UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE

Tesi di laurea:

LA CARTA CUSUM TABULARE AGGIUSTATA PER LA REGRESSIONE CON PARAMETRI STIMATI

Relatore: Prof. Capizzi Giovanna

Laureando:Fokou Joel
Matricola n. 521564

Anno accademico 2006/2007

Alla mia famiglia, ma soprattutto a mia Mamma che ha
Sempre fatto l’impossibile per mandare avanti tutta la famiglia.

A mia nonna che mi ha insegnato tante delle virtù della vita.

INDICE:

4.1 Simulazioni

4.1.1 Introduzione

4.1.2 Impatto della numerosità campionaria sulla stima dei parametri

4.1.3 Effetti della stima dei parametri sull'ARL della CUSUMREG e applicazioni della carta di controllo

4.2- Conclusioni.

<u>CAPITOLO 1</u>: CONTROLLO STATISTICO DI PROCESSO

1.1-INTRODUZIONE: Perché il controllo statistico di processo?

Sopratutto in ambito aziendale, le esigenze di mercato(aumento dei volumi di produzione per soddisfare la domanda, conformità dei prodotti, etc……) hanno portato ad un'automazione dei metodi produttivi.
Questi processi produttivi possono essere soggetti ad una grande variabilità che a volte può incidere sulla qualità del prodotto finale. Variabilità che può essere: casuale, dovuta alla performance dei macchinari, alla diversa qualifica degli operatori che ci lavorano, e a molte altre cause…
La riduzione di quella variabilità è uno degli obiettivi

principali delle aziende e può essere ottenuta sia con un miglioramento del processo produttivo(miglioramento dei macchinari) che con un buon monitoraggio dei processi produttivi.(es: l'uso di carte di controllo per controllare le variabili del processo che ci interessano.)

Tuttavia, la variabilità presente in un processo produttivo non può essere completamente eliminata perché quella (variabilità) casuale non può essere controllata.

Comunque per ottenere un miglioramento nella qualità dei prodotti, la variabilità diversa da quella casuale deve essere eliminata.

Quindi il controllo del processo produttivo deve essere continuo. La statistica attraverso le carte di controllo ci offre un ottimo strumento per raggiungere quel obiettivo.

1.2 - **Le carte di controllo**

1.2.1- **Caratteristiche generali**

Una carta di controllo consiste nella rappresentazione grafica in funzione del tempo (t) di alcuni valori chiamati statistiche di controllo.

Le statistiche di controllo sono trasformazioni delle osservazioni delle variabili di interesse.

Una carta di controllo comprende quindi:

- Una statistica Z_t funzione delle osservazioni del processo da monitorare.
- Una linea centrale (CL) che rappresenta il valore medio della statistica per il processo in controllo.
- Un limite superiore (UCL) valore soglia oltre al quale è poco probabile che cadano valori della statistica di controllo se il processo è in controllo.
- Un limite inferiore (LCL) valore soglia al di sotto del quale è poco probabile che cadano valori della statistica di controllo se il processo è in controllo.

Una carta di controllo è sostanzialmente una serie di test ad istanti t del tipo:

$$Ho : Z_t \in (UCL : LCL)$$

$$H_1 : Z_t \notin (UCL : LCL$$

Appena un valore di Z_t cade fuori dai limiti di controllo, scatta l'allarme.

1.2.2- I parametri

Di solito i parametri con i quali vengono disegnate delle carte di controllo semplici sono:

- ***L***: distanza tra il valore medio del processo e uno dei limiti di controllo. Di solito scelto uguale a 3 in caso di normalità dei dati
- ***σ*:** scarto quadratico medio della varianza della variabile.
- ***μ*:** Media della variabile di cui si disegna la carta di controllo.

Le carte di controllo sono classificate in 2 gruppi:

A)-<u>Le carte di controllo con memoria:</u>

Queste carte sfruttano tutte le informazioni a disposizione fino all'istante t. Quindi avendo a disposizione il comportamento della statistica Z all' istante t, possiamo dare delle informazioni sul comportamento dei campioni a disposizione "fino" all' istante t (cioè anche agli istanti precedenti t-1, t-2,….).

Uno *shift* (**δ**) è l' unità per la quale moltiplicare σ per ottenere lo scostamento di un'osservazione dalla media del processo.

Le carte di controllo con memoria sono sensibili non solo a degli shift maggiori di 1.5, ma anche a quelli minori di 1.5 (compresi tra 0 e 1.5).

=> **$\delta > 1.5$ e $0 < \delta < 1.5$**

B)-<u>Le carte di controllo senza memoria:</u>

Queste carte sfruttano solo le informazioni a disposizione all' istante t.

Questa loro caratteristica fa si che le carte senza memoria sono adatte per individuare solo degli shift (scostamenti della media) maggiori di 1.5 => **$\delta > 1.5$**

Sia le carte con memoria che quelle senza possono essere per **attributi** oppure per **variabili.**

a)-Le carte di controllo per attributi:

Queste carte si usano se le variabili che stiamo studiando sono delle variabili discrete (possono risultare da una classificazione es: numero di soggetti conformi o non conformi relativamente ad una caratteristica precisa di un oggetto prescelto).

b)-Le carte di controllo per variabili:

Queste carte si usano se le variabili soggette allo studio sono delle variabili continue (misurabili su scala numerica).

1.2.3 - Interpretazione

L' obiettivo delle carte di controllo è di rappresentare le statistiche di controllo in modo da individuare i loro

valori anomali.

Valori anomali della statistica di controllo implicano valori anomali della variabile sotto esame.

Un valore anomalo della statistica di controllo può essere: un valore che cade sopra UCL, un valore che cade sotto LCL oppure valori che indicano uno scostamento del valore centrale della media delle Z_t. Tali valori anomali ci vengono segnalati con l'attivazione di un allarme.

