2012.8	5	390.57	5	24.979	121.54	43.1
2012.7	7.1	2175	3	24.963	121.51	32.1
2012.7	34.5	623.47	7	24.979	121.54	40.3
2013.4	20.3	287.6	6	24.98	121.54	46.7

Figure 1-2: Ultime 8 righe del dataset

Capitolo 2
ANALISI DESCRITTIVA: LE ANALISI DI SINTESI

2.1 **Introduzione**

La statistica descrittiva è una disciplina che è responsabile della raccolta, dell'archiviazione, dell'ordinamento, della creazione di tabelle o grafici e del calcolo dei parametri di base sul set di dati.

La statistica descrittiva è, insieme all'inferenza statistica o statistica inferenziale, uno dei due grandi rami della statistica. Il suo stesso nome lo indica, cerca di descrivere qualcosa. Ma non descriverlo in alcun modo, ma in modo quantitativo.

2.2 **Le funzioni di analisi descrittiva nell'analisi del dataset**

In questa particolare analisi vengono utilizzati quattro differenti grafici, i quali riassumono con precisione, le evidenze descrittive del dataset. In particolare vengono utilizzati: istogrammi, boxplot, la matrice di correlazione ed il geoplot.

2.2.1 *La funzione Histogram*

Viene utilizzata la funzione histogram per visualizzare le variabili House age, Distance to the nearest MRT station, Number of convenience store e Price. Essa si configura come una rappresentazione grafica costituita da più rettangoli adiacenti (bins), ognuno dei quali ha per base un certo intervallo della variabile e un'altezza tale che la sua area rappresenti, nella scala prefissata, il relativo valore globale della funzione. Il codice Matlab è il seguente:

```
subplot(2,2,1)

histogram(X(:,1),25,FaceColor='b')
xlabel('House age (years)')

subplot(2,2,2)
```

```
histogram(X(:,2),25,FaceColor='g')
xlabel('Distance to the nearest MRT station')

subplot(2,2,3)
histogram(X(:,3),25,FaceColor='c')
xlabel('Number of convenience store')

subplot(2,2,4)
histogram(y,25,FaceColor='k')
xlabel('Price')
```

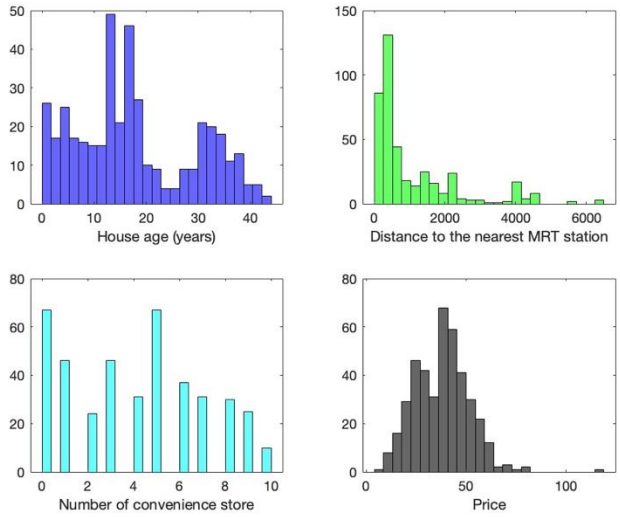


Figure 2-1: Histograms

Dagli istogrammi è possibile vedere in quale range di valori si distribuiscono maggiormente le diverse variabili. Di conseguenza, la variabile House age presenta 49 valori nell'intervallo [12,32 14,08]; Distance to the nearest MRT station presenta 131 valori nell'intervallo [260 520]; Number of convenience store mostra due gruppi principali da 67 valori negli intervalli [0 0.4] e [4,8 5,2]; infine, la variabile Price presenta la consueta distribuzione Normale (Gaussiana) con 68 valori nell'intervallo [36,2 40,8] e presenta un valore anomalo estremamente grande (117.5).

2.2.2 *La funzione Boxplot*

Viene utilizzata la funzione *boxplot* per osservare quali valori per ciascuna variabile possano essere considerati anomali (o presunti tali). Il boxplot è un grafico caratterizzato da quattro elementi principali:

1. Un segmento che indica la posizione della *mediana* della distribuzione,
2. Un rettangolo (*box*) la cui lunghezza indica il grado di dispersione del 50% dei valori centrali della distribuzione che si identificano come i valori compresi nella *distanza interquartile* ossia la differenza tra il terzo ed il primo quartile.
3. Due segmenti che, partendo dai lati minori del *box*, indicano fino a che valore si estendono le code (destra e sinistra) della distribuzione escludendo i presunti valori *anomali*.
4. Eventuali valori esterni considerati *anomali*.

I Boxplot vengono chiamati da Matlab tramite l'apposita funzione.

```
figure("Name","Boxplots")
subplot(2,2,1)
boxplot(tab.("House age"),"Labels","House age")
subplot(2,2,2)
boxplot(tab.("Distance to the nearest MRT station"),"Labels","Distance
nearest MRT station")
subplot(2,2,3)
boxplot(tab.("Number of convenience store"),"Labels","Number of convenience
store")
subplot(2,2,4)
boxplot(tab.("House price of unit area"),"Labels","House price of unit
area")
```

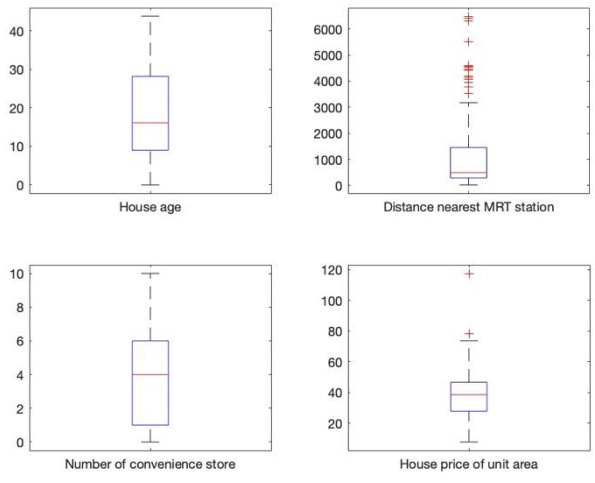


Figure 2-2: Boxplots

Come si evince dai 4 boxplots, le variabili *House age* e *Number of convenience store* non presentano particolari valori anomali presentando rispettivamente invece mediana di valore 16,1 e 4. Le variabili *Distance nearest MRT station* e *House price of unit area* presentano nel primo caso mediana uguale a 492,2313 e numerosi valori anomali di grande valore posizionati nell'intervallo [3529,564 6488,021], mentre nel secondo caso la mediana è uguale a 38,45 e si notano due valori anomali, uno molto vicino alla coda destra (78,3) ed uno, invece, molto distante (117,5).

