Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Analisi dei residui di un fit mediante metodi basati su random walk

Relatore:
Prof. Daniel Remondini

Correlatore:
Dott.ssa Alessandra Merlotti

Presentata da:
Luca Sammarini

Anno Accademico 2019/2020

Abstract

Lo scopo del lavoro svolto è studiare la bontà di alcuni fit eseguiti su distribuzioni di punti generati tramite MATLAB. Si vuole infatti capire se è possibile, tramite metodi basati su random walk applicati ai residui dei fit, identificare deviazioni sistematiche rispetto alle distribuzioni sperimentali, le quali non sono misurabili attraverso i classici indicatori di bontà quali r^2 o *root mean square error*. Si sono dunque svolti dei test su dati simulati e si è studiato come variano i parametri associati al random walk se vengono cambiati il numero di dati generati, il rumore applicato ai dati e l'intervallo di generazione degli stessi. Si è compreso che la variazione dei parametri associati al random walk non è significativa in nessuno dei tre casi se i dati sono fittati con la funzione corretta. Si è poi studiato come si trasforma il random walk per un fit eseguito con una funzione errata. Si è osservato che per dati con rumore piccolo lo studio dei residui si dimostra un indicatore accurato della deviazione del fit, mostrando la presenza di grandi sequenze di punti discostati dal fit, mentre perde di efficacia per valori di rumore molto grandi, dove lo scostamento dal fit risulta ridotto dal rumore e dunque i risultati sono equivalenti a quelli di un fit corretto.

Indice

Capitolo 1

Introduzione

Lo scopo del lavoro è comprendere se alcuni parametri associati a un random walk possono essere buoni indicatori della bontà di un fit, mostrandone la deviazione sistematica.

In una distribuzione i punti sono soggetti a un rumore gaussiano, dunque aleatorio e simmetrico. Perciò la probabilità che il punto si trovi sopra il fit, generando un residuo positivo, è uguale alla probabilità che il punto si trovi sotto il fit, generando un residuo negativo. Dunque il segno di ogni residuo può essere associato a una prova di Bernoulli, e da tali prove può essere costruito un random walk.

Lo studio dell'andamento del random walk può essere utile per lo studio dei residui di un fit. In particolare, essendo un indicatore del segno dei residui, può essere utilizzato per conoscere la posizione relativa del fit rispetto alla distribuzione e dunque può essere usato per studiare deviazioni sistematiche del fit rispetto alla distribuzione, deviazioni che non sarebbero misurabili con altri fattori.

Per testare l'utilità di questi indicatori eseguiremo degli studi su dati generati in modo aleatorio tramite MATLAB.

Inizialmente effettueremo delle generazioni seguendo un modello lineare, a cui sarà aggiunto un rumore gaussiano, e le fitteremo con una funzione rettilinea. Dopo aver ricavato il segno dei residui studieremo le sequenze di residui con segno concorde, dimostrando che la variazione del numero di punti generati, della quantità di rumore gaussiano e dell'intervallo dei punti non hanno effetto sulle sequenze.

Si eseguiranno successivamente delle generazioni secondo un modello parabolico e si fitteranno anch'esse con una funzione rettilinea, dunque errata. Anche in questo caso si ricaveranno i segni dei residui per studiare le sequenze di residui con segno concorde e osservare che all'aumentare del rumore gaussiano lo scostamento del fit risulta meno evidente e dunque la curva generata diventa sempre più simile a un random walk standard. Si osserverà quindi che per distribuzioni con rumori grandi il random walk non risulta più essere un parametro indicativo.

Capitolo 2

Random walk e distribuzione geometrica

2.1 Random walk

Un random walk può essere descritto come un percorso creato da un processo stocastico e formalmente può essere definito nel seguente modo.

Si consideri il reticolo d-dimensionale $\mathbb{Z}^d$. Sia e_i il vettore base standard con 1 alla i-esima coordinata e 0 altrove. Definiamo X_j come un vettore casuale con immagine $\pm e_i$ per un qualche $i \in 1, ..., d$. Si assumano $X_1, X_2, X_3, ...$ indipendenti e non identicamente distribuiti. Il random walk semplice a n passi, indicato con S_n, è definito come

$$S_n = x + \sum_{i=1}^{n} X_i.$$

Nella formula, x denota la posizione sul reticolo al tempo $n = 0$ e X_j rappresenta il movimento dal tempo j al tempo $j + 1$.

Se vale

$$\Pr(X_i = e_i) = \Pr(X_i = -e_i) = \frac{1}{2d}, \quad i = 1, 2, ..., d,$$

allora il random walk è definito simmetrico [1].

In Figura 2.1 sono riportati alcuni random walk a titolo esemplificativo.

2.2 Distribuzione geometrica

Nel corso dello studio, al fine di valutare le deviazioni sistematiche della curva stimata tramite il metodo di fit dalla distribuzione sperimentale, verrà ricavata la distribuzione delle lunghezze delle sequenze di passi nello stesso verso, le quali in un random walk simmetrico seguono l'andamento della distribuzione geometrica, descritta di seguito.

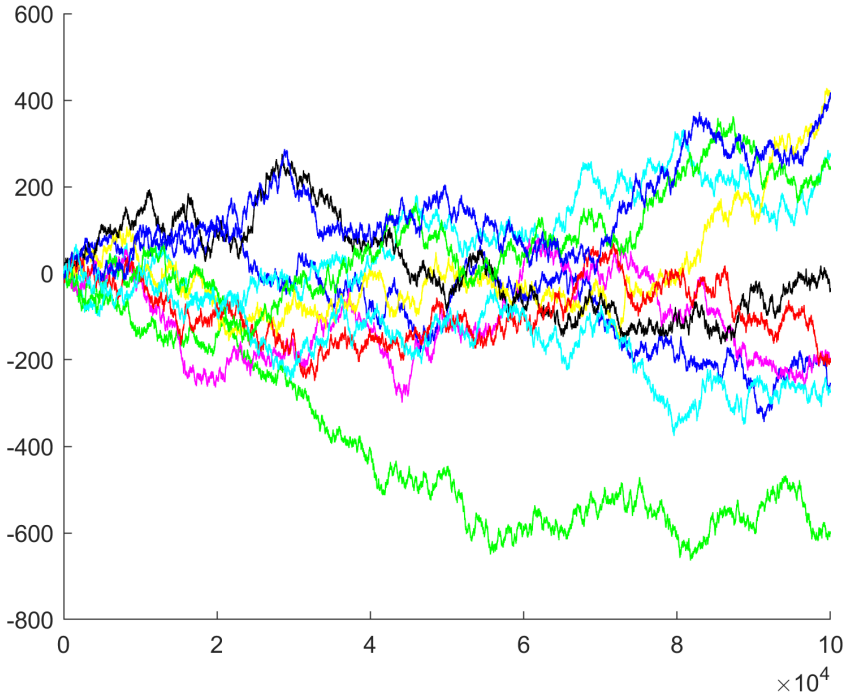


Figura 2.1: Random walk unidimensionali simmetrici.

Per trattare la distribuzione geometrica è prima necessario introdurre la *Prova di Bernoulli*.

Prove ripetute e indipendenti di un esperimento con esattamente due esiti sono definite prove di Bernoulli. Si chiami uno dei due esiti *successo* e l'altro *fallimento*. Sia p la probabilità di successo in una prova di Bernoulli, e q la probabilità di fallimento. Allora le due probabilità hanno somma 1, poiché sono eventi complementari: *successo* e *fallimento* sono mutuamente escludentesi ed esaustivi. Si hanno perciò le seguenti relazioni: $p = 1 - q$, $q = 1 - p$, $p + q = 1$ [2].

È ora possibile introdurre la distribuzione geometrica. La distribuzione geometrica può essere descritta con le seguenti due distribuzioni:

- la distribuzione di probabilità del numero X di prove di Bernoulli necessarie per ottenere un successo, ha come supporto l'insieme $\{1, 2, 3, \ldots\}$;
- la distribuzione di probabilità del numero $Y = X - 1$ di fallimenti antecedenti il primo successo, ha come supporto l'insieme $\{0, 1, 2, \ldots\}$.

La prima forma della distribuzione geometrica indica la probabilità che il primo successo richieda k prove indipendenti, ognuna con una probabilità di successo p. Sia dunque p la probabilità di successo di ogni prova, allora la probabilità che il k-esimo

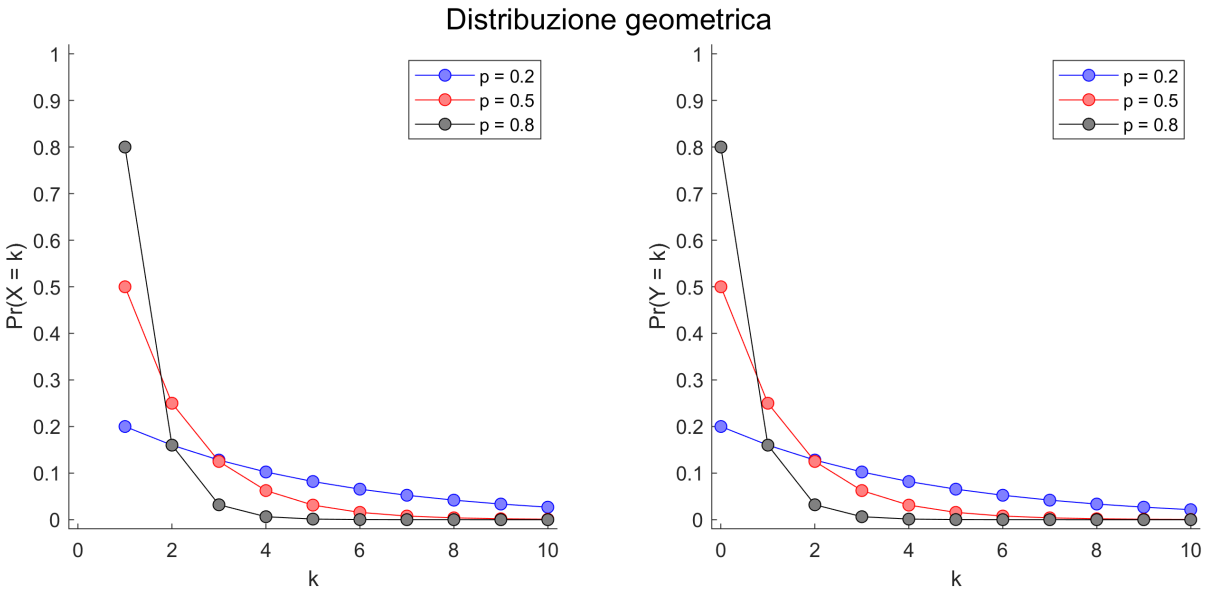


Figura 2.2: Distribuzione geometrica.

tentativo (su k tentativi) sia il primo successo è

$$\Pr(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \ldots.$$

La forma precedente della distribuzione geometrica viene usata per modellare il numero di tentativi fino al primo successo compreso. Invece, la forma seguente della distribuzione geometrica è usata per modellare il numero di fallimenti precedenti il primo successo:

$$\Pr(Y = k) = \Pr(X = k + 1) = (1-p)^{k}p, \quad k = 0, 1, 2, \ldots.$$

In entrambi i casi la successione delle probabilità è una successione geometrica.

In Figura 2.2 sono riportati i grafici di entrambe le forme della distribuzione.