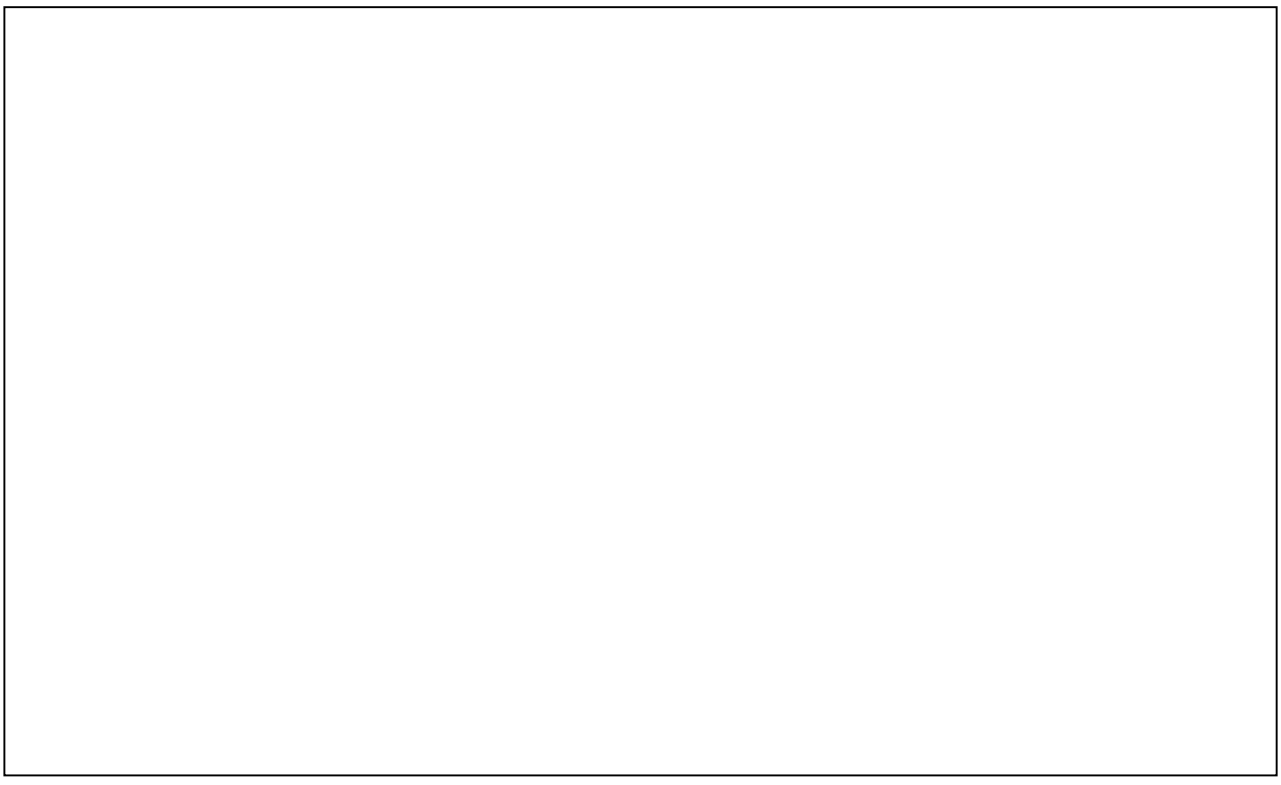
Ovviamente, nessun processo essendo perfetto, oltre ai momenti in cui le carte ci daranno delle informazioni esatte (che rifletteranno il comportamento esatto della variabile :sotto controllo o fuori controllo), può capitare che le carte ci diano delle informazioni sbagliate. Cioè che scatti l'allarme quando il processo è sotto controllo (errore del 1° tipo) oppure che non scatti l'allarme quando il processo non lo è più (errore del 2° tipo).

Nei casi più semplici, diremo che una variabile è sotto controllo se le sue osservazioni cadono dentro i limiti di controllo. Poi ci sono delle carte più complesse come quella CUSUM dove oltre alla condizione già richiesta, dobbiamo osservare un andamento del tipo "*RANDOM-WALK*"

(passeggiata casuale) intorno alla linea centrale della carta di controllo dei valori di Z_t e non ci devono essere 6 valori consecutivi di Z_t sopra o sotto la linea centrale per poter dire che la variabile è sotto controllo.

Vediamo sul grafico un semplice esempio di carta di controllo in cui le osservazioni sono rappresentate dai punti neri. La carta è stata disegnata con parametri: L=3, $\boldsymbol{\sigma}$=1, $\boldsymbol{\mu}$=20.

Dall' istante 1 all'istante 6, la variabile è in controllo perché i valori della statistica cadono dentro i limiti di controllo LCL=17 e UCL=23. Dall' istante 7 in poi, la variabile è fuori controllo.



1.3- La carta CUSUM

1.3.1 Generalità

La carta CUSUM includendo informazioni di più campioni consecutivi, incorpora tutte le informazioni disponibili dall'inizio del processo fino all'istante (t) di riferimento.
Questa carta è molto efficace quando la dimensione campionaria è molto bassa, soprattutto se è unitaria(n=1) per quello è un ottimo strumento di controllo per i processi chimici e tutti i processi dove sono fatte misurazioni su singoli pezzi di prodotti.
Esistono diversi modi di rappresentare le carte CUSUM:

- Le CUSUM tabulari o ALGORITMICHE.
- La carta CUSUM bilaterale standardizzata.
- La maschera a V per le CUSUM.

Noi ci limiteremo allo studio delle proprietà della carta CUSUM tabulare per la regressione quando i parametri sono stimati.

Avendo a disposizione campioni di dimensione $n \geq 1$, calcoliamo la media $\overline{Xj}$ per ogni campione. Supponendo che la vera media del processo (μ_o) sia nota.

La carta di controllo CUSUM è costruita considerando la somma cumulata fino all'i-esimo campione delle $\overline{Xj}$.

Cioè la quantità:

$$C_i = \sum_{i=1,\ldots,n} (X_j - \mu_o) = (X_i - \mu_o) + \sum (X_j - \mu_o) = (X_i - \mu_o) + C_{i-1}$$

La carta CUSUM tabulare è la forma più usata nella pratica ed è basata sulle statistiche $C+$ e $C-$ rispettivamente CUSUM unilaterale superiore e CUSUM unilaterale inferiore che vengono calcolate secondo le formule:

$$C_i^+ = max[0, X_i - (\mu_o + K) + C^+_{i-1}]$$

$$C_i^- = max[0, (\mu_o - K) - X_i + C^-_{i-1}]$$

dove per definizione $C_o^+ = C_o^- = 0$

Se decidiamo di esprimere gli scostamenti dalla media in unità di σ, avremo: $\mu_1 = \mu_o + \delta\sigma$.

Ci sono due parametri molto importanti che entrano in gioco nella fase del disegno della carta CUSUM: K e H

- La quantità **K=kσ** è solitamente detta valore di tolleranza ed è di norma pari alla metà dello scostamento tra il valore obiettivo μ_o e il valore μ_1 assunto quando osserviamo un fuori controllo.

Quindi : $$K = (\frac{\delta}{2})\sigma = \frac{|\mu_1 - \mu_o|}{2}$$

È da notare che le statistiche C^+ e C^- accumulano solo le deviazioni del valore obiettivo μo di ampiezza superiore a K.

- H è un valore relativo all'intervallo di decisione da considerare che serve per definire i limiti di controllo.

Quindi : **UCL=-LCL=H=h** σ

L'*Average Run Length* (***ARL***) è il numero medio di osservazioni prima di osservare un allarme.

Per la progettazione della carta di controllo, è consigliato scegliere opportunamente il valore di riferimento di **k** e dell'intervallo di decisione **h** in modo da ottenere buone prestazioni in termini di **ARL**. Molti studi teorici sono stati svolti a proposito e suggeriscono alcuni criteri in base ai quali viene effettuata la scelta di **k** e di **h**.

Comunque la scelta di **k** e di **h** dipende da cosa aspettiamo della nostra carta di controllo in termini di prestazioni.

1.3.2- *Run-length* e *Average Run Length* di una carta CUSUM

Per calcolare la performance di una carta di controllo, si possono usare indici come la *Run-Length(***RL***)* e l'*Average Run Length(***ARL***).*

L'**RL** è il primo istante in cui si osserva un fuori controllo.

L'**ARL** è il numero medio di osservazioni prima di osservare un allarme.

Si distinguono:

- L'ARL_0 : numero medio di osservazioni prima di osservare un falso allarme (va massimizzato).
- L'ARL_1 : numero medio di osservazioni prima di osservare un vero allarme (va minimizzato).

L'ARL complessivo di una carta CUSUM bilaterale può essere ricavato a partire delle ARL delle statistiche C_i^+ (ARL^+) e C_i^- (ARL^-) con la seguente formula:

$$\frac{1}{ARL} = \frac{1}{ARL^+} + \frac{1}{ARL^-}$$

In fase di progettazione della carta, dobbiamo fissare un valore di ARL_0 e dunque minimizzare ARL_1 per l' individuazione di un salto di media pari a δ.