2.2.3 *La funzione Corrplot*

La *matrice di correlazione* (*corrplot*) indica i coefficienti di correlazione di tutte le coppie di variabili nella matrice input. Ogni *subplot* al di fuori della diagonale contiene uno *scatterplot* di una specifica coppia di variabili a cui viene associata una linea di riferimento basata sui minimi quadrati (retta di regressione), mentre ogni subplot sulla diagonale rappresenta un istogramma della distribuzione della variabile. Ogni coefficiente di correlazione lineare [-1 1] è ottenuto calcolando la *covarianza* calcolata sugli scostamenti standardizzati secondo la formula:

$$r_{xy} = \frac{COV(X,Y)}{\sqrt{VAR(X)VAR(Y)}}$$

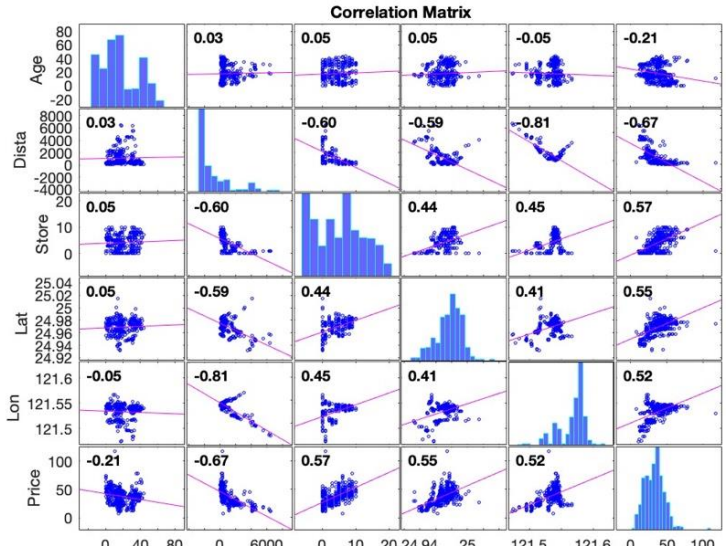


Figure 2-3: Matrice di correlazione

```
varnames={'Age' 'Distance MRT' 'Store' 'Lat' 'Lon' 'Price'};
figure
[R,PValue]=corrplot(tab{:,2:7},'varnames',varnames);
```

Nella matrice in considerazione notiamo sia relazioni negative ($r_{xy} < 0$), sia positive ($r_{xy} > 0$), sia tendenti a valori prossimi a 0 esprimendo una non correlazione. La funzione *corrplot* permette anche il calcolo dei relativi *p-values*.

2.2.4 *La funzione Geoplot*

La funzione *geoplot* crea una cartina geografica, basata su latitudine e longitudine, che permette di identificare le osservazioni (abitazioni) su una mappa geografica. Il grafico è utile per analizzare se la posizione più o meno centrale in una determinata città (nel nostro caso Taipei) possa influenzare il prezzo per unità di superficie. In questo caso, la maggioranza delle abitazioni si trova nel centro del distretto di Xindian ed osservando la matrice di correlazione tra prezzo/latitudine e tra prezzo/longitudine, si può notare come al crescere della longitudine si verifichi un aumento del prezzo così come per la latitudine; infatti, all'aumentare delle due variabili ci si sposta verso il centro del distretto.

```
figure('Name','House distribution in Taipei')
geoplot(tab.Latitude,tab.Longitude,"o");
geobasemap streets-light
title('House distribution in Taipei')
```

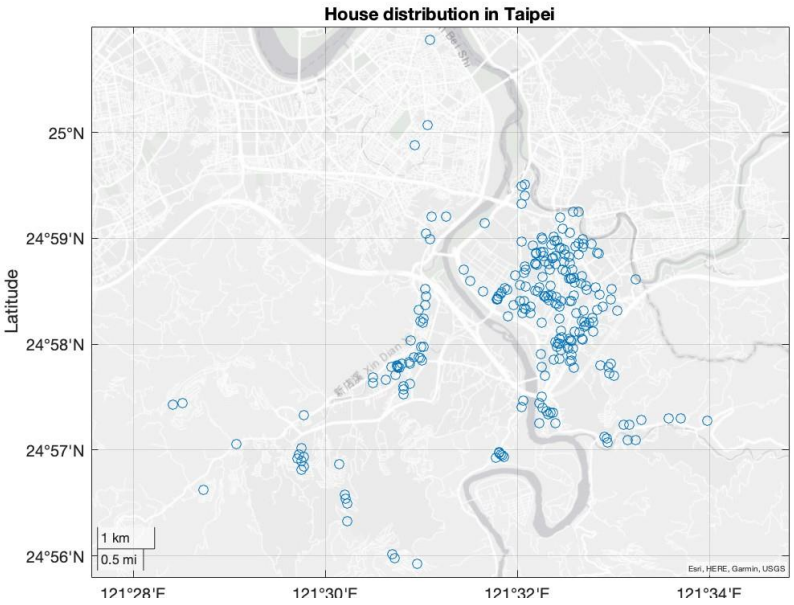


Figure 2-4: Geoplot

Capitolo 3
LA REGRESSIONE: MODELLO SEMPLICE E ROBUSTO

3.1 **Introduzione**

Tramite i modelli di regressione è possibile capire se ciascun valore osservato dalla variabile dipendente sia esprimibile come funzione lineare del corrispondete valore della variabile esplicativa (indipendente) più un residuo che mostra l'incapacità del modello di riprodurre con esattezza la realtà. Date *n* coppie di osservazioni dei valori (x_i, x_j) di due fenomeni quantitativi *X* ed *Y*, la regressione lineare è esprimibile da una retta con equazione:

$$y_i = a + bx_i + e_i \qquad \text{con } i = 1,2,\dots,n$$

dove:

- y_i ed x_i sono valori della variabile dipendente e della variabile idipendente
- *a* è l'intercetta della retta di regressione con l'asse *y* (punto di origine)
- *b* è il coefficinete angolare (o *di regressione*).
- e_i è il *residuo*.

La funzione lineare dei valori x_i rappresenta la retta di regressione:

$$\widehat{y_i} = a + bx_i \qquad \text{con } i = 1,2,\dots,n$$

dove l'intercetta a è il valore che assume $\widehat{y_i}$ quando *X* è pari a 0, mentre i residui sono considerati come la differenza tra i valori teorici $\widehat{y_i}$ ed i valori osservati y_i.