2.2.1 Proprietà

Il valore atteso del numero di prove indipendenti per ottenere il primo successo e la varianza di una variabile aleatoria geometricamente distribuita X sono

$$\mathrm{E}(X) = \frac{1}{p}, \qquad \mathrm{var}(X) = \frac{1-p}{p^2}.$$

Allo stesso modo, il valore atteso e la varianza della variabile aleatoria geometricamente distribuita $Y = X - 1$ sono

$$\mathrm{E}(Y) = \frac{1-p}{p}, \qquad \mathrm{var}(X) = \frac{1-p}{p^2}.$$

Le mediane nei due casi valgono rispettivamente

$$\text{median}(X) = \left\lceil \frac{-1}{\log_2(1-p)} \right\rceil, \qquad \text{median}(Y) = \left\lceil \frac{-1}{\log_2(1-p)} \right\rceil - 1$$

se $-1/\log_2(1-p)$ non è intero, mentre non sono uniche altrimenti.

La distribuzione geometrica possiede mancanza di memoria. Ciò significa che se si è intenzionati a ripetere un esperimento fino al primo successo, allora, se il primo successo non si è ancora verificato, la distribuzione di probabilità condizionale del numero di prove addizionali non dipende da quanti fallimenti sono stati osservati.

$$\Pr(X > m + n \mid X > n) = \Pr(X > m)$$

La distribuzione geometrica è l'unica distribuzione discreta che gode di questa proprietà [3].

Capitolo 3

Analisi

3.1 Modello lineare

Si sono eseguiti degli studi con numeri generati mediante metodi aleatori tramite il software MATLAB.

Inizialmente si sono generati n numeri casuali, distribuiti uniformemente nell'intervallo tra 0 e 100. Ad essi si è aggiunto un rumore gaussiano. Si sono poi fittati i valori ottenuti con una funzione polinomiale $ax+b$. Dopo aver ricavato i residui del fit eseguito, sono state determinate le posizioni relative di ogni punto generato rispetto al fit (sopra o sotto il fit).

Essendo stato aggiunto un rumore gaussiano (dunque aleatorio e simmetrico) la probabilità che il punto si trovi sotto al fit è pari alla probabilità che il punto si trovi sopra lo stesso ($p = 0.5$). Anche in un random walk simmetrico la probabilità di effettuare un passo +1 o -1 è la stessa, perciò il segno di ogni residuo può essere associato ad un passo di un random walk. Estendendo il concetto a tutto il fit, dalla successione dei segni di tutti i residui generati si può costruire un random walk unidimensionale che possiede informazioni analoghe. Si sono allora analizzate alcune caratteristiche di quest'ultimo, allo scopo di ricavare le proprietà della distribuzione dei residui del fit.

Inizialmente si sono ottenute tutte le sequenze di punti consecutivi che si trovano dallo stesso lato del fit, cioè i passi consecutivi del random walk che vengono percorsi nella stessa direzione, in seguito chiamate escursioni[1]. Si è dunque studiata la distribuzione di queste escursioni, dalla quale si è poi trovata l'escursione massima e quella minima, quella media e quella mediana. Poiché la distribuzione geometrica presuppone di stabilire a priori quale dei due esiti sia successo e quale fallimento, si è prima studiata la distribuzione delle escursioni dei residui negativi, ponendo l'evento "residuo positivo"

[1]Con escursioni di un random walk sono solitamente indicate le parti dello stesso comprese tra due ritorni all'origine (come in E. Csáki e Y. Hu [4] e in E. Csáki, P. Erdős e P. Révész [5]). In questo lavoro invece si è inteso il termine escursione come sequenza di passi consecutivi eseguiti nella stessa direzione.

n	a	b	r^2	rmse	max neg. exc.	min neg. exc.	mean neg. exc.	median neg. exc.
1×10^2	9.998 ± 0.006	5.0 ± 0.3	1.0000	0.86	6	0	0.92	0
1×10^3	9.998 ± 0.002	5.1 ± 0.1	1.0000	0.90	12	0	1.00	0
1×10^4	9.9999 ± 0.0006	5.02 ± 0.03	1.0000	0.89	11	0	1.00	1
1×10^5	10.0002 ± 0.0002	4.99 ± 0.01	1.0000	0.89	15	0	1.00	0
1×10^6	$9.99998 \pm 6 \times 10^{-5}$	5.000 ± 0.003	1.0000	0.89	20	0	1.00	0
1×10^7	$10.00000 \pm 2 \times 10^{-5}$	5.000 ± 0.001	1.0000	0.89	21	0	1.00	1

(a) Analisi dei residui negativi.

n	a	b	r^2	rmse	max pos. exc.	min pos. exc.	mean pos. exc.	median pos. exc.
1×10^2	9.998 ± 0.006	5.0 ± 0.3	1.0000	0.86	7	0	1.06	1
1×10^3	9.998 ± 0.002	5.1 ± 0.1	1.0000	0.90	10	0	0.99	0
1×10^4	9.9999 ± 0.0006	5.02 ± 0.03	1.0000	0.89	11	0	1.00	1
1×10^5	10.0002 ± 0.0002	4.99 ± 0.01	1.0000	0.89	18	0	1.00	1
1×10^6	$9.99998 \pm 6 \times 10^{-5}$	5.000 ± 0.003	1.0000	0.89	17	0	1.00	0
1×10^7	$10.00000 \pm 2 \times 10^{-5}$	5.000 ± 0.001	1.0000	0.89	22	0	1.00	0

(b) Analisi dei residui positivi.

Tabella 3.1: Analisi residui negativi e positivi del fit variando n, mantenendo un rapporto tra segnale e rumore costante a 1. I punti sono generati in un intervallo tra 0 e 100.

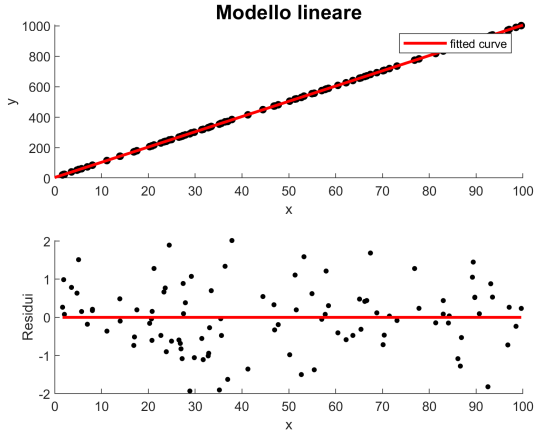
come successo e studiando quindi la distribuzione degli insuccessi. Si è poi studiata la distribuzione delle escursioni dei residui positivi ponendo come successo il "residuo negativo" e dunque studiando la distribuzione degli insuccessi.

Le prime generazioni sono state eseguite variando il numero di valori generati n e tenendo fisso l'intervallo di distribuzione dei valori tra 0 e 100 e il rapporto tra segnale e rumore $snr = 1$ (Figura 3.1).

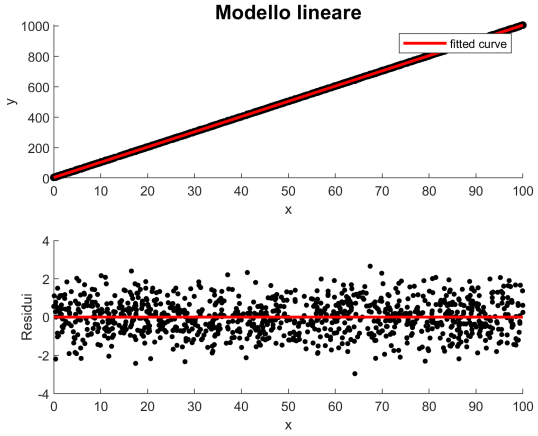
Per ogni generazione si sono prodotti due grafici rappresentanti rispettivamente l'istogramma, normalizzato a 1, della lunghezza delle escursioni dei residui positivi e dei residui negativi, per studiare la distribuzione delle stesse (Figura 3.2). È possibile osservare che la distribuzione dei punti segue una distribuzione geometrica con $p = 0.5$, come descritto nel Capitolo 2, con una precisione crescente all'aumentare del numero di punti generati.

Successivamente si sono ricavati i parametri sulle escursioni citati in precedenza, riportati in Tabella 3.1.

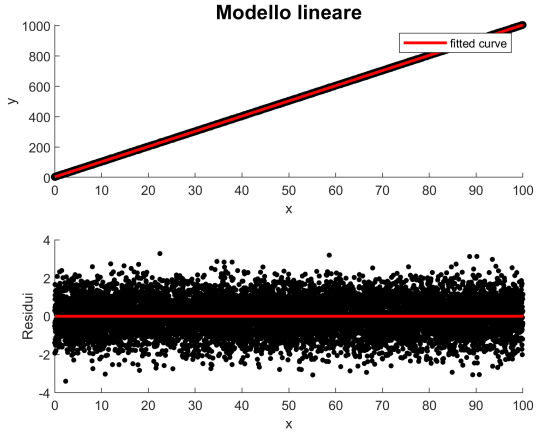
È possibile osservare che l'escursione minima non dipende dal numero di punti, attestandosi sempre a un valore di 0. Questo è infatti il valore più probabile per la distribuzione e si verifica sempre. Il valore massimo cresce invece con il crescere del numero di punti, abbiamo infatti che aumentando i tentativi vengono prodotti risultati sempre meno probabili. La distribuzione media si attesta sempre attorno ad un valore di 1.00 punti, con una precisione crescente all'aumentare dei punti generati. Il valore è in li-