CAPITOLO 2: *REGRESSIONE E STIMA DEI PARAMETRI*

2.1- Il modello di regressione

Supponiamo di avere a disposizione un numero n di dati relativi alle variabili Y_i e X_{ij} (i=1,..,n; j=1,..,m) con le variabili X ed Y correlate , un modello è una relazione che ci consente di prevedere i valori della variabile risposta Y_i in funzione delle variabili esplicative $X_{i,j}$.

In un modello di regressione è presente anche una componente erratica ε con media nulla e varianza costante ignota perché non sappiamo di quanto il modello sottostimi o sovrastimi il valore vero.

Le variabili esplicative possono essere di due tipi:

- qualitative .
- quantitative.

Ma per l'aspetto da noi trattato, tutte le variabili saranno quantitative visto che nei processi produttivi sono quelle più controllate.
Esistono diverse tipologie di modelli di regressione, ma noi tratteremo e useremo per semplicità il modello di regressione lineare semplice:

$Y_i=\beta_0+\beta_1 X_1+ \beta_2 X_2+ .....+ \beta_j X_{j+}\varepsilon_i$ oppure $Y=f(X, \beta)+ \varepsilon$

Questo modello si dice lineare perché l'equazione f(X, β) è lineare nei suoi parametri.
Quello suppone che i dati di Y se estrapolati bene, dovranno avere un andamento lineare seppure alcuni scostamenti dalla retta possono essere comunque osservati a causa dell'errore ε.

2.2- I parametri e la loro stima.

I parametri che sono incognite sono βj (coefficienti delle variabili esplicative) e la varianza della componente erratica σ^2.

Riusciamo a stimare i parametri βj a partire da valori storici

di (X_{ij},Y_i) con la seguente formula:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Si assume in questo caso che l' errore ε_i abbia una distribuzione normale con media nulla e varianza costante non nota σ^2 cioè

$$\varepsilon_i \sim N(0,\sigma^2) \quad i=1,\ldots\ldots\ldots,n$$

Siccome σ^2 è ignoto, lo dobbiamo stimare. Anche se non tratteremo molto della stima di σ perché il nostro studio si focalizzerà sul monitoraggio della media, notiamo comunque che è ottenuto con la formula:

$$S^2 = \frac{(\|Y\|^2 - \|\hat{Y}\|^2)}{(n-j)}$$

Per quello che li riguarda, i residui, sono ottenuti dalla formula:

$$e = Y - \hat{Y}$$

La qualità della stima dei parametri dipende molto dalla numerosità campionaria n.

Questo punto sarà discusso nel capitolo successivo.

CAPITOLO 3:

CARTE DI CONTROLLO AGGIUSTATA PER LA REGRESSIONE: LA CUSUMREG

3.1- **Introduzione al concetto di carta di controllo per la regressione**

Inizialmente proposte da MENDEL nel 1969, le carte di controllo aggiustate per la regressione sono un effettivo strumento di controllo statistico di processo.
Il suo concetto base è di rimuovere gli effetti della correlazione esistente tra la variabile risposta e le variabili esplicative usando l'aggiustamento per regressione e poi disegnare la carta di controllo sui residui di regressione.

Questo concetto è stato poi sviluppato da diversi autori come HAWKINS (1991).
In pratica, il modello di regressione che spiega la variabile risposta in funzione di quella esplicativa è raramente conosciuto e quindi deve essere stimato.
In questo capitolo, vedremo che impatto avrà la stima dei parametri del modello di regressione sulla *Run-Length* e sull'*Average Run Length* della variabile risposta del processo.

3.2 - La carta CUSUMREG

3.2.1- Introduzione

In un processo produttivo dove la variabile risposta (*output*) e la variabile esplicativa (*input*) sono correlate, quando l'input assume un valore elevato, l'output può sembrare fuori controllo quando non lo è. Per questo motivo, conviene controllare i residui di regressione e non direttamente l' *output.*

Per il controllo dei residui di regressione, possiamo usare la carta SHEWHART, la carta EWMA , la carta CUSUM.
Noi useremo la carta CUSUM, e definiamo quindi la carta CUSUMREG come la carta CUSUM aggiustata per la regressione.
La carta CUSUMREG quasi le stesse caratteristiche di una carta CUSUM normale. La sua particolarità è che le statistiche di controllo rappresentati (Z_t) sono dei residui di regressione e la carta di controllo viene progettata in 2 fasi:
1°fase: Stima di un modello lineare spiegando le Y_i a partire dalle X_i usando osservazioni storiche delle (X_i, Y_i). Questa fase implica la stima dei parametri del modello di regressione.
2°fase: Progettazione vera e propria della carta di controllo.

3.2.2- I parametri di regressione e le loro stime

Nella fase di regressione, faremo le seguenti assunzioni:

- La variabile risposta e quindi l'errore hanno distribuzione normale.

- La variabile risposta ha varianza costante.
- La relazione che lega la media della variabile risposta e la variabile esplicativa è lineare nei parametri.

Noi ci limiteremo a studiare la carta di controllo sulla media ed il modello di regressione che useremo sarà il modello lineare semplice:

$Y_i=\beta_0+\beta_1 X_1+\varepsilon_i \qquad i=1,\ldots,n$

Dalla formula generale: $\hat{\beta} = (X^T X)^{-1} X^T Y$ ricaviamo le stime di β_0 e di β_1

Per convenienza, useremo una trasformazione dell'espressione del modello lineare nella quale $\hat{\beta}'_0$ e $\hat{\beta}_1$ non sono correlati.

$$Y_i=\beta'_0+\beta_1(X_i-\bar{X})+\varepsilon_i$$

con $\beta'_0=(\beta_0+\beta_1\times(\bar{X}))$

La stima dei minimi quadrati di β'_0 e β_1 nell'equazione scelta

È :

$$\hat{\beta}'_o = \bar{Y} \quad => \quad \hat{\beta}'_o \sim N(\beta'_{0,} \frac{\sigma^2}{n})$$

$$\hat{\beta}_1 \sim N(\beta_{1,} \frac{\sigma^2}{\sqrt{var(X)}})$$

per quanto riguarda σ^2 la sua stima ci viene data da

$$\hat{\sigma} = \sqrt{MSE}$$

dove $$MSE = \frac{\sum(e_i^2)}{(n-2)}$$

Ed i residui che ci interessano particolarmente perché su di loro saranno disegnate le carte di controllo sono dati da:

$$e_t = Y_t - [\hat{\beta}'_o + \hat{\beta}_1 (X_t - \bar{X})] \sim N(0, \hat{\sigma}^2)$$

che possono essere standardizzati => $\frac{e_t}{\hat{\sigma}} \sim N(0, 1)$.

La fase successiva alla stima dei parametri è il disegno della carta di controllo. Però facciamo prima un approfondimento sulle implicazioni di questa stima.

La qualità della stima dei parametri dipende molto dalla numerosità del campione che abbiamo a disposizione ed ha

lo stesso effetto sia sulla distribuzione della *RUN-LENGTH* che sull'*ARL*.

3.2.3- Impatto della numerosità campionaria sulla stima dei parametri .