A livello pratico, prima di costruire un modello di regressione lineare, quello che è importante fare è:

- Statistiche descrittive e grafici per le singole variabili (ad esempio istogrammi)
- Controllare la presenza di valori impossibili (ad esempio, un'età negativa) o inusuali

- Costruire i diagrammi di dispersione per controllare la presenza di combinazioni inusuali di valori (in gergo, outliers bivariati).

Una volta costruito il modello, è invece importante:

- Valutare graficamente i residui (cioè le stime osservate degli errori) per capire in primo luogo se sono tra loro i.i.d. ed in secondo luogo se hanno una distribuzione Normale (per verificarlo, puoi costruire un istogramma dei residui o un grafico quantile-quantile), come vedremo tra poco.
- Esaminare le misure di influenza (distanza di Cook, DFBETA, DFITS, residui studentizzati, …). Questo perché, in generale, un'osservazione è considerata influente se la sua eliminazione dal campione provoca un cambio sostanziale nei risultati della regressione.

3.2 **I modelli di regressione nel dataset**

Di seguito viene eseguito il modello di regressione semplice tra il prezzo della casa per unità di superficie e le tre principali variabili indipendenti: *House age, Distance to the nearest MRT station* e *Number of convenience store.* Date le rette di regressione semplici, si osserva come varia la regressione andando ad utilizzare le rette di regressione robusta, le quali non considerano eventuali "*outliers*".

```
figure("Name", "Regressione semplice e robusta")

subplot(1,3,1)
plot(tab.("House age"),y,"o")
ls1=lsline;
ls1.Color='k';
ls1.LineWidth=2;
[outLTSage]=LXS(y,tab.("House age"));
b = outLTSage.beta;
hold('on')
plot(tab.("House age"),b(1)+b(2)*tab.("House age"), 'r','LineWidth' ,1.5);
legend({'Points' 'Least squares fit' 'Robust fit'},'Location','best');
xlabel("House age")
ylabel("House price of unit area")

subplot(1,3,2)
plot(tab.("Distance to the nearest MRT station"),y,"o")
ls2=lsline;
ls2.Color='k';
ls2.LineWidth=2;
[outLTSmrt]=LXS(y,tab.("Distance to the nearest MRT station"));
b = outLTSmrt.beta;
hold('on')
plot(tab.("Distance to the nearest MRT station"),b(1)+b(2)*tab.("Distance to
the nearest MRT station"), 'r','LineWidth' ,1.5);
legend({'Points' 'Least squares fit' 'Robust fit'},'Location','best');
```

```
xlabel("Distance to the nearest MRT station")
ylabel("House price of unit area")

subplot(1,3,3)
plot(tab.("Number of convenience store"),y,"o")
ls3=lsline;
ls3.Color='k';
ls3.LineWidth=2;
[outLTSmrt]=LXS(y,tab.("Number of convenience store"));
b = outLTSmrt.beta;
hold('on')
plot(tab.("Number of convenience store"),b(1)+b(2)*tab.("Number of
convenience store"), 'r','LineWidth' ,1.5);
legend({'Points' 'Least squares fit' 'Robust fit'},'Location','best');
xlabel("Number of convenience store")
ylabel("House price of unit area")
```

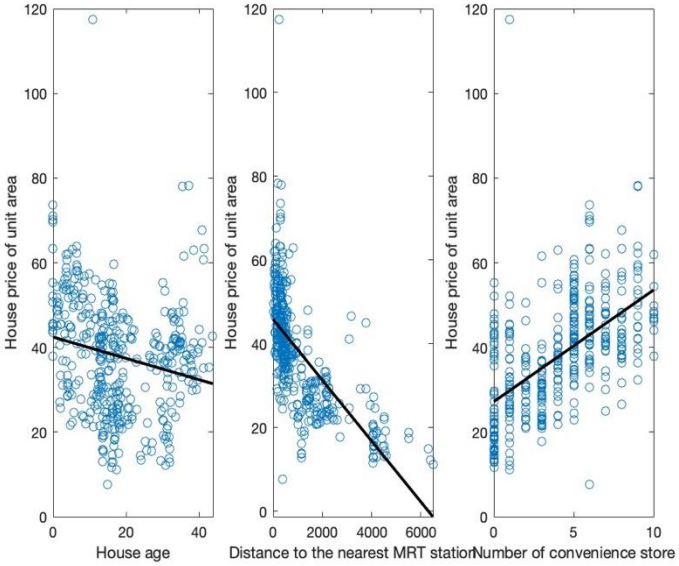


Figure 3-1: Regressione semplice

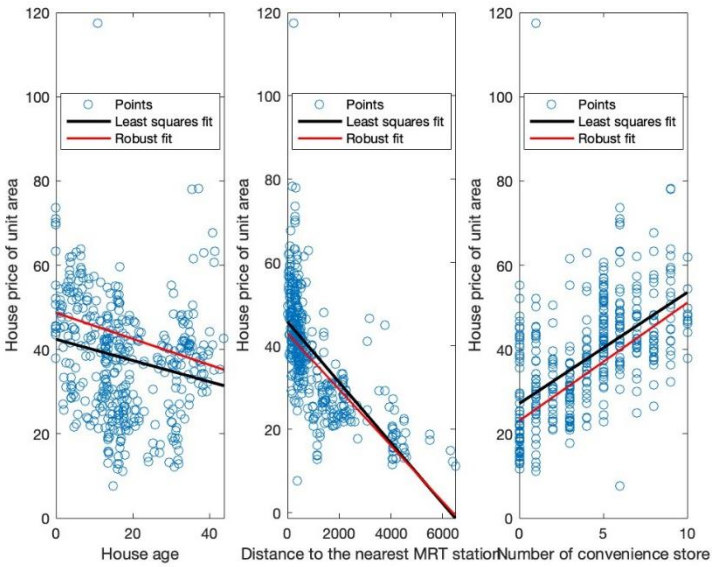


Figure 3-2: Regressione semplice contro Robusta

Come osservabile dai grafici la retta di regressione robusta non varia molto la precisione del modello.