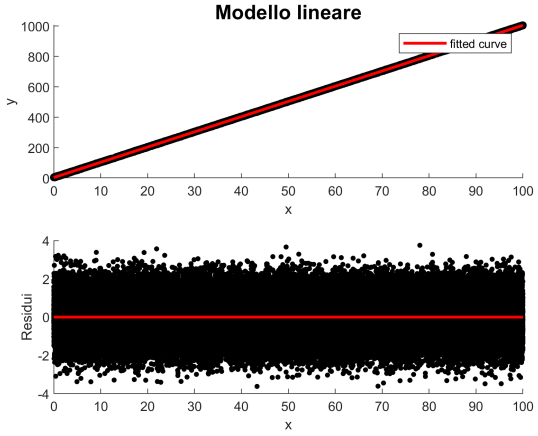
(a) $n = 1 \times 10^2$



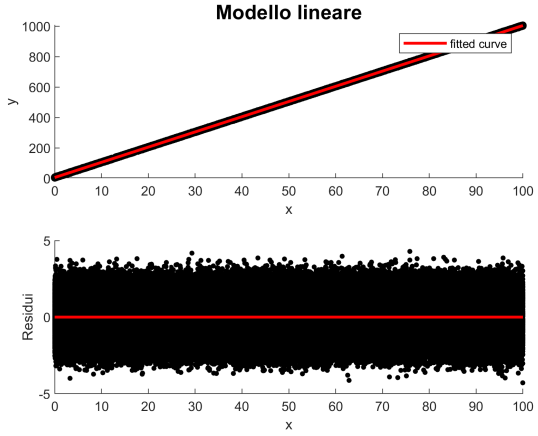
(b) $n = 1 \times 10^3$



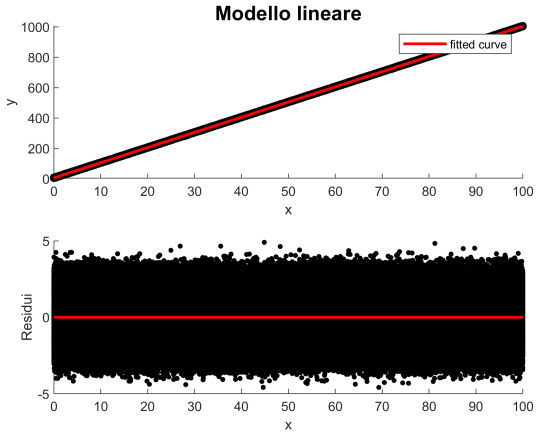
(c) $n = 1 \times 10^4$



(d) $n = 1 \times 10^5$

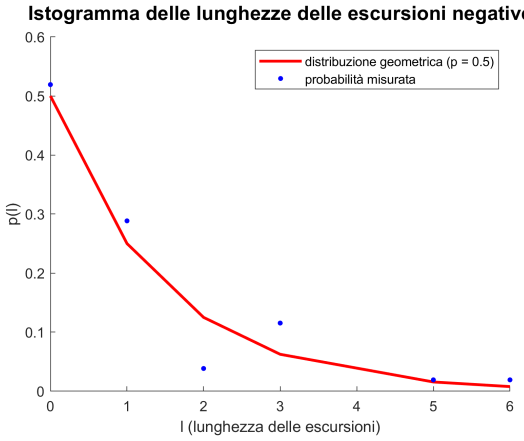


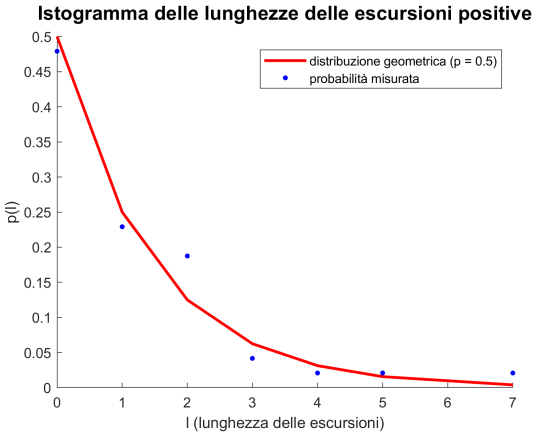
(e) $n = 1 \times 10^6$



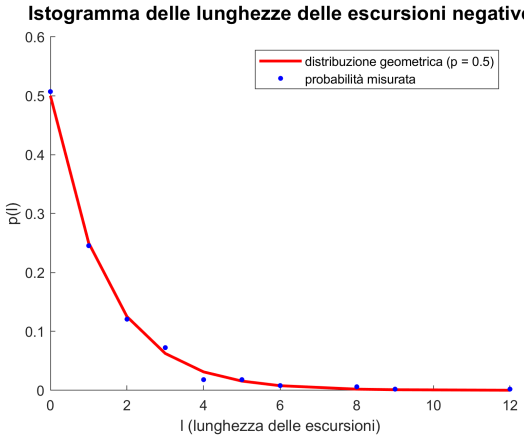
(f) $n = 1 \times 10^7$

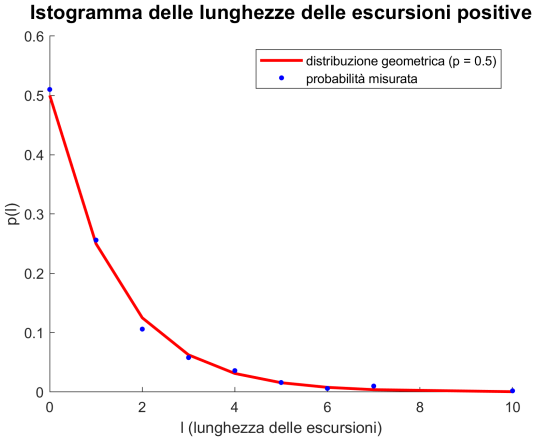
Figura 3.1: Generazioni del modello lineare variando il parametro n.



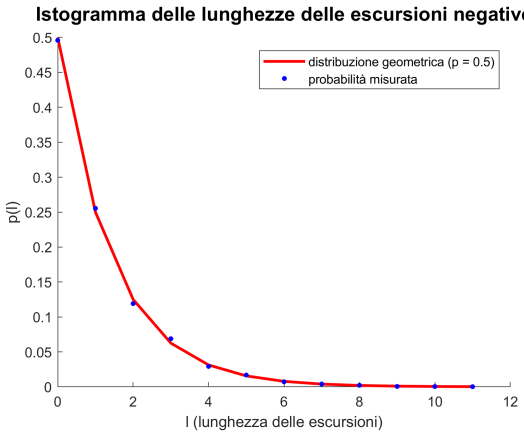


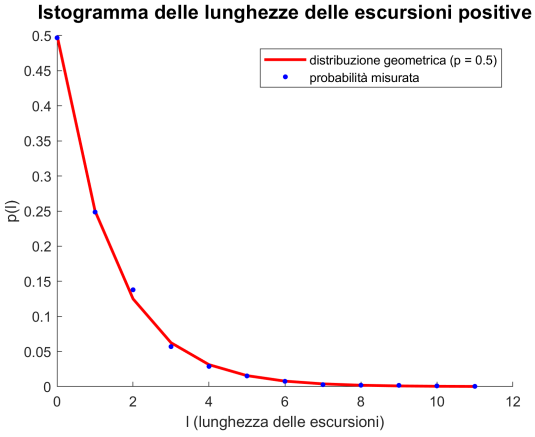
(a) $n = 1 \times 10^2$





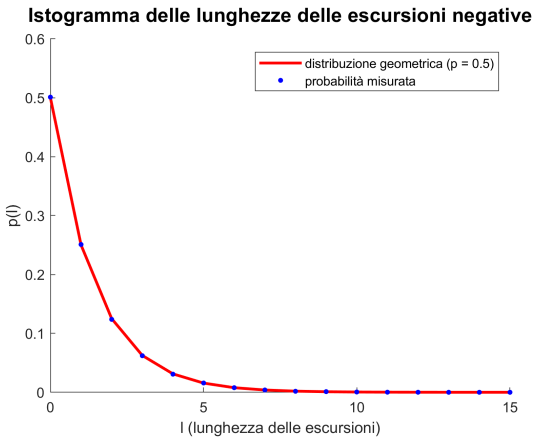
(b) $n = 1 \times 10^3$

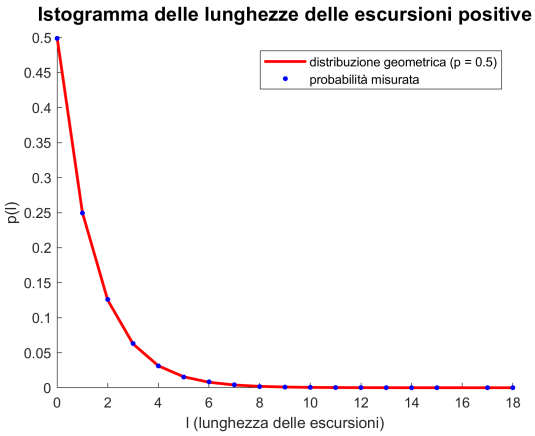




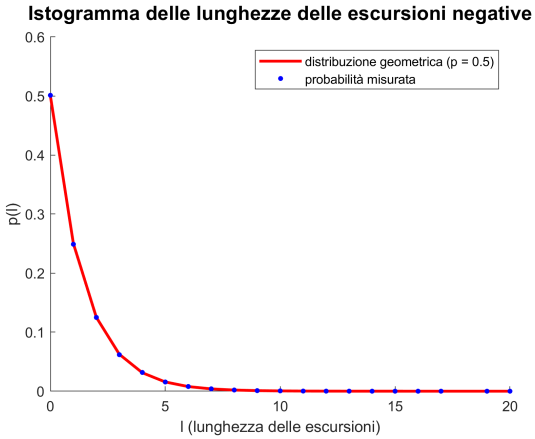
(c) $n = 1 \times 10^4$

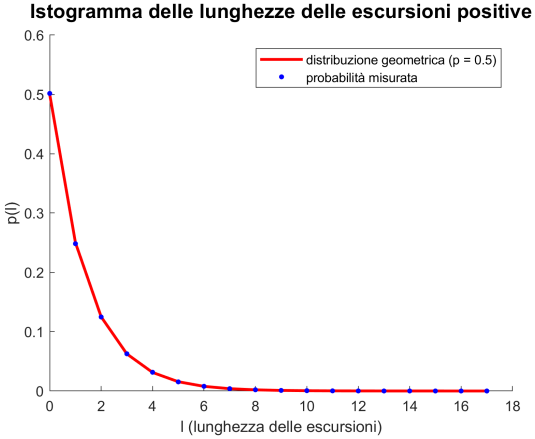
Figura 3.2: Istogrammi delle lunghezze delle escursioni con diversi n.



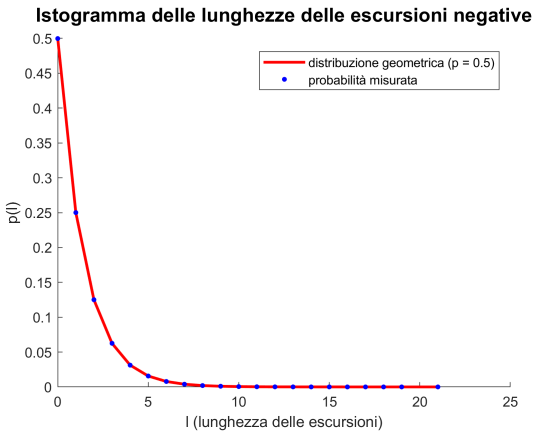


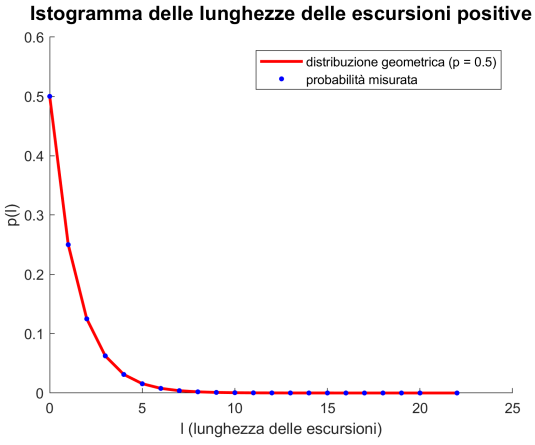
(d) $n = 1 \times 10^5$





(e) $n = 1 \times 10^6$





(f) $n = 1 \times 10^7$

Figura 3.2: Istogrammi delle lunghezze delle escursioni con diversi n.

snr	a	b	r^2	rmse	max neg. exc.	min neg. exc.	mean neg. exc.	median neg. exc.
100	$10.0000000001 \pm 2 \times 10^{-10}$	$4.99999999 \pm 1 \times 10^{-8}$	1.0000	1.00×10^{-5}	25	0	1.00	0
50	$9.99999997 \pm 7 \times 10^{-8}$	$5.000002 \pm 4 \times 10^{-6}$	1.0000	0.0032	23	0	1.00	0
10	$10.000002 \pm 7 \times 10^{-6}$	4.9997 ± 0.0004	1.0000	0.32	20	0	1.00	0
1	$10.00000 \pm 2 \times 10^{-5}$	5.000 ± 0.001	1.0000	0.89	23	0	1.00	1
0.1	$9.99999 \pm 2 \times 10^{-5}$	5.000 ± 0.001	1.0000	0.99	23	0	1.00	0
0.01	$10.00001 \pm 2 \times 10^{-5}$	5.000 ± 0.001	1.0000	1.00	22	0	1.00	1

(a) Analisi dei residui negativi.

snr	a	b	r^2	rmse	max pos. exc.	min pos. exc.	mean pos. exc.	median pos. exc.
100	$10.0000000001 \pm 2 \times 10^{-10}$	$4.99999999 \pm 1 \times 10^{-8}$	1.0000	1.00×10^{-5}	25	0	1.00	0
50	$9.99999997 \pm 7 \times 10^{-8}$	$5.000002 \pm 4 \times 10^{-6}$	1.0000	0.0032	23	0	1.00	1
10	$10.000002 \pm 7 \times 10^{-6}$	4.9997 ± 0.0004	1.0000	0.32	21	0	1.00	0
1	$10.00000 \pm 2 \times 10^{-5}$	5.000 ± 0.001	1.0000	0.89	21	0	1.00	0
0.1	$9.99999 \pm 2 \times 10^{-5}$	5.000 ± 0.001	1.0000	0.99	22	0	1.00	0
0.01	$10.00001 \pm 2 \times 10^{-5}$	5.000 ± 0.001	1.0000	1.00	22	0	1.00	0

(b) Analisi dei residui positivi.