Nella prima fase, trattiamo un problema di regressione e come per tutti i problemi di regressione, la numerosità campionaria è un fattore fondamentale per la qualità della stima dei parametri.
La variabilità *Between* è la variabilità all'interno di un gruppo. Se generiamo 10000 volte dei dati di numerosità campionaria n, e calcoliamo per ogni campione l'ARL, otterremo un gruppo di ARL di 10000 valori. Vedremo, ed è da lì che inizieremo il capitolo delle simulazioni che più è importante la numerosità campionaria n, più è accurata la stima dei parametri perché la variabilità *between* dell' ARL è piccola.
Quindi la numerosità campionaria avrà un impatto considerevole sulla qualità della stima dei parametri.

3.2.4-Effetti della stima dei parametri sull' RL e Sull'ARL della CUSUMREG

a) Stima individuale:

Supponiamo di sapere qual è il valore vero dei parametri β'_0 e σ, possiamo benissimo stimare β_1.

Questa stima individuale di β_1 genera più falsi allarmi rispetto a quando questo parametro è noto indipendentemente dal fatto che abbiamo sovrastimato o sottostimato il parametro β_1. Questo riduce il valore dell'ARLo al quale ci dobbiamo aspettare dalla scelta di k e h. Gli effetti sarebbero stati gli stessi se al posto di β'_0 avessimo avuto i veri valori di β_1 e σ ed avessimo stimato β'_0 .

Per quanto riguarda il parametro σ , una sua sottostima genera più falsi allarmi. Mentre una sua sovrastima genera meno allarmi.

b)- Stima combinata di più parametri

La stima combinata di tutti e tre i parametri è più difficile da generalizzare a causa della dipendenza della

distribuzione dell' RL (e quindi di dell'ARL) dai tre parametri.

Se sovrastimiamo σ, l'effetto è meno chiaro sulle carte di controllo comunque l'osservazione di più o meno falsi allarmi dipende dall'errore di stima di β'_0 e β_1. Mentre una sottostima di σ implica un aumento del numero di falsi allarmi non importa se sottostimiamo o sovrastimiamo β'_0 e β_1 .

In linea generale, noi stimiamo i parametri perché non sappiamo qual è il loro vero valore. Non siamo quindi in grado di dire quanto sovrastimiamo o sottostimiamo i parametri. Ed è questo errore casuale che crea una distorsione nella distribuzione della RL.

Gli effetti della stima di μ e σ in una carta CUSUM sulla distribuzione della *Run-Length* sono gli stessi di quelli della stima di β'_0 e σ in CUSUMREG.

3.3- Tipologie di scostamenti dalla media

Nella regressione lineare classica, assumiamo che i regressori siano non casuali. Perché se lo fossero, la regressione sarebbe considerata come condizionatà al valore del regressore.

Come già accennato prima, noi ci limiteremo a disegnare la carta di controllo sulla media.

Nel caso del controllo della media, due tipi di shift o scostamenti dalla media sono da considerare:

- Un cambiamento medio delle X
- Un cambiamento medio della Y condizionato da X.

Dobbiamo notare che entrambi i cambiamenti in β'_0 e β_1 possono risultare in un cambiamento di **E(Y/X)**.

Ma visto che un cambiamento in β_1 influenza la variabilità , e che ci sono delle procedure standard della CUSUM per il controllo della variabilità di un processo, noi concentreremo l'attenzione sulla performance di una carta

CUSUM per degli shift in E(Y/X) dovuti a cambiamenti nell'intercetta β'_0.

Se μ_x e σ^2_x sono rispettivamente i valori della media e della varianza di X quando il processo è in controllo, allora i residui per un futuro valore di X sono dati da:

$$e_t = (\beta'_o - \hat{\beta}'_o) + (\beta_1 - \hat{\beta}_1)(X_t - \overline{X}) + \varepsilon_t$$

quando c'è uno shift di dimensione **aσ** in E(Y/X) e uno shift di dimensione $\mathbf{b\sigma_x}$ in E(X), i residui condizionati ai valori di $\hat{\beta}'_o$, $\hat{\beta}_1$, e $\overline{X}$ sono normalmente distribuiti con media :

$$E(e_t) = a\sigma + (\beta'_o - \hat{\beta}'_o) + (\beta_1 - \hat{\beta}_1)(\mu_x + b\sigma_x - \overline{X})$$

e varianza:

$$V(e_t) = (\beta_1 - \hat{\beta}_1)^2 \sigma^2_x + \sigma^2$$

L'obbiettivo dell' aggiustamento per regressione è di creare uno strumento per la diagnosi più sensibile rimuovendo gli effetti delle perturbazioni presenti a monte.

Dall'equazione del valore medio dei residui,

viene:

$$E(e_t) = a\sigma + (\beta'_o - \hat{\beta}'_o) + (\beta_1 - \hat{\beta}_1)(\mu_x - \overline{X}) + b(\beta_1 - \hat{\beta}_1)\sigma_x$$

Si può vedere che se b è diverso da zero, allora osserviamo uno scostamento dalla media dei residui pari a:

$$b(\beta_1 - \hat{\beta}_1)\sigma_x$$

Questo è uno degli effetti indesiderati, perché idealmente, i residui dovrebbero avere media nulla.

Dato che **σ** è diverso da 0, anche un valore non nullo di **a** genera uno spostamento nella media dei residui.

In uno shift del tipo **aσ** può anche risultare uno shift del tipo **b.** Quindi valori diversi di **a** e **b** avranno quasi gli stessi effetti sui residui perché tutti e due generano uno spostamento della loro media.

CAPITOLO 4: CARTA CUSUMREG APPLICAZIONI.

4.1 Simulazioni e applicazione della CUSUMREG.

4.1.1-Introduzione

In questo capitolo, noi dimostreremo con delle simulazioni alcune affermazioni fatte nei capitoli precedenti. Dimostreremo per primo che la numerosità campionaria influenza la qualità della stima dei parametri. Poi faremo delle simulazioni per vedere quanto detto sugli effetti della stima dei parametri sull'RL di una carta CUSUMREG. E per finire, disegneremo la carta CUSUMREG per il vettore Y di dati che abbiamo a disposizione.

4.1.2-Impatto della numerosità campionaria sulla stima dei parametri

Per studiare l'impatto della numerosità campionaria, faremo delle simulazioni con diverse numerosità campionarie.

La procedura da seguire è la seguente:

i)-Generiamo un vettori X con numerosità campionaria $\mathbf{n_1=50}$, poi $\mathbf{n_2=100}$ e finalmente $\mathbf{n_3=500}$ di media 0 e varianza unitaria. Poi generiamo un vettore E con gli stessi criteri con i quali abbiamo generato X.

ii)-Per ogni numerosità campionaria $\mathbf{n_i}$, calcoliamo il vettore Y=20+2.5*X+E e stimiamo il modello lineare che lega X e Y.
Useremo quindi come valori veri dei parametri di stima:
Beta_0=20 **Beta_1=2.5**

iii)-Ripetiamo **N=10000** volte i passi i) e ii) per rassicurarci della consistenza dei risultati che otterremo.

iv)- Calcoliamo il valore medio e la varianza dei vettori di 10000 valori $\hat{\beta}_o$ e $\hat{\beta}_1$ ottenuti.
Tutti i passi sopra elencati, li facciamo con la funzione STIMA(n_i,N)

Applicazione:

per n=50

Beta0_hat: 19.99592 Varianza: 0.02074769

Beta1_hat: 2.50111 Varianza: 0.02143349

per n=100

Beta0_hat: 20.00085 Varianza: 0.009957442

Beta1_hat: 2.499580 Varianza: 0.01044286

per n=500

Beta0_hat: 20.00009 Varianza: 0.001992644

Beta1_hat: 2.49944 Varianza: 0.0020069

Osserviamo come all'aumentare del valore n (numerosità campionaria), le stime sono molto più vicine ai valori veri sia di Beta0 che di Beta1. Anche la variabilità dei valori osservati dai parametri nei 10000 campioni diventa più piccola più n è grande.