3.3 **La funzione *yXplot* e *fitlm***

Tramite la funzione *yXplot* possiamo visualizzare le tre variabili indipendenti contro la variabile dipendente.

```
yXplot(y,X,'nameX',{'House age' 'Distance station' 'Convenience
store'},'namey',{'Hose price'});
```

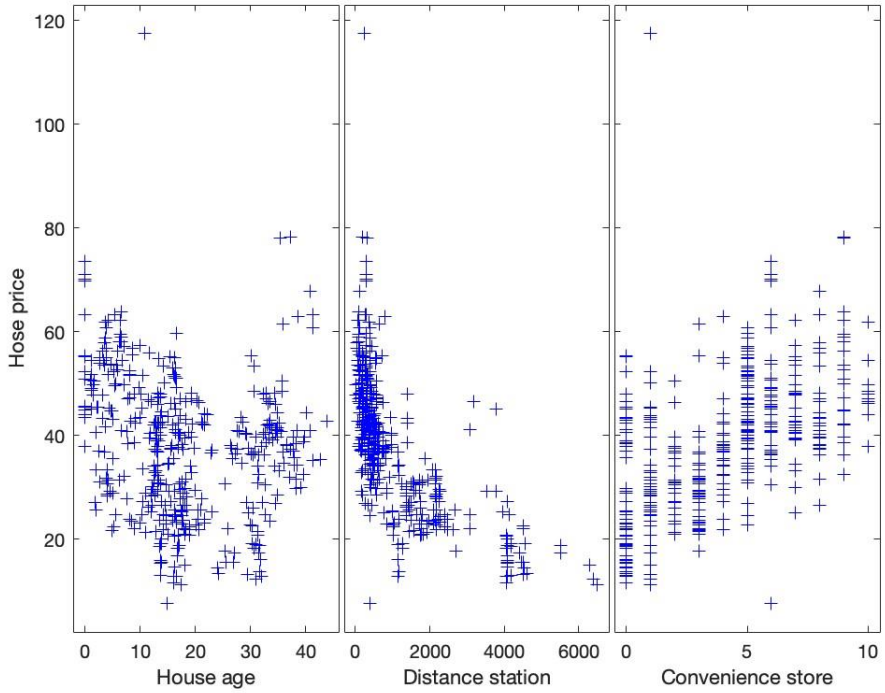


Figure 3-3: yXplot

I grafici ci permettono di osservare come sono posizionati i valori nella loro relazione con la variabile dipendente. Nel primo grafico (*House age*) notiamo come si formino due distinti gruppi divisi da un valore di *x* prossimo al 23. Nel secondo grafico (*Distance to the nearest MRT station*) osserviamo una grande quantità di valori vicini a valori compresi tra 0 e 2000, andando a delineare una relazione negativa tra le due variabili. Nel terzo ed ultimo grafico (*Number of convenience store*) osserviamo una leggera relazione positiva tra il numero di discount ed il valore della casa. Tutti e tre i grafici mostrano un evidente valore anomalo nella parte superiore del grafico.

Per verificare la bontà di adattamento del modello (R^2) ed il relativo p-value si utilizza la funzione *fitlm.*

```
out=fitlm(X,y);

disp(out)
```

```
Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ________    __________    _______    __________

    (Intercept)      42.977       1.3845      31.041    1.0856e-109
    x1              -0.25286     0.040105    -6.3048     7.4705e-10
    x2             -0.0053791    0.00045303  -11.874     3.7641e-28
    x3               1.2974      0.19429      6.6779     7.9085e-11


Number of observations: 414, Error degrees of freedom: 410
Root Mean Squared Error: 9.25
R-squared: 0.541,  Adjusted R-Squared: 0.538
F-statistic vs. constant model: 161, p-value = 5.44e-69
```

Figure 3-4: Fitlm

La funzione fornisce una bontà di adattamento $R^2 = 0.541$ ed un $p - value = 5.44e - 69$. Il primo indica che il modello spiega il 54,1% della variabilità del prezzo delle case per unità di superficie, mentre il p-value la significatività del test e che per valori inferiori generalmente a 0,05 permette di evidenziare una fortissima relazione lineare tra le variabili.

3.4 **La trasformazione del modello di regressione**

3.4.1 *Definizione, calcolo e valutazione di lambda*

È opportuno trasformare i dati in modo tale da ottenere una bontà di adattamento maggiore. Per trasformare i dati è opportuno calcolare i vari valori di *Lambda* (Λ) la quale è definita come misura della varianza percentuale nelle variabili dipendenti non spiegata dalle differenze nei livelli della variabile dipendente.

```
out=Score(y,X);


lam="lambda="+([-1:0.5:1]');
disp(array2table(out.Score,'RowNames',lam,"VariableNames","Score test"));
[outfan]=FSRfan(y,X,'plots',1,'init',5);
n=length(y);
```

	Score test
lambda=-1	25.589
lambda=-0.5	13.923
lambda=0	4.2308
lambda=0.5	-4.9648
lambda=1	-14.329

Figure 3-5: Valore di Lambda

Trovati i valori di Lambda, la funzione *FSRfan* permette di monitorare i valori dei *Test Statistici* per ogni lambda ottenendo:

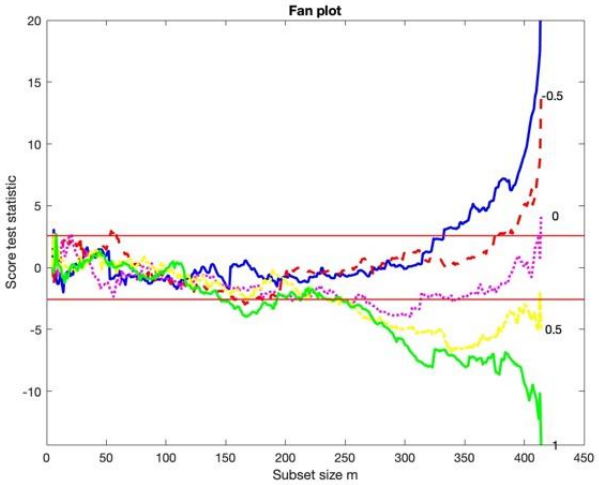


Figure 3-6: FSRfan

In particolare, con la funzione *fanBIC* è possibile identificare il miglior valore di Lamba con cui procedere alla trasformazione dei dati.

```
[out]=fanBIC(outfan);
```