Tabella 3.2: Analisi dei residui negativi e positivi del fit variando il rapporto tra segnale e rumore. Ogni generazione è composta da 1×10^7 valori compresi nell'intervallo tra 0 e 100.

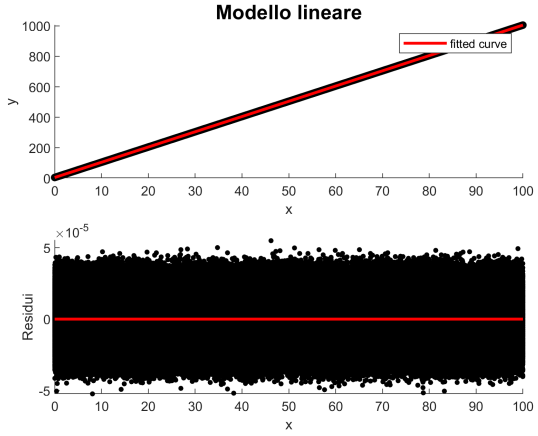
nea con il quello atteso per una distribuzione geometrica: vale infatti $\mathrm{E}(Y) = \frac{1-p}{p} = 1$. L'escursione mediana assume due valori, 0 e 1, senza alcuna legge evidente.

Si è poi proceduto ad eseguire delle generazioni variando il rapporto tra segnale e rumore, mantenendo fisso $n = 1 \times 10^7$. Le generazioni sono riportate in Figura 3.3, mentre le distribuzioni delle escursioni in Figura 3.4. Anche in questo caso gli istogrammi rispecchiano la distribuzione geometrica attesa. I parametri ottenuti dallo studio delle distribuzioni sono in Tabella 3.2.

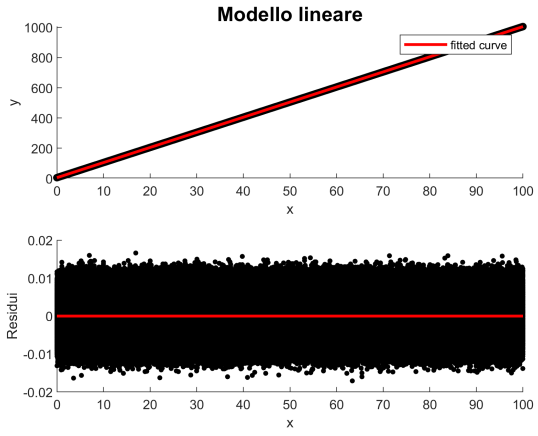
È possibile osservare che la variazione del rapporto tra segnale e rumore non influenza il valore di nessuno dei parametri relativi alle escursioni, mentre è influenzato il *root mean square error*. L'escursione minima e media rimangono infatti equivalenti a quelle della generazione precedente. Anche l'escursione massima rimane attorno ai valori 21 e 22 osservati nell'ultima generazione rispettivamente delle Tabelle 3.1a e 3.1b, corrispondenti a un valore di $n = 1 \times 10^7$ come in questo caso. La mediana assume i medesimi valori osservati in precedenza, 0 e 1, con una prevalenza del valore 0.

Infine si sono eseguite ulteriori generazioni fissando i precedenti parametri, $n = 1 \times 10^7$ e $snr = 1$, e variando l'intervallo di estensione della retta (Figura 3.5). Analogamente ai casi precedenti si mostrano in Figura 3.6 le distribuzioni delle escursioni. Anche in questi grafici si vede rispettata la distribuzione attesa. I dati ottenuti dall'analisi delle distribuzioni sono in Tabella 3.3.

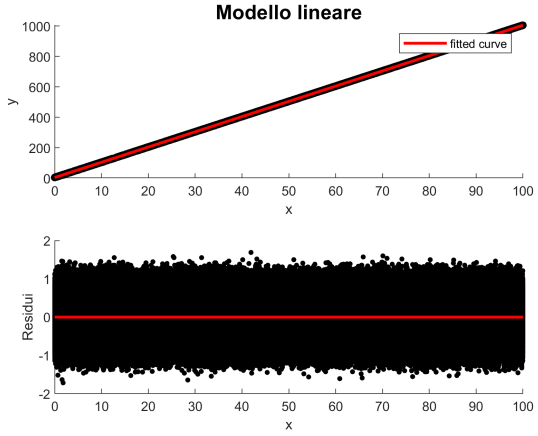
Analogamente alle generazioni precedenti i parametri rimangono pressoché costanti in tutte le generazioni.



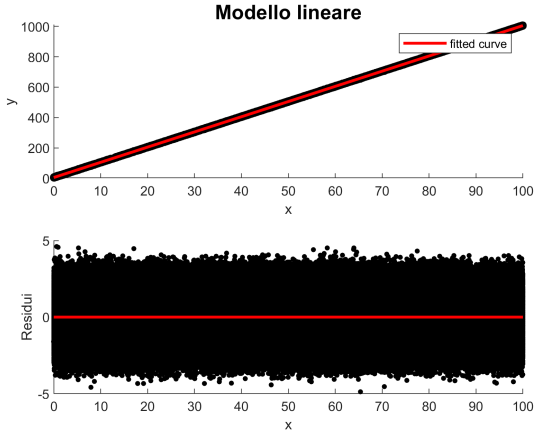
(a) $snr = 100$



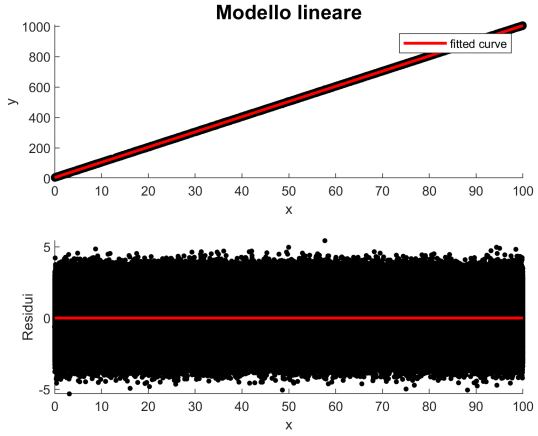
(b) $snr = 50$



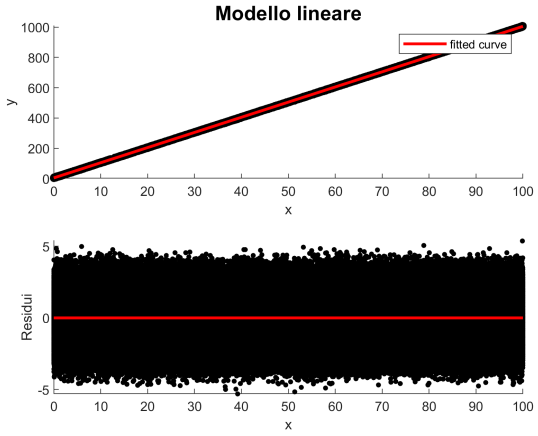
(c) $snr = 10$



(d) $snr = 1$

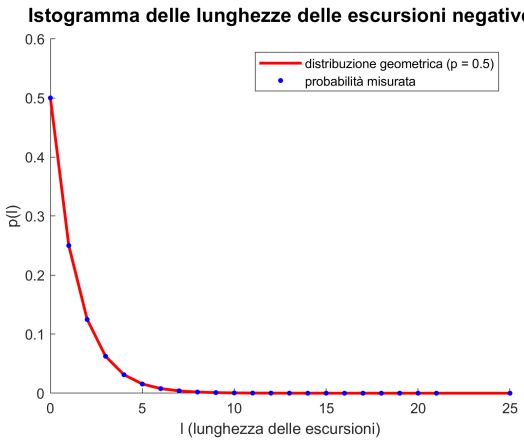


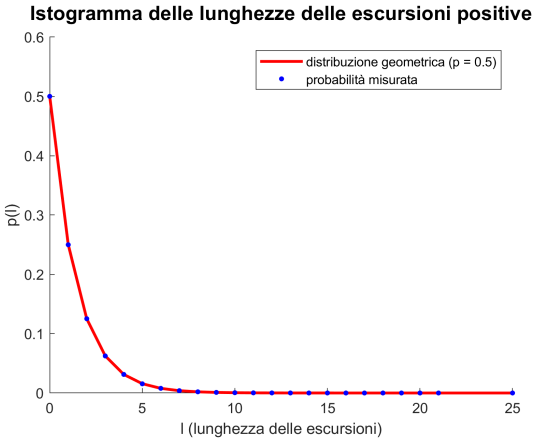
(e) $snr = 0.1$



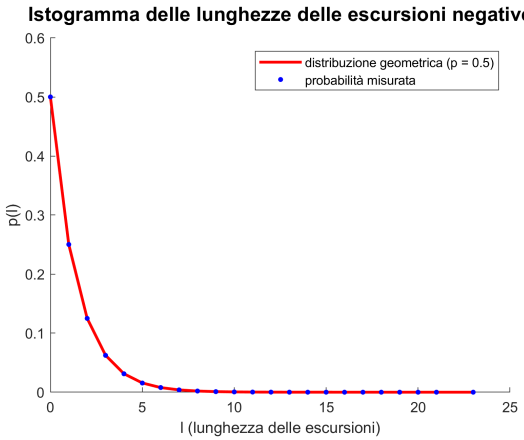
(f) $snr = 0.01$

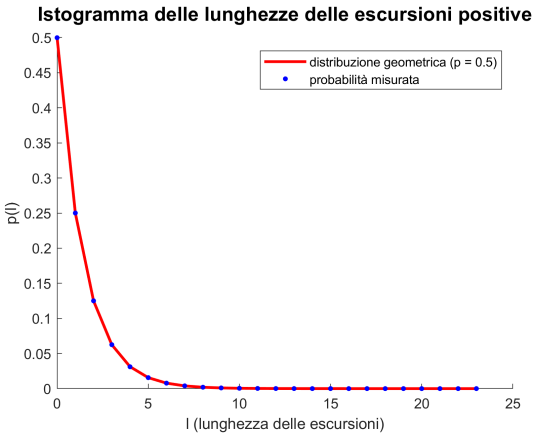
Figura 3.3: Generazioni del modello lineare con diversi rapporti tra segnale e rumore.