Quindi per degli studi di regressione, conviene sempre nella misura del possibile, usare delle numerosità campionarie elevate per essere vicino ai valori veri dei parametri.

4.1.3- Effetti della stima dei parametri sull'ARL della CUSUMREG e applicazioni della carta di controllo

Noi non faremo delle simulazioni per dimostrare gli effetti della stima di σ e terremo per buone le affermazioni fatte nel capitolo precedente su quel parametro. Inoltre esistono delle procedure standard per il controllo della variabilità di un processo.
Per le simulazioni che faremo, useremo la numerosità campionaria che ci ha dato dei valori soddisfacenti cioè n=500. Useremo la seguente procedura per studiare l'effetto della stima dei parametri sull'ARL della CUSUMREG

i)- Abbiamo due vettori X ed Y con numerosità 500 e correlati. In controllo, la variabile Y ha media 20 e varianza unitaria. X invece ha media nulla e varianza unitaria.

ii)-Generiamo un vettore E di 500 dati da una normale di media 0 e varianza 1. E stimiamo i parametri del modello lineare legando X a Y.

iii)-In realtà, sappiamo come sono stati generati i vettori X ed Y, quindi sappiamo quali sono i valori veri dei parametri anche se ci abbiamo aggiunto dei valori anomali. Usiamo i valori veri di $\hat{\beta}_1$ e σ e poi il valore stimato di $\hat{\beta}'_o$ per calcolare i residui della regressione. Calcoliamo la probabilità di fuori controllo e l'ARL del processo.

Per fare queste simulazioni, usiamo la funzione: **CUSUMREG** che potrete vedere nell'appendice e di cui useremo solo l'outpu numerico e non il grafico per commentare i risultati. Visto che vogliamo avere un ARLo pari a 370, la tabella di HAWKINS ci suggerisce di usare valori di k=0.5 e h=4.77.

- Per la stima di β'_o abbiamo ottenuto i risultati seguenti:

Prob fuori controllo: 0.302

ARL del processo: 203

- Per la stima di $\hat{\beta}_1$ abbiamo ottenuto come risultato:

Prob fuori controllo: 0.22

ARL del processo: 203

Vediamo che la probabilità di fuori controllo è abbastanza alta in entrambi i casi. L'ARLo del processo è abbastanza bassa rispetto ai 370 ai quali ci aspettavamo vista la scelta di k e di h. L'allarme si attiva allo stesso istante t=203 in entrambi i casi.

- Per la stima di entrambi i parametri β_o' e $\hat{\beta}_1$ abbiamo ottenuto come risultato:

Prob fuori controllo: 0.222

ARL del processo: 203

Anche in questo caso il valore dell'ARL è piccolo rispetto ai 370 fissati e l'allarme si attiva dopo all'istante t=203.

Adesso, disegniamo la carta CUSUMREG per il vettore Y che abbiamo.

Verifica delle assunzioni per le quali possiamo applicare la regressione e per conseguenza la CARTA CUSUMREG:

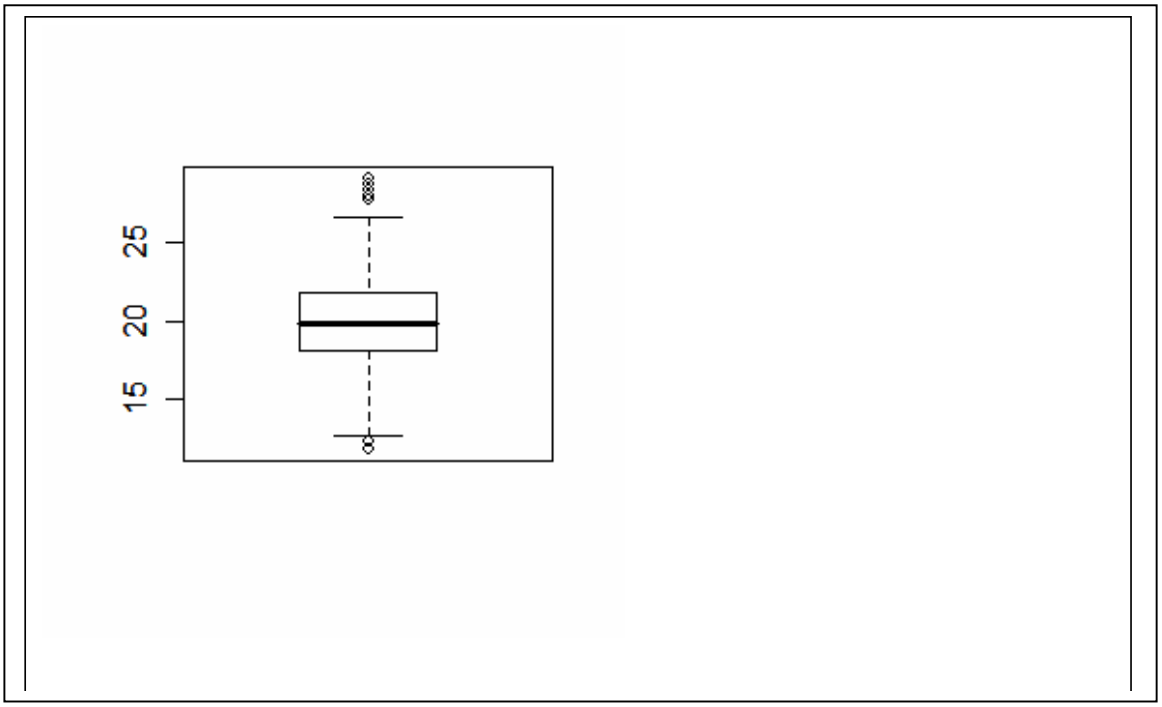
- Normalità della variabile risposta.
- Esistenza di una relazione lineare tra variabile risposta e variabili esplicative.
- Omoschedasticità della variabile risposta

(che è verifica per costruzione dei dati relativi al vettore Y)

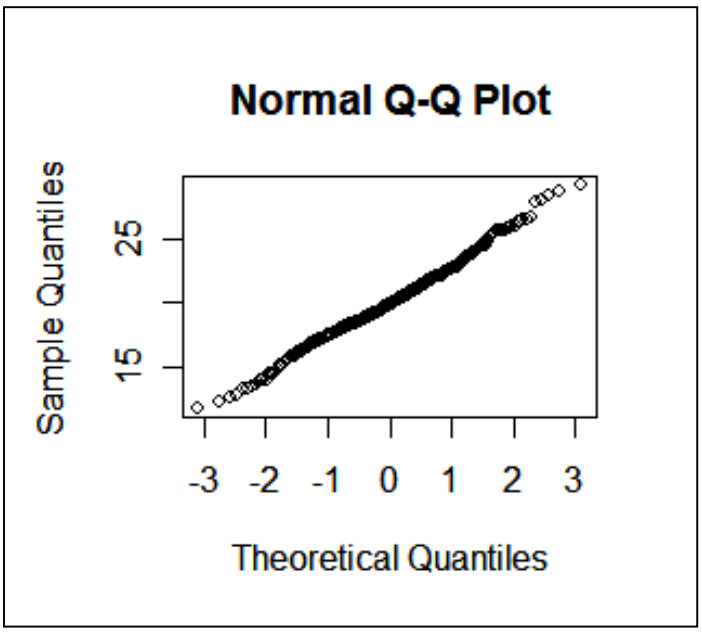
<u>Analisi esplorativa e verifica delle assunzioni:</u>