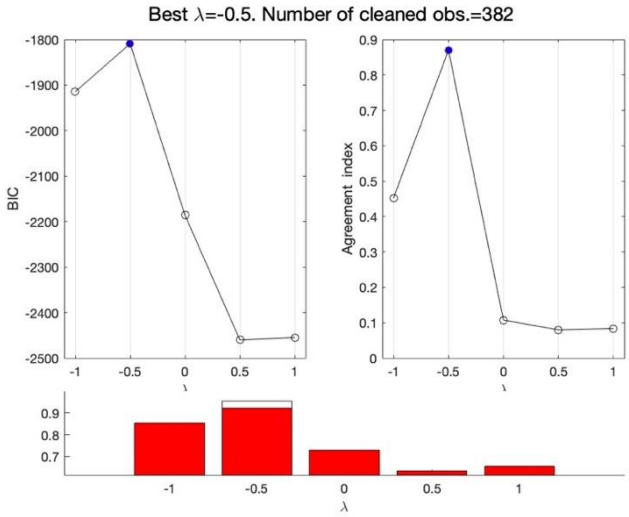


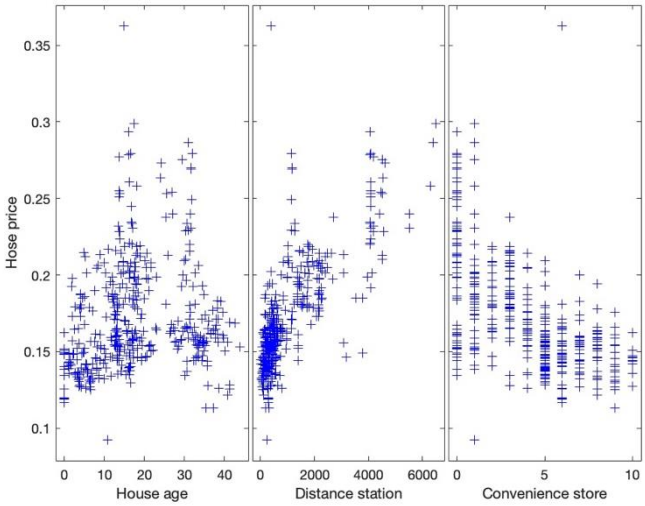
Figure 3-7: fanBIC

In questo caso il valore migliore di Lambda è -0.5, quindi:

$$y_{Lambda} = y^{-0.5}$$

3.4.2 *La variazione di yXplot e fitlm*

Si osserva la variazione dell'*yXplot* procedendo nuovamente con la funzione *fitlm*. Trasformando i dati otteniamo una bontà di adattamento maggiore $R^2 = 0.632$ ed un valore di $p - value = 1.31e - 88$. Il modello ora spiega il 63,2% della variabilità del prezzo delle case per unità di superficie con un *p-value* inferiore.



Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	0.1531	0.003321	46.102	2.6655e-164
x1	0.00053285	9.6197e-05	5.5391	5.4439e-08
x2	1.8178e-05	1.0867e-06	16.728	2.9389e-48
x3	-0.0026264	0.00046603	-5.6358	3.2462e-08

Number of observations: 414, Error degrees of freedom: 410
Root Mean Squared Error: 0.0222
R-squared: 0.632, Adjusted R-Squared: 0.629
F-statistic vs. constant model: 235, p-value = 1.31e-88

Figure 3-8/9: yXplot e fitlm con Lambda -0.5

Capitolo 4
ANALISI DEI RESIDUI: IDENTIFICAZIONE E VALUTAZIONE DEGLI OUTLIERS

4.1 **Introduzione**

Considerando il margine di imprecisione, nei modelli di regressione si aggiunge un termine di errore, che è indiato dalla lettera greca Epsilon (ε).

La variabile risposta (la y) nell'equazione di regressione è quindi determinata, come già annunciato, dai valori delle variabili esplicative (le x) più un termine d'errore (ε).

Affinché il modello di regressione riesca ad avere un buon potere predittivo, questo errore deve essere una variazione imprevedibile nella variabile risposta.

Per verificare se è effettivamente così, nella costruzione di un modello di regressione bisogna fare alcune verifiche su come si distribuiscono i residui.

I valori residui in un'analisi di regressione rappresentato proprio la parte di errore di previsione del modello di regressione.

I residui, detti anche scarti, rappresentano infatti le differenze tra i valori osservati nel dataset e i valori stimati calcolati con l'equazione di regressione. In altre parole, i residui indicano la variabilità dei dati attorno alla retta di regressione.

4.2 **Le funzioni di analisi degli outliers**

Ecco su cosa devi concentrarti analizzando i residui:

- I residui hanno una distribuzione normale?
- Le variabili indipendenti sono incorrelate con l'errore?
- La varianza dei residui è omogenea?
- La distribuzione dei residui è lineare?
- Ci sono degli outliers che influenzano la pendenza della retta?
- I residui sono tra loro correlati?

Il modello di regressione lineare non richiede né che la variabile dipendente (la y) né che le variabili indipendenti (le x) abbiano una distribuzione Normale.

Quello che in realtà richiede il modello è che gli errori siano tra loro indipendenti ed identicamente distribuiti (in gergo tecnico i.i.d.) in modo approssimabile ad una distribuzione Normale con media pari a 0 e varianza pari a σ^2:

$$\varepsilon \sim \text{i.i.d. } N(0, \sigma^2)$$

Di questa formula, la parte più importante (ma anche meno verificata) è quella relativa agli errori i.i.d. Il fatto che gli errori abbiano una distribuzione Normale invece è quella meno importante in quanto basta che ci sia un'approssimazione a tale distribuzione.

4.2.1 *La funzione qqplotFS*

Tramite le proprietà della funzione *fitlm* è possibile andare a ricavare i residui. Una volta individuati è possibile costruire il *qqplotFS* (quantile-quantile plot). Esso visualizza i quantili dei dati del campione rispetto ai valori quantilici teorici da una distribuzione normale. L'idea è che se i residui avessero una distribuzione normale, i loro quantili dovrebbero coincidere con quelli della distribuzione Normale. A livello visivo, questo significa che i punti dovrebbero disporsi lungo la bisettrice, indicata dalla retta presente nel grafico. Se il campione è distribuito normalmente il risultato risulta una retta.

```
outLM=fitlm(X,y1,'exclude','');
res=outLM.Residuals{:,3};
qqplot(res)

qqplotFS(res,'X',X,'plots',1);
title('qqplot of stud. res.')
```

Nella pratica, non capita quasi mai che i punti si dispongano esattamente lungo la bisettrice. Per poter dire che gli errori hanno una distribuzione normale basti quindi che i punti siano vicino alla linea presente nel grafico.

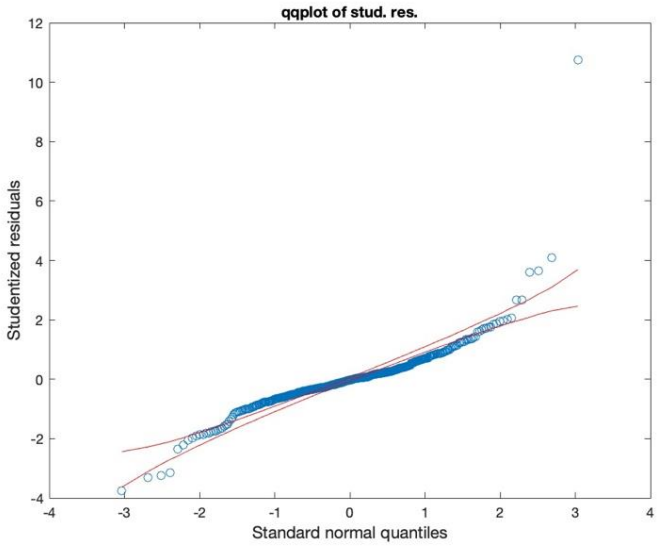


Figure 4-1: qqplot

Le bande (envelopes) sono formate da una matrice con due colonne corrispondenti ai valori di confidenza rispettivamente inferiori e superiori.