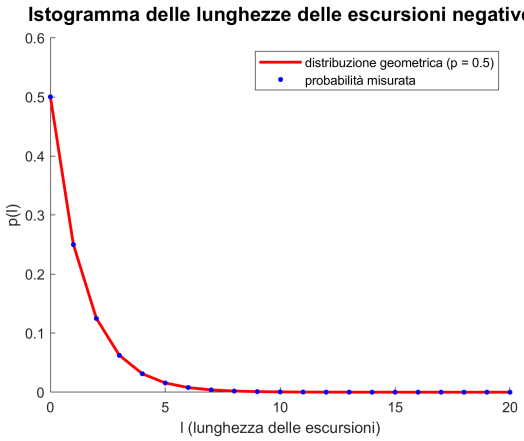


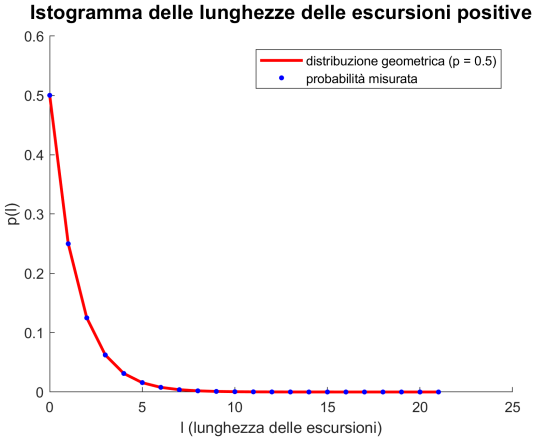
(a) $snr = 100$





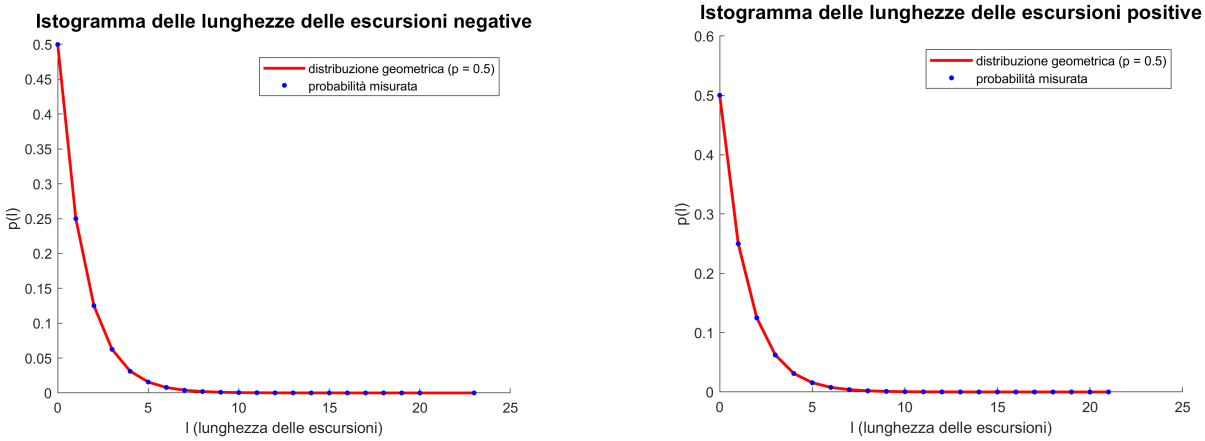
(b) $snr = 50$



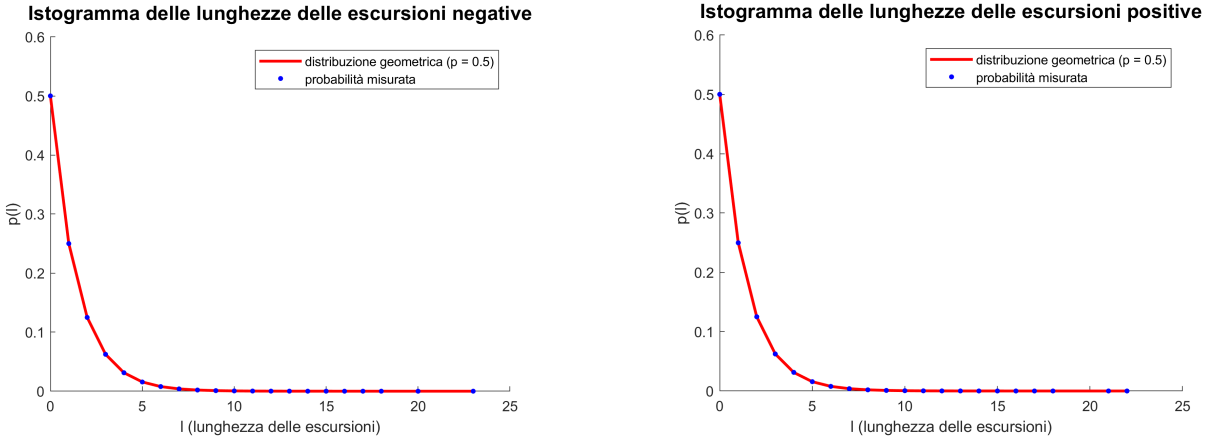


(c) $snr = 10$

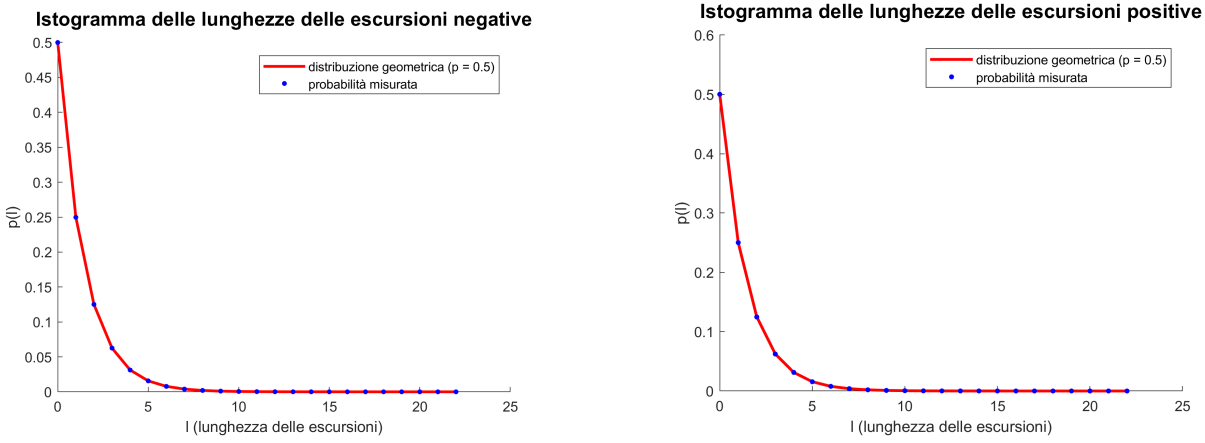
Figura 3.4: Istogrammi delle lunghezze delle escursioni con diversi rapporti tra segnale e rumore.



(d) $snr = 1$

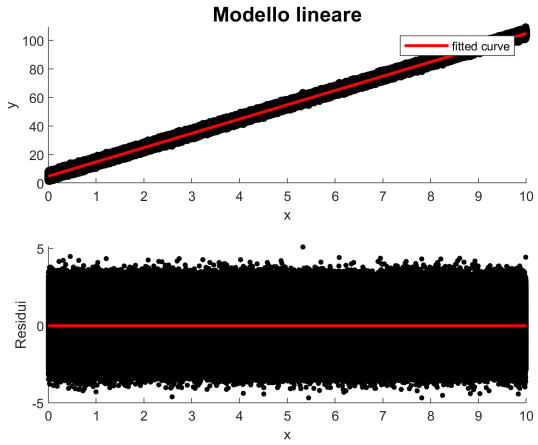


(e) $snr = 0.1$

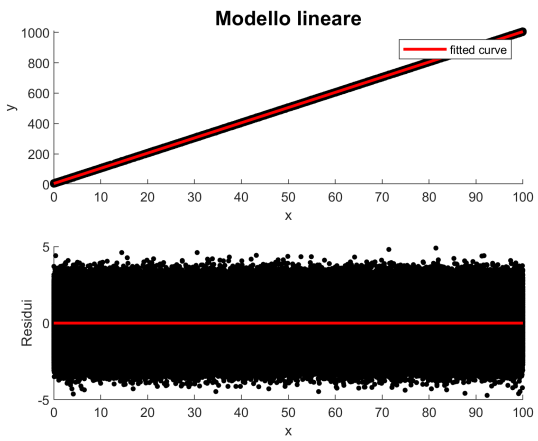


(f) $snr = 0.01$

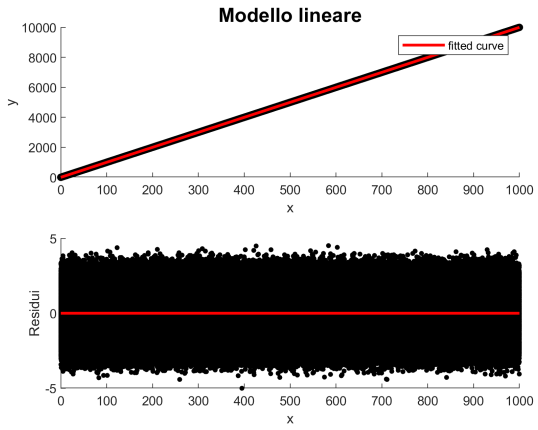
Figura 3.4: Istogrammi delle lunghezze delle escursioni con diversi rapporti tra segnale e rumore.