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
11.79	18.14	19.80	19.97	21.81	29.12

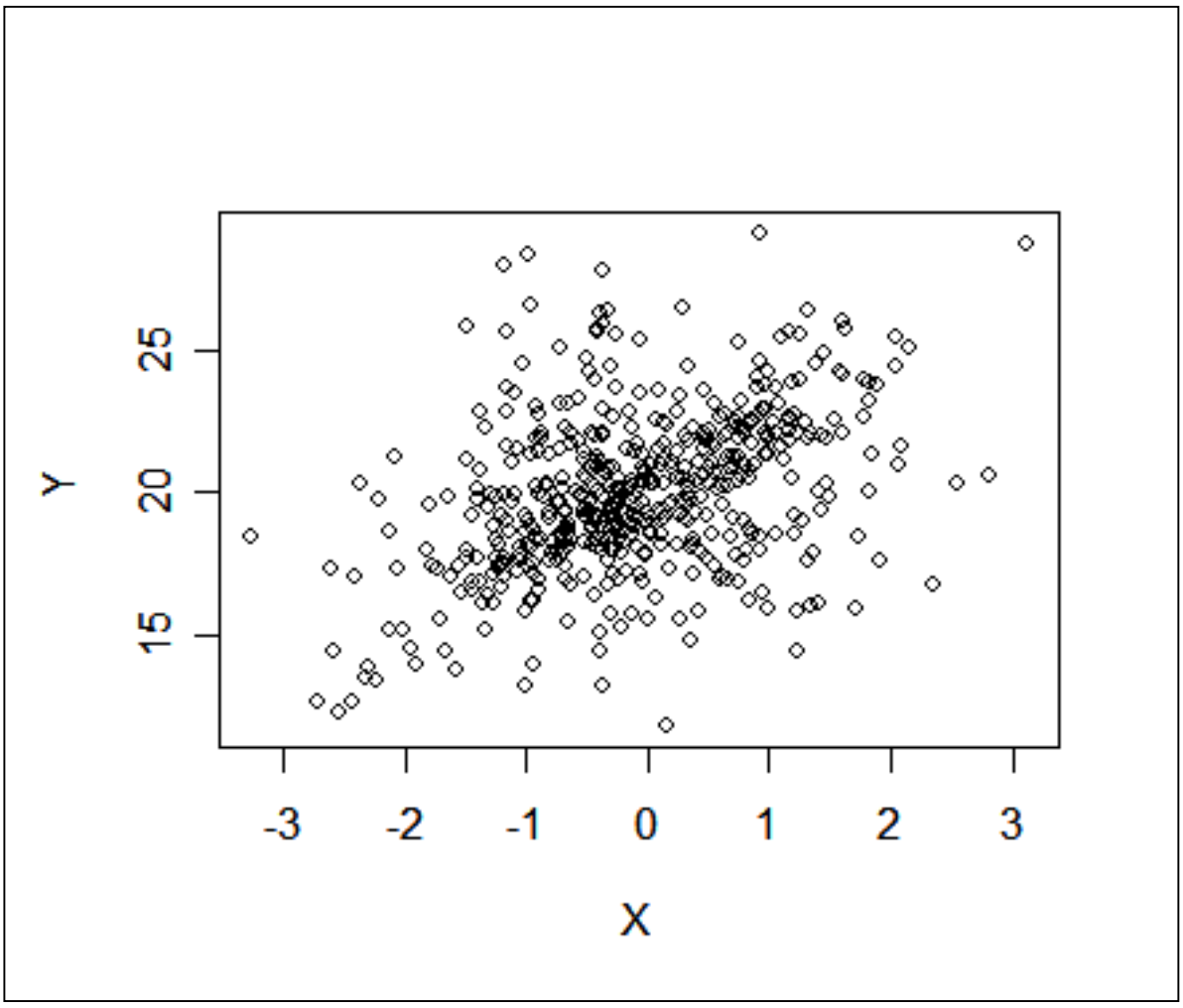
I dati sembrano avere media 20



Possiamo notare la simmetria dei dati. In effetti, le code hanno la stessa lunghezza e la scatola sembra presentare dei valori che oscillano intorno alla media del processo 20. Dobbiamo anche notare la presenza dei valori anomali sia sopra che sotto.



I dati del vettore Y sembrano essere normali.



La relazione esistente tra X e Y può essere considerata lineare visto l'andamento dei punti nello schema.

Quindi usiamo il modello lineare.

stima dei parametri:

Call:

lm(formula = Y ~ X)

Coefficients:

(Intercept)	X
20.085	1.144

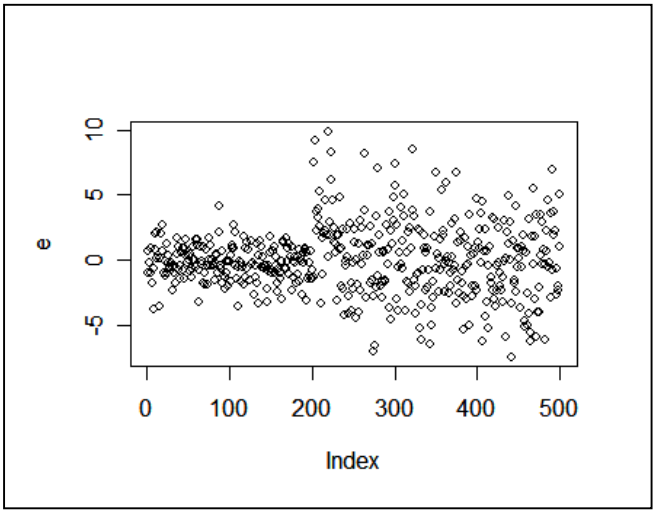
Abbiamo ottenuto come valore dei parametri: β'_0=20.085

e β_1 =1.14

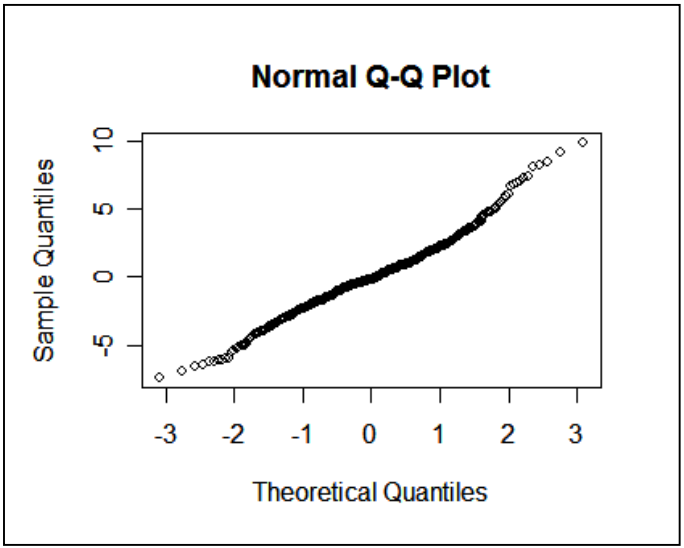
<u>Calcolo dei residui:</u>

con la formula $e = Y - \hat{Y}$

Calcoliamo i residui e ne facciamo il grafico



Vediamo se i residui hanno una distribuzione normale.