4.2.2 *La dispersione dei residui*

Per analizzare gli outliers identificati automaticamente e per verificare l'ipotesi di omogeneità delle varianze dei residui, è necessario creare un grafico a dispersione.

Si procede alla rappresentazione grafica della dispersione dei residui, sull'asse y, e dei valori adattati (risposte stimate), sull'asse x. Il grafico viene utilizzato per rilevare non linearità, varianze di errore disuguali e valori anomali. Se c'è omogeneità della varianza dei residui, i punti saranno dispersi in modo simile sia nella parte sinistra che in quella destra del grafico.

```
figure


plot(outLM.Fitted,res,'o')
xlabel('Fitted values')
ylabel('Residuals')
lsline
```

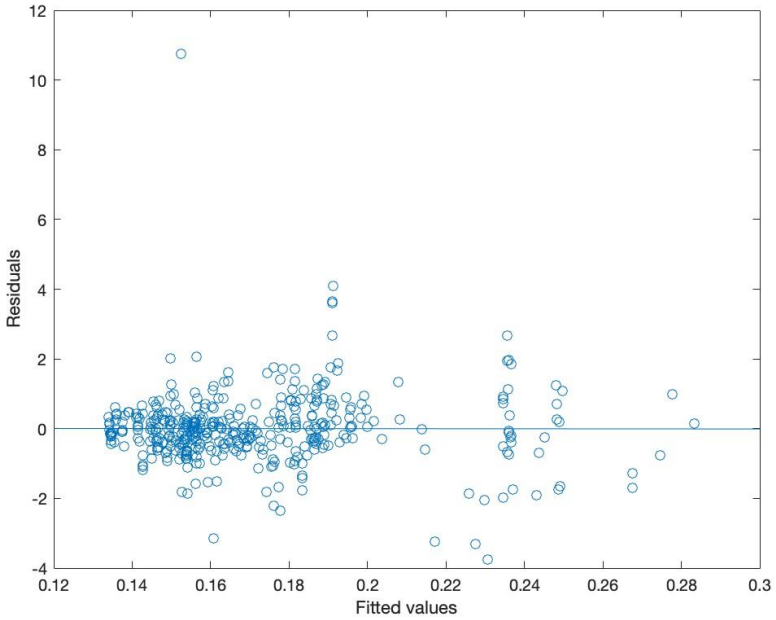


Figure 4-2: Plot di dispersione dei residui

Come è possibile osservare, in questo grafico è riportata una linea orizzontale tratteggiata in corrispondenza dei residui con media zero. I residui di un modello di regressione costruito con il metodo dei minimi quadrati (OLS), infatti, hanno per definizione sempre media zero.

Secondo l'ipotesi di linearità, i dati devono infatti distribuirsi in modo casuale intorno allo 0.

4.2.3 *L'identificazione automatica degli outliers*

È di conseguenza possibile identificare automaticamente i valori anomali utilizzando la funzione *FSR*.

```
[outFSR]=FSR(y.^-0.5,X,'init',7);
```

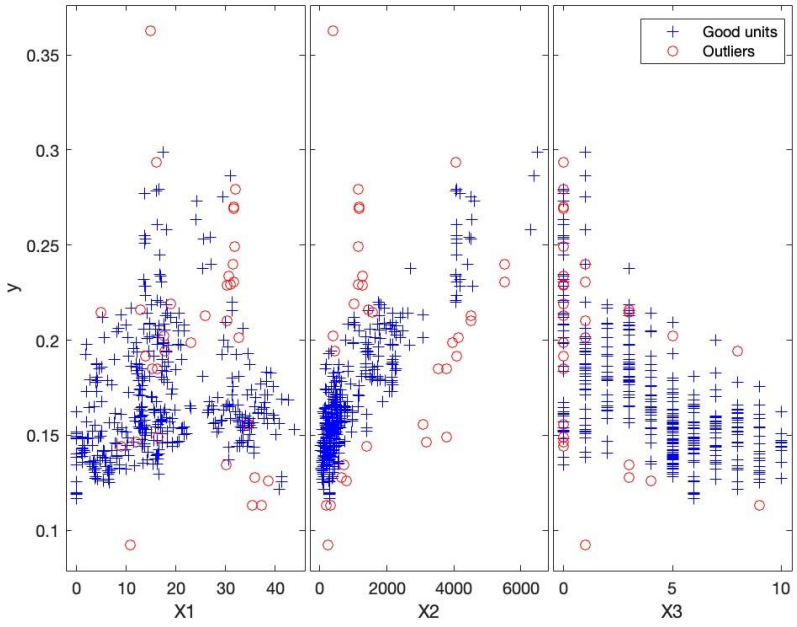


Figure 4-3: Identificazione degli outliers

La funzione evidenzia 33 valori anomali (*outliers*).

Tuttavia, per verificare se ci sono outliers in un modello di regressione, ti consiglio di non basarti solo sul grafico, ma di utilizzare anche le seguenti misure:

- *I punteggi di leva*: sono compresi tra 0 ed 1. Un punteggio elevato di leva è quindi un valore vicino ad 1.
- *I residui studentizzati*: si considerano valori elevati quelli maggiori di 3 o minori di -3.
- *La distanza di Cook*: si considerano elevati i valori superiori ad 1

4.2.4 *La variazione dei residui*

Tramite la funzione *Sreg* vengono calcolati gli stimatori in un modello di regressione lineare, i quali vengono analizzati dalla funzione residenxplot la quale traccia i residui di un'analisi di regressione rispetto qualsiasi variabile. Vengono definiti due diversi breakdown point (0.25 e 0.5). Essi sono i punti dopo i quali uno stimatore diventa inutile. Il breakdown point è una misura di robustezza; maggiore è il punto di rottura, migliore è lo stimatore.

```
conflev=[0.95 0.99];

figure;

h1=subplot(2,1,1);

bdp=0.25;

[out]=Sreg(y,X,'nsamp',3000,'bdp',bdp);
resindexplot(out,'h',h1,'conflev',conflev);
ylabel(['Breakdown point =' num2str(bdp)])
h2=subplot(2,1,2);
bdp=0.5;
[out]=Sreg(y,X,'nsamp',3000,'bdp',bdp);
resindexplot(out,'h',h2,'conflev',conflev);
ylabel(['Breakdown point =' num2str(bdp)])
cascade;
```