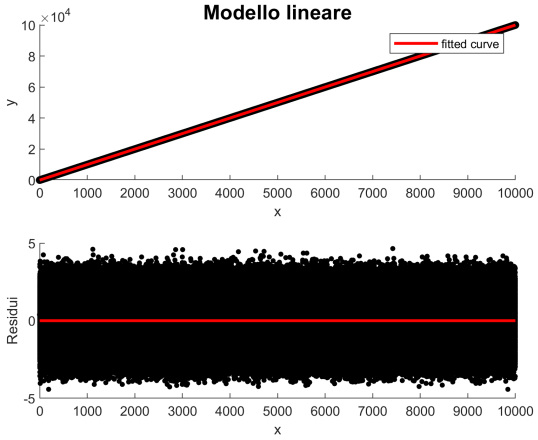
(a) Intervallo tra 0 e 10



(b) Intervallo tra 0 e 50

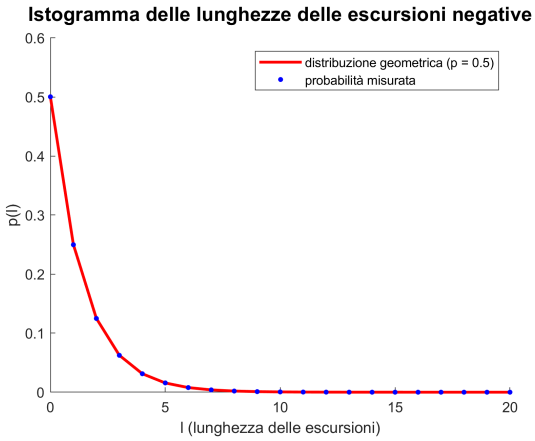


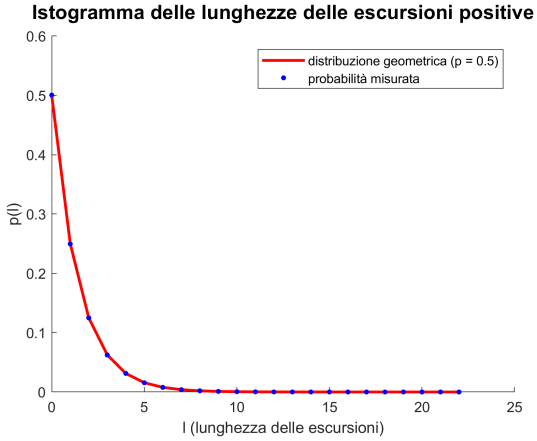
(c) Intervallo tra 0 e 100



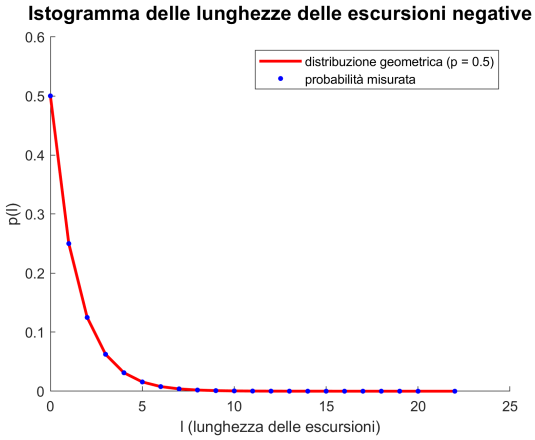
(d) Intervallo tra 0 e 1000

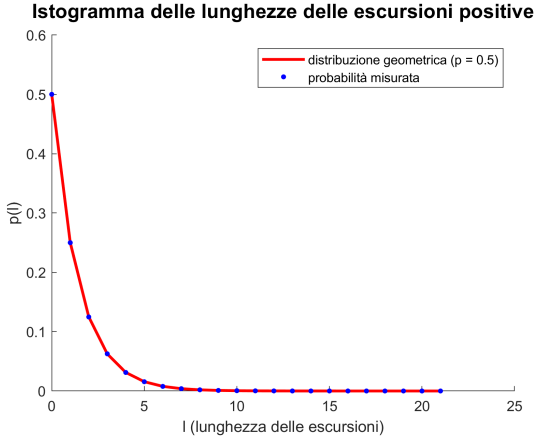
Figura 3.5: Generazioni del modello lineare eseguite su diversi intervalli.



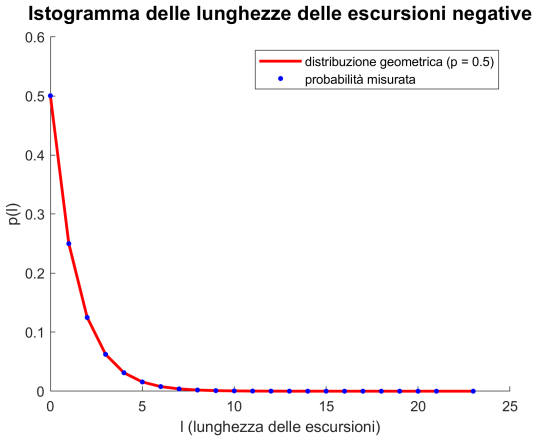


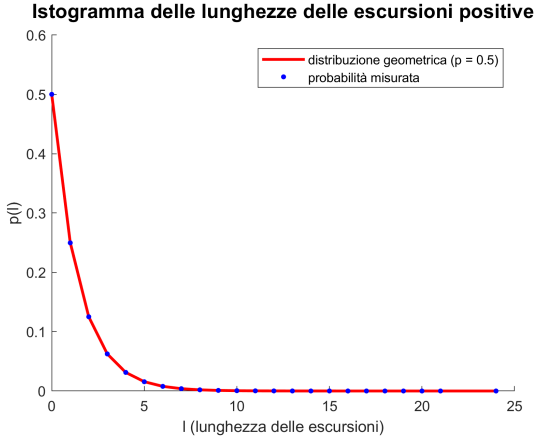
(a) Intervallo tra 0 e 10





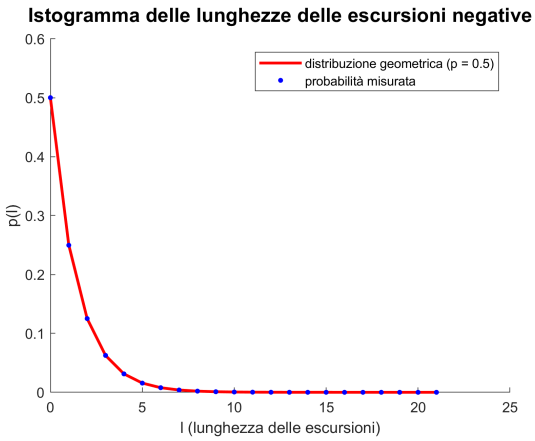
(b) Intervallo tra 0 e 50

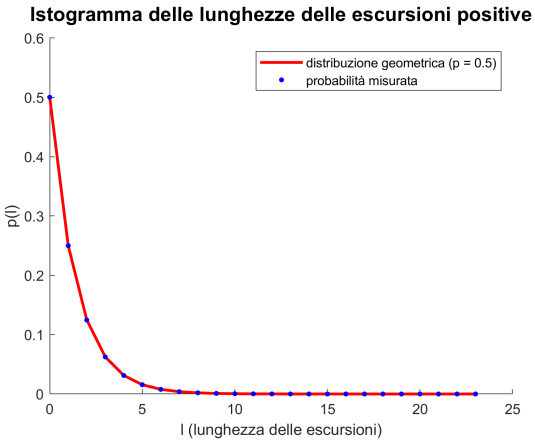




(c) Intervallo tra 0 e 100

Figura 3.6: Istogrammi delle lunghezze delle escursioni di generazioni eseguite su diversi intervalli.





(d) Intervallo tra 0 e 1000

Figura 3.6: Istogrammi delle lunghezze delle escursioni di generazioni eseguite su diversi intervalli.

intervallo	a	b	r^2	rmse	max neg. exc.	min neg. exc.	mean neg. exc.	median neg. exc.
0-10	10.00000 ± 0.0002	5.000 ± 0.001	0.9990	0.89	20	0	1.00	0
0-100	$9.99999 \pm 1.9126e-05$	5.000 ± 0.001	1.0000	0.89	22	0	1.00	0
0-1000	$9.999999 \pm 2 \times 10^{-6}$	5.000 ± 0.001	1.0000	0.89	23	0	1.00	0
0-10000	$9.9999999 \pm 2 \times 10^{-7}$	5.001 ± 0.001	1.0000	0.89	21	0	1.00	0

(a) Analisi residui negativi.

intervallo	a	b	r^2	rmse	max pos. exc.	min pos. exc.	mean pos. exc.	median pos. exc.
0-10	10.00000 ± 0.0002	5.000 ± 0.001	0.9990	0.89	22	0	1.00	0
0-100	$9.99999 \pm 1.9126e-05$	5.000 ± 0.001	1.0000	0.89	21	0	1.00	0
0-1000	$9.999999 \pm 2 \times 10^{-6}$	5.000 ± 0.001	1.0000	0.89	24	0	1.00	0
0-10000	$9.9999999 \pm 2 \times 10^{-7}$	5.001 ± 0.001	1.0000	0.89	23	0	1.00	0

(b) Analisi residui positivi.

Tabella 3.3: Analisi residui negativi e positivi del fit variando l'intervallo dei dati fittati. Ogni generazione è composta da 1×10^7 valori, con un rapporto tra segnale e rumore costante e pari a 1.

snr	a	b	r^2	rmse	max neg. exc.	min neg. exc.	mean neg. exc.	median neg. exc.
100	$-9.9990001 \pm 2 \times 10^{-7}$	$9.9983340 \pm 9 \times 10^{-7}$	1.0000	7.5×10^{-4}	5659576	0	2.65	0
50	$-9.9990005 \pm 7 \times 10^{-7}$	$9.998334 \pm 4 \times 10^{-6}$	1.0000	0.0032	31	0	1.00	0
10	$-9.99898 \pm 7 \times 10^{-5}$	9.9981 ± 0.0004	0.9999	0.32	21	0	1.00	0
1	-9.9990 ± 0.0002	9.998 ± 0.001	0.9990	0.89	23	0	1.00	0
0.1	-9.9989 ± 0.0002	9.998 ± 0.001	0.9988	0.99	22	0	1.00	0
0.01	-9.9989 ± 0.0002	9.998 ± 0.001	0.9988	1.00	20	0	1.00	0

(a) Analisi dei residui negativi.

snr	a	b	r^2	rmse	max pos. exc.	min pos. exc.	mean pos. exc.	median pos. exc.
100	$-9.9990001 \pm 2 \times 10^{-7}$	$9.9983340 \pm 9 \times 10^{-7}$	1.0000	7.5×10^{-4}	2056498	0	0.38	0
50	$-9.9990005 \pm 7 \times 10^{-7}$	$9.998334 \pm 4 \times 10^{-6}$	1.0000	0.0032	37	0	1.00	0
10	$-9.99898 \pm 7 \times 10^{-5}$	9.9981 ± 0.0004	0.9999	0.32	23	0	1.00	0
1	-9.9990 ± 0.0002	9.998 ± 0.001	0.9990	0.89	22	0	1.00	0
0.1	-9.9989 ± 0.0002	9.998 ± 0.001	0.9988	0.99	22	0	1.00	0
0.01	-9.9989 ± 0.0002	9.998 ± 0.001	0.9988	1.00	22	0	1.00	0

(b) Analisi dei residui positivi.

Tabella 3.4: Analisi dei residui negativi e positivi del fit sulla generazione parabolica.

3.2 Modello parabolico

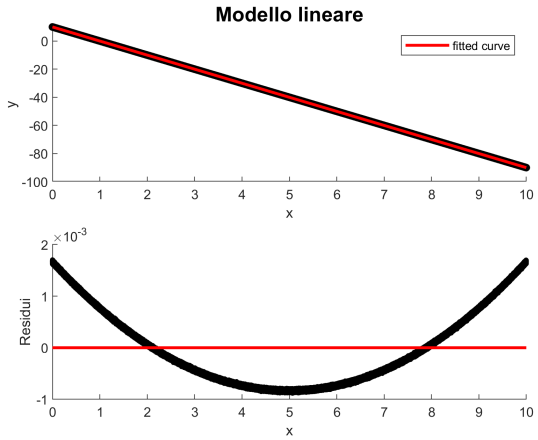
Per studiare i residui per una distribuzione fittata con la funzione sbagliata, si sono eseguiti fit lineari di generazioni di $n = 1 \times 10^7$ elementi generati secondo una funzione parabolica $ax^2 + bx + c$ in un intervallo tra 0 e 10 con un rapporto tra segnale e rumore variabile. Le generazioni sono state eseguite utilizzando come valori in ingresso $a = 0.0001$, $b = -10$ e $c = 10$. In Figura 3.7 sono riportati i grafici dei fit con i rispettivi residui.

Sono riportati i dati ottenuti dallo studio delle escursioni in Tabella 3.4 e gli istogrammi delle stesse in Figura 3.8.