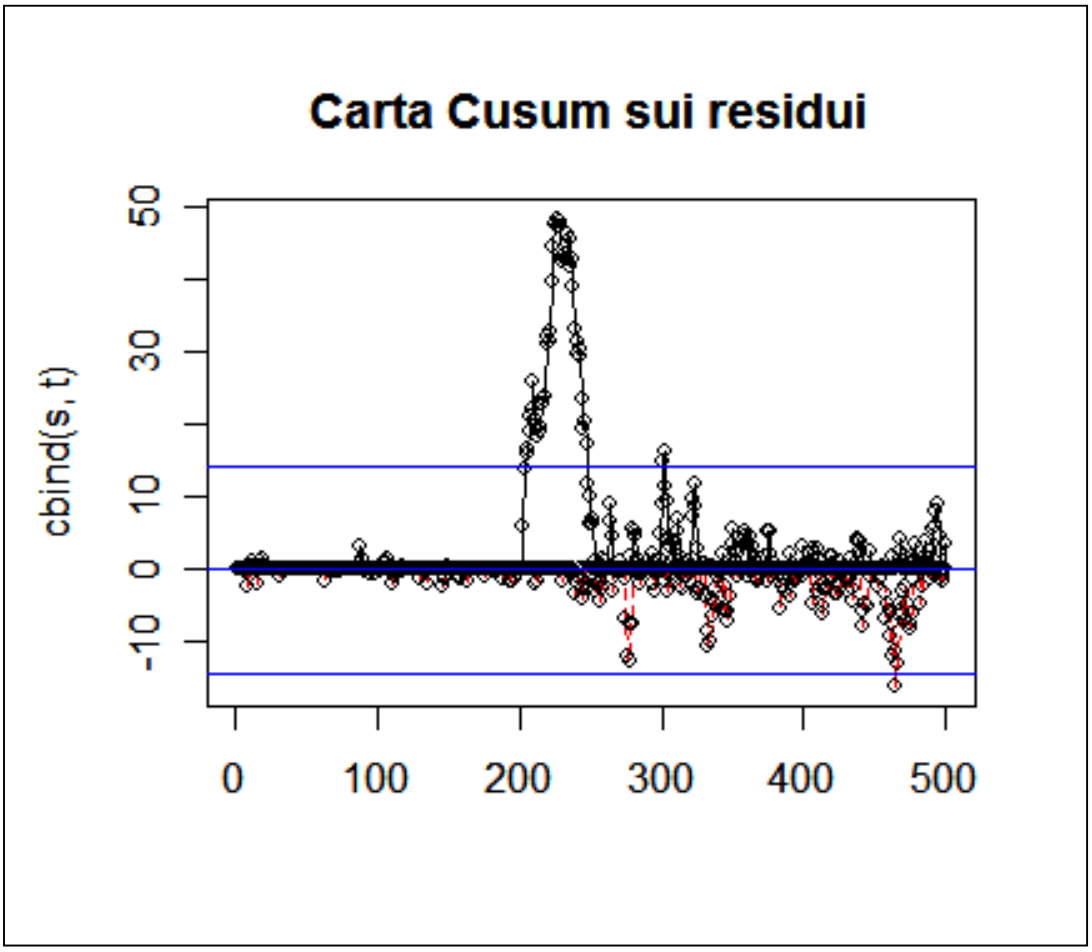
L’andamento dei valori del grafico ci consentono di assumere la normalità dei residui.

Facciamo il grafico della CUSUMREG e calcoliamo la probabilità di fuori controllo e la *Run-Lenght* con la funzione CUSUMREG.

Risultati:

Prob fuori controllo: 0.222

ARL del processo: 203



La carta di controllo mi segnala che il 22,2% dei dati sono fuori controllo.
Il primo fuori controllo viene segnalato all'istante t=203. E infatti, dall' istante t=201 a t=225 i dati X sono stati generati con una media più alta. Questo fatto genera l'aumento del valore medio dei valori di Y e quindi dei residui.
Vediamo quando risulta considerevole il salto sui residui.
La media dei residui compresi tra 201 e 225 è di 3.19 che è abbastanza grande come valore.
[1] 3.191124
Dalle nostre analisi sappiamo che questo può essere dovuto al fatto che su quei campioni, sono presenti salti sia nella media di X che in β'_0. E infatti, c'è un salto in media di X per costruzione dei dati.
Gli ultimi 25 dati sono stati ottenuti con valori normali di X, ma con un salto risultante sul valore di β'_0 e questo si nota anche con dei valori alti dei residui e uno scostamento del valore medio dei residui.
Vediamo infatti quanto è la media di e per gli ultimi 25 valori dell'errore.

[1] 0.7284854

Se proviamo a togliere questi valori che creano uno scostamento considerevole della media dei residui, otteniamo:

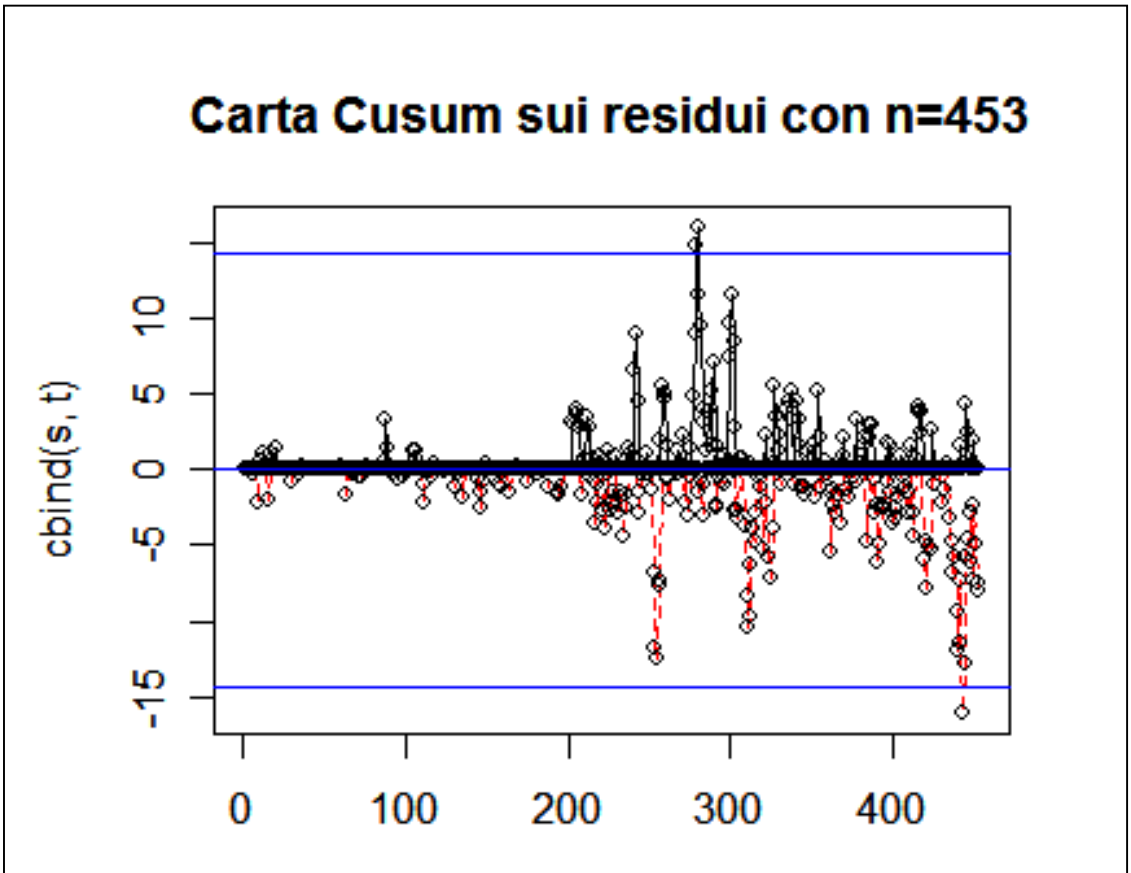
-Prob fuori controllo: 0.1236203

-ARL del processo: 241

Il primo fuori controllo ci viene segnalato questa volta al 241-esimo campione. E ovviamente la probabilità di fuori controllo diminuisce.