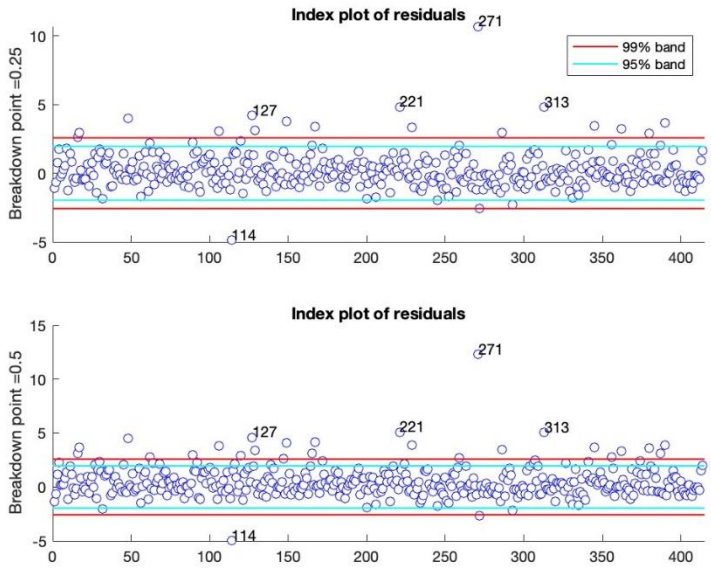


Figure 4-4: Identificazione degli outliers al variare del breakdown point

Tramite le impostazioni di Matlab è possibile selezionare i valori anomali dal grafico evidenziati dal numero di osservazione (127, 221, 313, 271, 114) e visualizzarli su un *yXplot* evidenziati in rosso.

```
bdp=0.5;
% due livelli di confidenza
conflev=[0.95 0.99];
[out]=Sreg(y,X,'nsamp',3000,'bdp',bdp,'yxsave',1);
resindexplot(out,'conflev',conflev,'databrush',1);
```

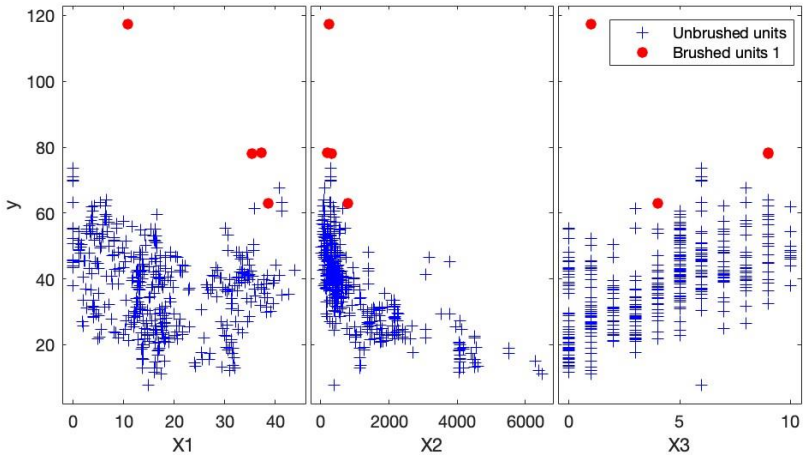


Figure 4-5: yXplot con l'opzione brush

Tramite la funzione *Sregeda* vengono calcolati gli stimatori in un modello di regressione lineare per una serie di valori di Breakdown point.

```
[out]=Sregeda(y,X,'nsamp',1000);

fground=struct;
sel=[127 221 313 271 114]';
fground.funit=sel;
fground.FontSize=1;

LineStyle=[ repmat({'-.'},6,1); repmat({'--'},9,1); repmat({':'},2,1)];
Color= [ repmat({'r'},6,1); repmat({'k'},9,1); repmat({'b'},2,1)];
fground.Color=Color;  % different colors for different foreground
trajectories
fground.LineWidth=3;
fground.LineStyle=LineStyle;

resfwdplot(out,'fground',fground);
```

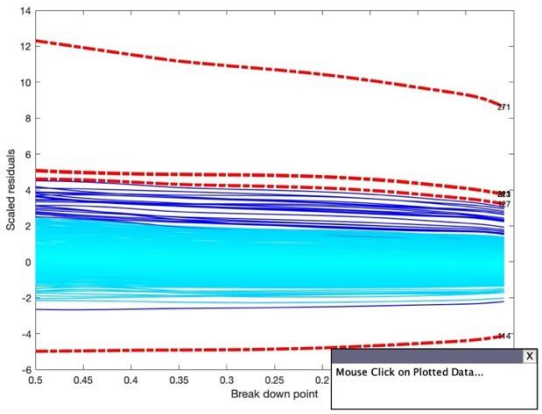


Figure 4-6: funzione Sregeda

Il grafico mostra come variano i residui al variare del breakdown point ed è possibile analizzare ogni residuo singolarmente in ogni punto. Molto utile è la funzione resfwdplot, la quale, grazie alle sue varie proprietà, permette di evidenziare l'ammontare di residui alla variazione di osservazioni comprese.

```
[out]=LXS(y,X,'nsamp',10000);
[out]=FSReda(y,X,out.bs);
out1=out;
% Calcolare i residui su base quadratica
out1.RES=out.RES.^2;

% plot minimum deletion residual with personalized options
mdrplot(out,'ylimy',[1 4.2],'xlimx',[10
60],'FontSize',14,'SizeAxesNum',14,'lwdenv',2);

% Persistent brushing on the plot of the scaled residuals. The plot is:
fground.flabstep='';        % without labels at steps 0 and n
fground.fthresh=3.5^2;      % threshold which defines the trajectories in
foreground
fground.LineWidth=1.5;      % personalised linewidth for trajectories in
foreground
fground.Color={'r'};        % personalised color (red lines) for
trajectories in foreground

databrush=struct;
databrush.bivarfit='';
databrush.selectionmode='Rect'; % Rectangular selection
databrush.persist='on'; % Enable repeated mouse selections
databrush.Label='on'; % Write labels of trajectories while selecting
databrush.RemoveLabels='off'; % Do not remove labels after selection
databrush.Pointer='hand'; % Hand cursor point while selecting
databrush.FlagSize='8'; % Size of the brushed points
databrush.RemoveTool='on'; % Remove yellow selection after finishing
brushing
resfwdplot(out1,'fground',fground,'databrush',databrush);
```