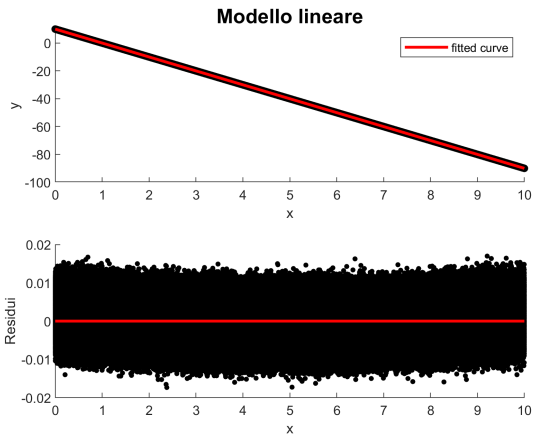
Dai grafici della distribuzione si nota che per $snr = 100$ la distribuzione delle escursioni si discosta altamente da una distribuzione geometrica. Nei grafici per $snr = 50$ si nota solamente una piccola differenza nei valori di lunghezza più bassi, mentre per rumori superiori lo scostamento è trascurabile.

I dati nelle tabelle mostrano che nella generazione con $snr = 100$ le escursioni massime negativa e positiva sono molto maggiori di quelle ottenute dai fit nel modello lineare. È quindi palese che il fit lineare si discosta per intervalli molto grandi dai punti della distribuzione. Questo è un chiaro indicatore del fatto che il fit non è corretto per la distribuzione generata.

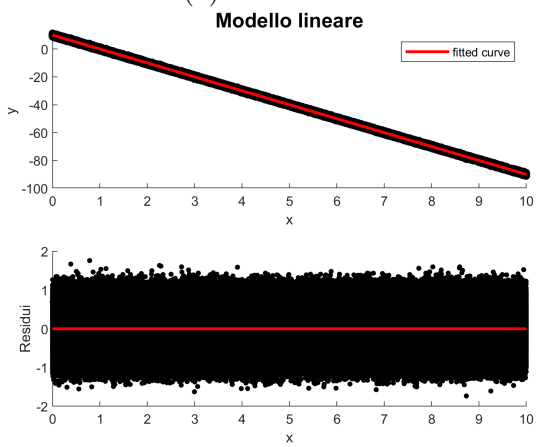
Anche il valore dell'escursione media si discosta in entrambi i casi dal valore osservato in precedenza. In particolare esso risulta maggiore nel caso dei residui negativi, mentre risulta minore nel caso dei residui positivi. I due valori sono dunque concordi nell'indicare che il caso negativo è quello con le escursioni più significative, che dunque alzano la



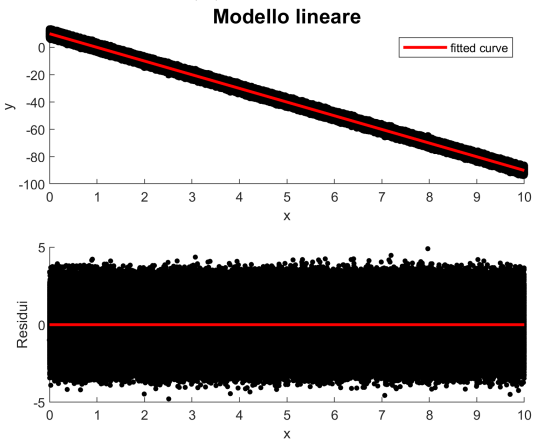
(a) $snr = 100$



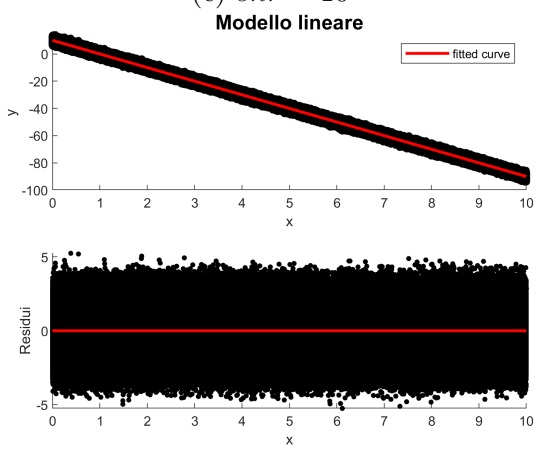
(b) $snr = 50$



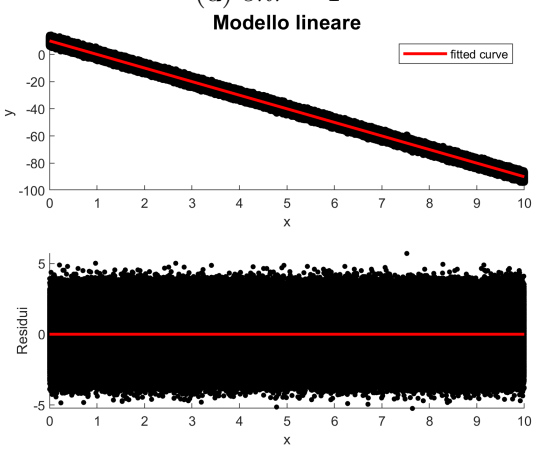
(c) $snr = 10$



(d) $snr = 1$

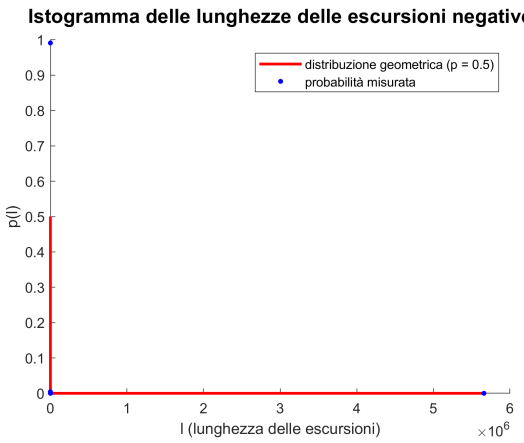


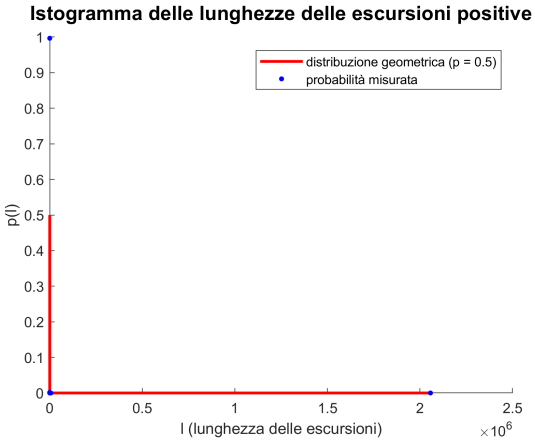
(e) $snr = 0.1$



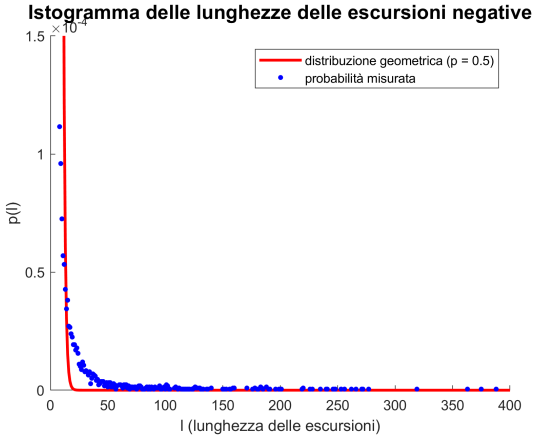
(f) $snr = 0.01$

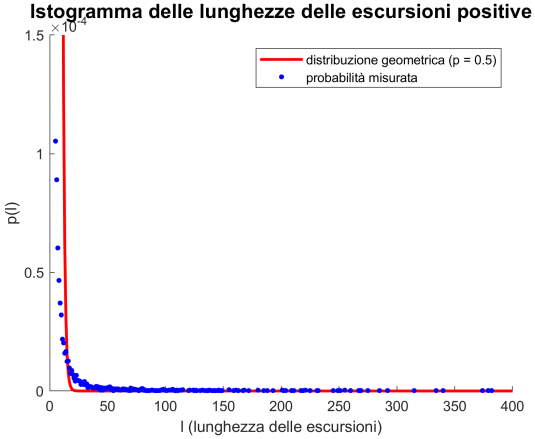
Figura 3.7: Generazioni paraboliche e rispettivi fit.



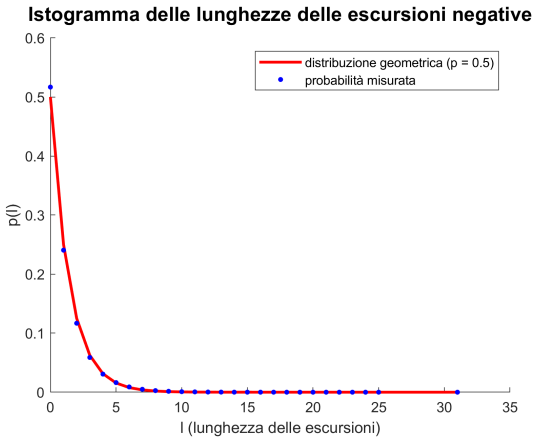


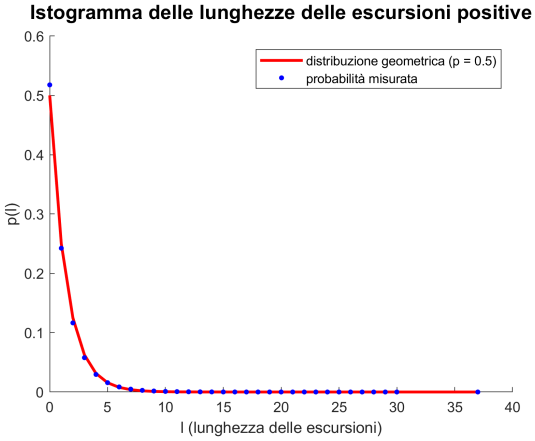
(a) $snr = 100$





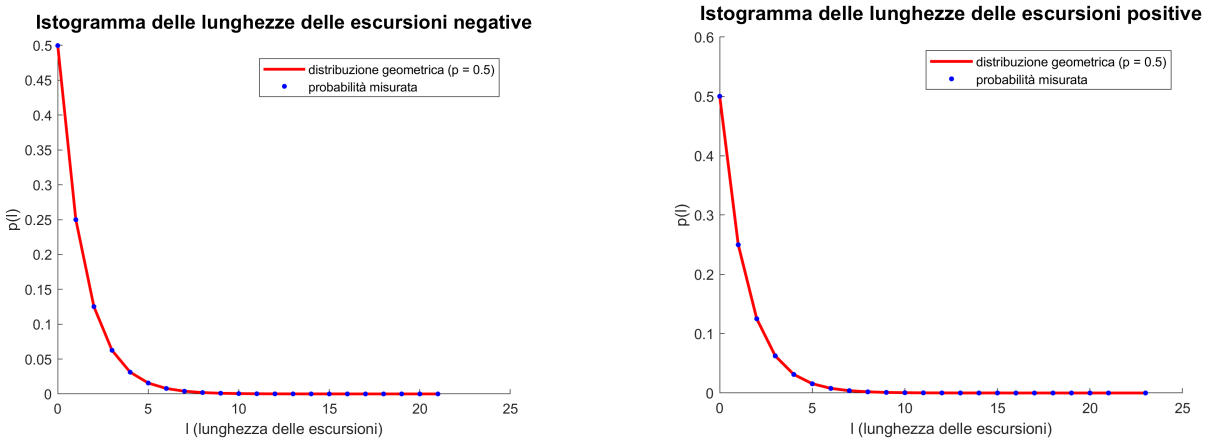
(b) $snr = 100$, dettaglio sui punti più prossimi all'origine al fine di mostrarne l'andamento