Siccome non vediamo valori oltre i limiti di controllo sino al 241-esimo campione , allora l'errore segnalato sarà dovuto ad un lieve scostamento della media dei residui.

- Ed il grafico seguente con 453 dati



4.2- CONCLUSIONI:

In conclusione, possiamo dire che nell'ambito del controllo statistico di processo, le carte di controllo sono un potente strumento. La carta aggiustata per la regressione grazie allo studio della relazione tra due o più variabili consente di restringere lo studio al controllo dei residui del modello di stima usato. Ovviamente come per ogni problema di regressione, il risultato ottenuto dipende dalla qualità del modello scelto che al suo posto è fortemente condizionato da uno buona stima dei parametri.

Noi possiamo stimare bene i parametri, cioè avvicinarci molto ai valori veri che non conosciamo solo se riusciamo ad avere una numerosità campionaria considerevole. La sovrastima o sottostima del valore medio delle variabili esplicative risultano in una sovra o sottostima del parametro β'_0.

Comunque anche se riuscissimo a stimare bene i parametri, rimarrebbe sempre il fatto che un risultato migliore sarebbe ottenuto con i valori veri.

La stima della variabilità sigma non è stata trattata in questo articolo in cui abbiamo tenute per buone le affermazioni, però gioca un ruolo molto importante nella qualità dei risultati.

APPENDICE

Funzioni:

- Per disegnare il primo grafico:

```
 >esempio<-function (dati,L,mu,sigma,tit)
{
LCL<-mu-(L*sigma)
LC<-mu
UCL<-mu +(L*sigma)
cat("LCL = ", round(LCL,digits=4), "\n")
cat("LC = ", round(LC,digits=4), "\n")
cat("UCL = ", round(UCL,digits=4), "\n")
plot(dati,ylim=c(LC-5,LC+5),main=tit,type="l")
points(dati,pch=20)
abline(a=LCL,b=0,col="red")
abline(a=LC,b=0,col="blue")
abline(a=UCL,b=0,col="red")
}
```

- Funzione per simulazioni sulla numerosità campionaria:

```
>Stima<-function (n,N)
{
 beta_0<-
matrix(c(rep(0,N)),nrow=N,ncol=1,byrow=T
RUE)
 beta_1<-
matrix(c(rep(0,N)),nrow=N,ncol=1,byrow=T
RUE)
 for(i in 1:N){
 fit<-0
 X<-rnorm(n,0,1)
 E<-rnorm(n,0,1)
 Y<-20+2.5*X+E
 fit<-lm(Y~X)
 beta_0[i]<-fit$coef[1]
 beta_1[i]<-fit$coef[2]
    }
beta_0_hat<-mean(beta_0)
beta_1_hat<-mean(beta_1)

cat('Beta0_hat:',beta_0_hat,'  ')
cat('Varianza:',var(beta_0[,1]),'\n')
cat('Beta1_hat:',beta_1_hat,'  ')
cat('Varianza:',var(beta_1[,1]),'\n')
}
```

Disegno CUSUM –calcolo RL –calcolo prob. Fuori controllo:

```
>CUSUMREG<-
function(dati,k,sigma,mu0,h,tit)
{
s<-cusumu(dati,k,sigma,mu0)
t<-cusuml(dati,k,sigma,mu0)
LCL<--h*sigma
UCL<-h*sigma
matplot(cbind(s,t),type="l",ylim=c(min(c
(min(t),LCL)),max(c(max(s),UCL))),lty=1:
2,main=tit)
points(s)
points(t)
abline(a=LCL,b=0,col="blue")
abline(a=UCL,b=0,col="blue")
abline(a=0,b=0,col="blue")
w<-cbind(s,t)

fuori<-
length(which(abs(w)>h))/length(dati)
rl<-min(which(abs(w)>h))

if (rl>length(dati)){

rl<-rl-length(dati)

}
cat('Prob fuori controllo:',fuori,'\n')
cat('ARL del processo:',rl,'\n')
```

```
}
```

CUSUM superiore:

```
>Cusumu<-function(dati,k,sigma,mu0)
{
K<-k*sigma
n<-length(dati)
s<-c(rep(0,n))
for (t in 1:length(dati))
{
if (t==1)
{
s[1]<-max(0,(dati[1]-mu0-K))
}
else
{
s[t]<-max(0,(s[t-1]+dati[t]-mu0-K))
}
}
t<-t+1
s
}
```

CUSUM inferiore :

```
>cusuml<-function(dati,k,sigma,mu0)
    {
K<-k*sigma
n<-length(dati)
```

```
t<-c(rep(0,n))
for (i in 1:length(dati))
{
if (i==1)
{
t[1]<-min(0,(dati[1]-mu0+K))
}
else
{
t[i]<-min(0,(t[i-1]+dati[i]-mu0+K))
}
}
i<-i+1
t
}
```

Analisi esplorativa delle assunzioni:

> summary(Y)
>boxplot(Y)
> qqnorm(Y)
>plot(X,Y)
> lm(Y~X)
> e<-Y-(20.085+1.14*X+E)
> plot(e)
> qqnorm(e)
> CUSUMREG(e,0.5,3,0,4.77,"Carta Cusum sui residui")

> mean(e[201:225])

> mean(e[475:500])

```
> e1<-e[1:202]
> e2<-e[225:475]

> e3<-c(e1,e2)
> CUSUMREG(e3,0.5,3,0,4.77,"Carta Cusum sui residui con n=453")
```

BIBLIOGRAFIA

Douglas C. Montgomery, *controllo statistico della qualità*, terza edizione, McGraw-Hill.

Lianjie Shu, Fugee Tsung, Kwok-Leung Tsui, "Run-Length performance of regression control charts with estimated parameters" *Journal of Quality technology*, vol. 36, No. 3, July 2004 Pg 280-292.

Willis A.Jensen, L.Allison Jones-Farmer, Charles W.champ and William H.Woodall, "Effects of parameter estimation on control chart properties: A literature review", *Journal of Quality technology*, vol. 38, No.4, October 2006

Alberto Lacobini , *Il controllo statistico della qualità, teoria e metodi*,nuova edizione, EUROMA, Editrice universitaria di Roma-LA GOLIARDICA

G. Barrie Wetherill and Don W.Brown, *Statistical process control*, teory and practice, 1991, CHAPMAN and HALL.