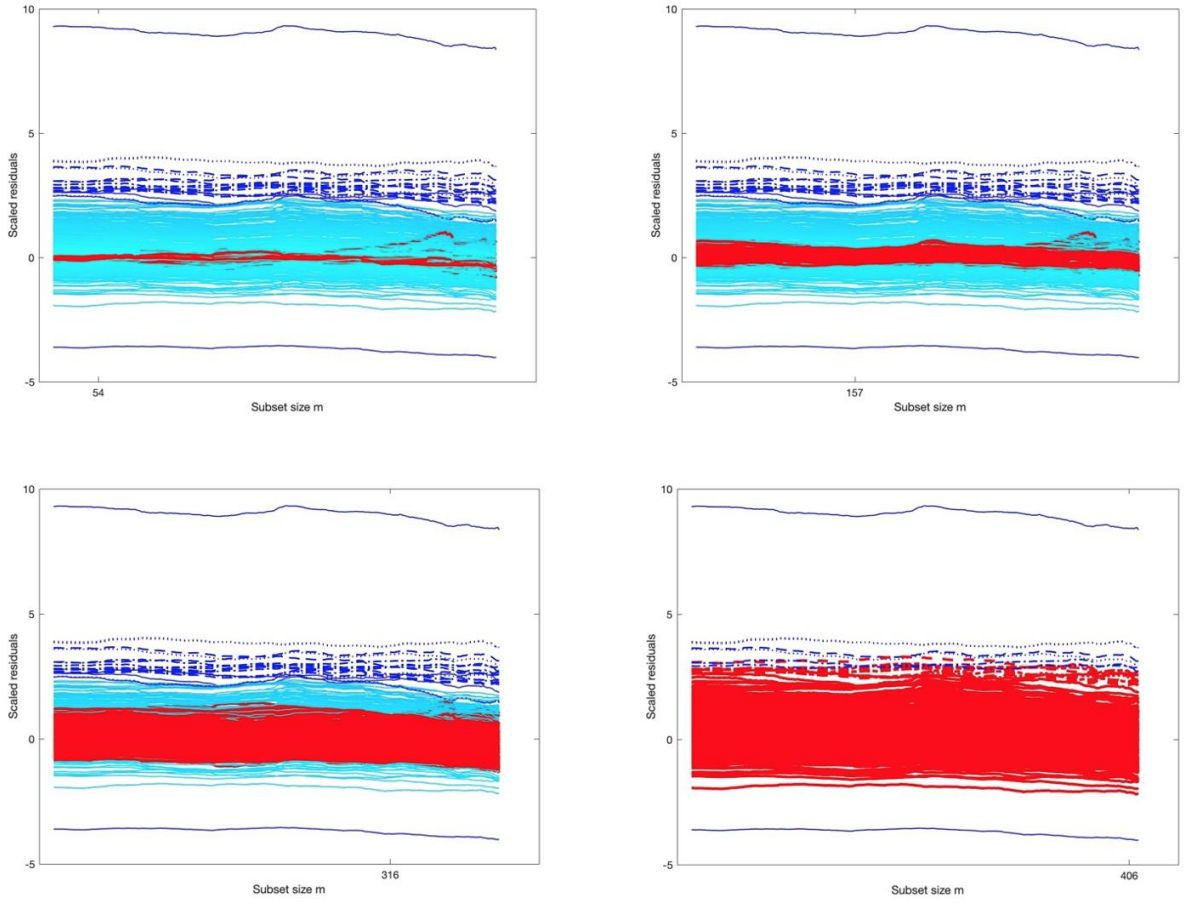


Figure 4-7: Variazione dei residui al variare delle unità nel campione

Tramite l'impostazione "*brushing*" è possibile selezionare una moltitudine di punti e visualizzarli in un *yXplot* che rappresenti la posizione dell'osservazione nel dataset. All'aumentare del numero di osservazioni prese in considerazione, la funzione FSRmdr calcola l'eliminazione minima residua e altre quantità di regressione lineare di base in ogni fase della ricerca.

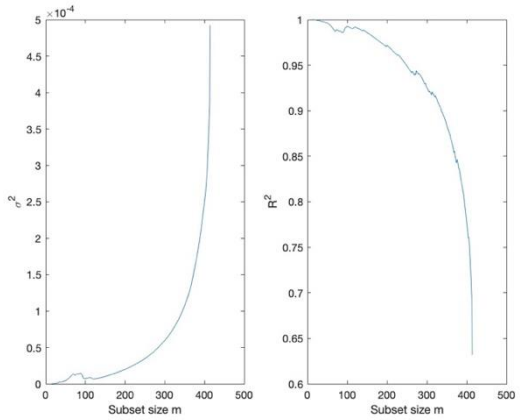


Figure 4-8: FSRmdr

BIBLIOGRAFIA

Milioli, M., Riani, M., Zani, S.,2014. La relazione lineare tra fenomeni. In: Milioli, M., Riani, M., Zani, a cura di *Introduzione all'analisi dei dati statistici*. Bologna: Pitagora, pp. 213-215

RINGRAZIAMENTI

In conclusione, questo elaborato sancisce il termine del mio percorso di laurea triennale e tengo a fare dei doverosi ringraziamenti.

In primis ringrazio il mio relatore, Prof. Marco Riani per avermi accompagnato, guidato e seguito con le sue conoscenze nello sviluppo di questo elaborato di ricerca, il quale spero possa avviarmi verso un migliore approfondimento nei miei prossimi studi magistrali.

Un grazie di cuore alla mia famiglia e soprattutto ai miei genitori che mi hanno sempre sostenuto durante tutto il percorso universitario non solo come genitori, ma anche come "amici" con cui potermi confidare.

Ringrazio la mia fidanzata Ioana per avermi supportato e sopportato in tutti i momenti di gioia e di difficoltà incontrati, con l'augurio che possa continuare a farlo anche in futuro così come, con amore, ha sempre fatto.

Voglio ringraziare particolarmente i miei coinquilini Michele, Stefano, Viola, Cristina ed Erika con i quali abbiamo condiviso ore di studio, quarantene, lockdown e cibi spazzatura, nonché tutti i ragazzi che, a Parma e non, hanno fatto parte di questo magnifico percorso senza i quali non sarebbe stato lo stesso: Francesco, Kristian, Micheal, Tommaso.

Per concludere voglio ringraziare tutte quelle persone che non ho potuto citare per motivi di spazio, ma a cui penso e mando un grande ringraziamento.

Luca Ratti