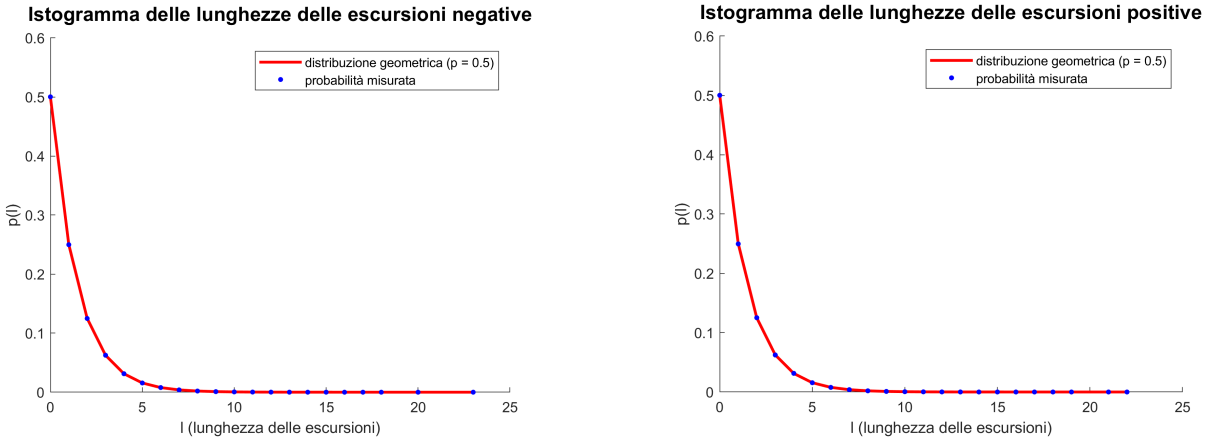


(c) $snr = 50$

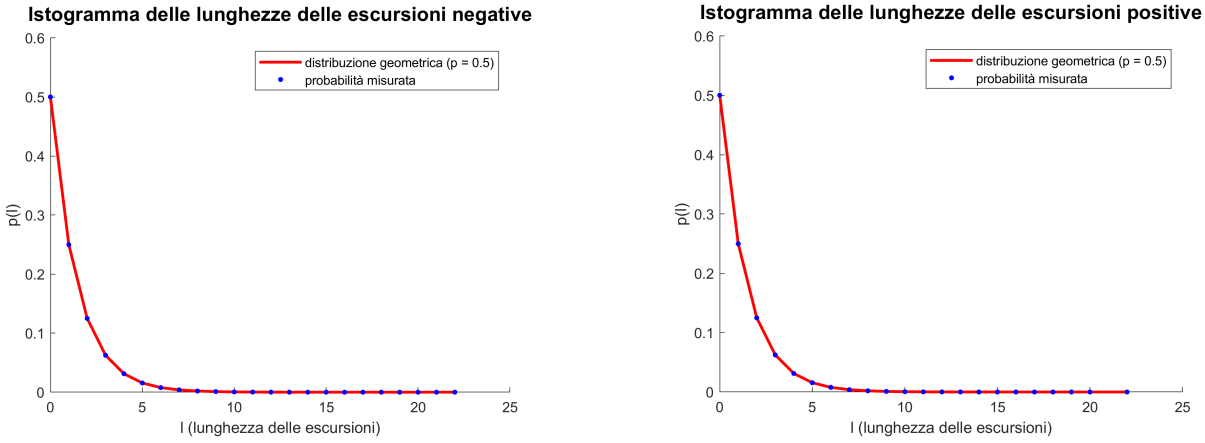
Figura 3.8: Istogrammi delle lunghezze delle escursioni delle generazioni paraboliche.



(d) $snr = 10$

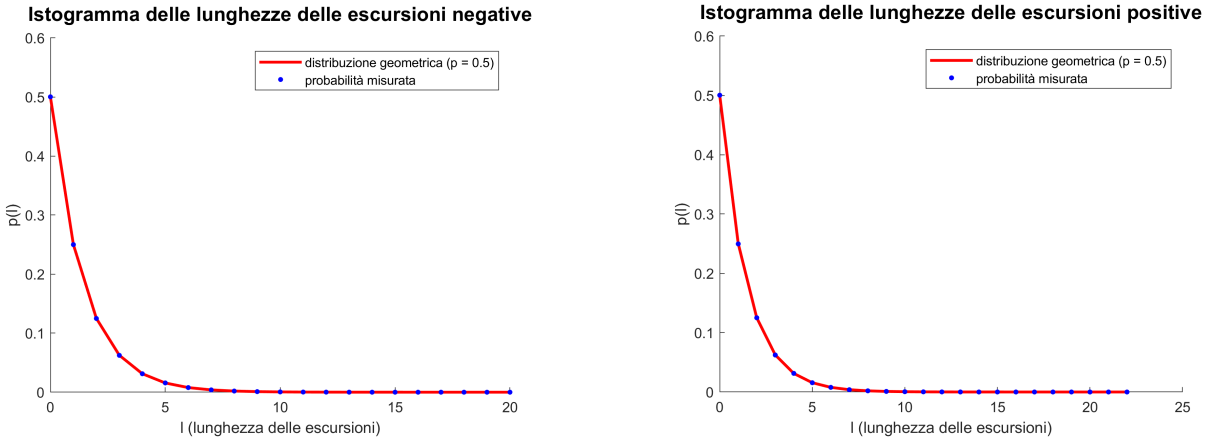


(e) $snr = 1$



(f) $snr = 0.1$

Figura 3.8: Istogrammi delle lunghezze delle escursioni delle generazioni paraboliche.



(g) $snr = 0.01$

Figura 3.8: Istogrammi delle lunghezze delle escursioni delle generazioni paraboliche.

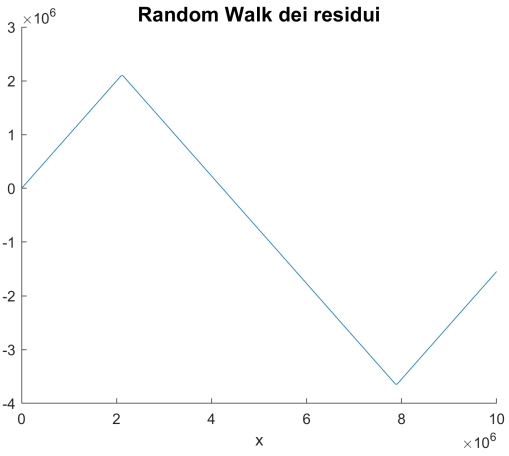
media nel caso negativo mentre abbassano la media nel caso positivo perché sono fonte di numerose escursioni con valore 0. La presenza di una quantità di esiti con valore 0 superiore al normale è osservabile anche dai grafici in Figura 3.8a e 3.8b.

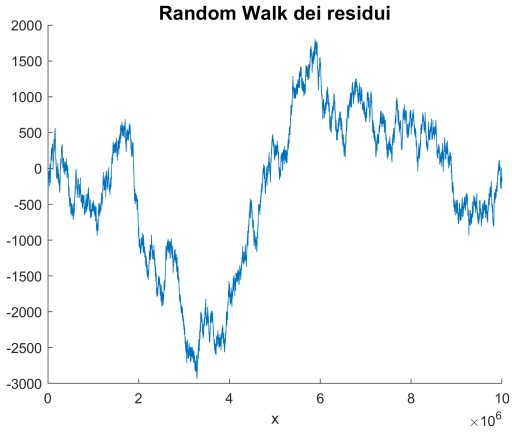
Lo scostamento dei parametri da quelli generati dal modello lineare va tuttavia perdendosi con l'aumento del rumore sui dati generati. Infatti, per $snr \leq 10$ i valori misurati sulle escursioni sono perfettamente in linea con quelli ottenuti nella Sezione 3.1. Lo studio del random walk dei residui perde dunque di efficacia. L'unico parametro che rimane anomalo in tutte le generazioni è l'escursione mediana, che assume solamente il valore 0.

Risulta interessante osservare l'andamento dei classici indicatori di bontà del fit r^2 e $rmse$. Quest'ultimo rileva un'anomalia del fit solamente nella prima generazione. Si osserva infatti che solo in questo caso il suo valore è un ordine di grandezza più grande rispetto alla generazione del modello lineare con stesso valore di snr. Risulta dunque meno efficace dello studio dell'escursione massima, che rivela un'anomalia anche per $snr = 50$. Il valore di r^2, al contrario, non rileva anomalie nei primi due casi, ma mostra una leggera decrescita nei casi successivi, dove tutti gli altri indicatori rimangono invece invariati.

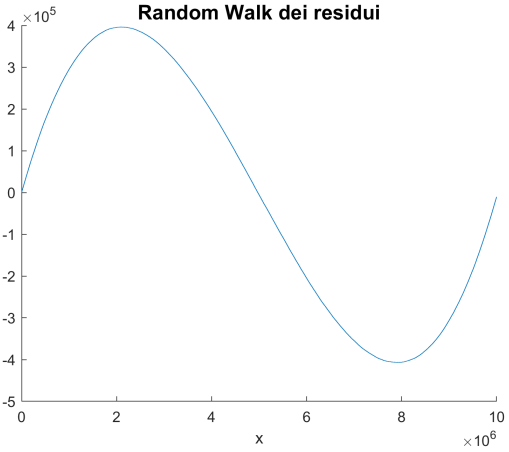
Una certa anomalia è osservabile anche dai grafici del random walk generato, riportati in Figura 3.9. I random walk costruiti sul modello parabolico (a sinistra) sono confrontati con quelli costruiti sulle generazioni di Figura 3.3 (a destra).

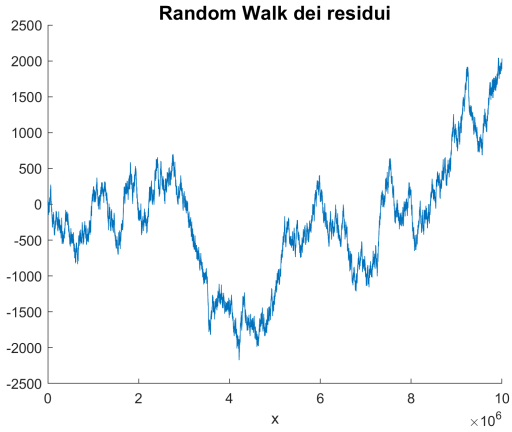
Mentre per $snr \geq 10$ è ancora chiaramente visibile l'andamento della curva, prima crescente, poi decrescente e infine ancora crescente, per snr inferiori il dato si perde progressivamente, nonostante il random walk rimanga in tutti i casi crescente nei primi punti e leggermente decrescente in tutti gli altri.



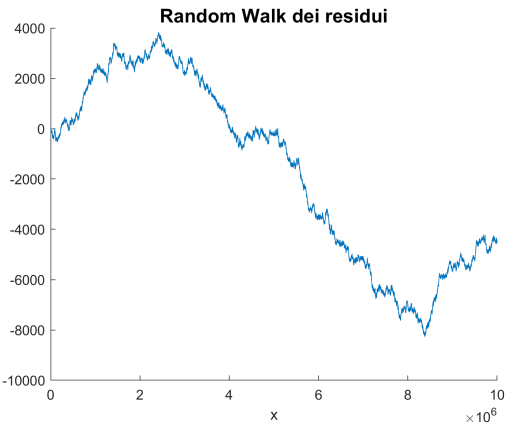


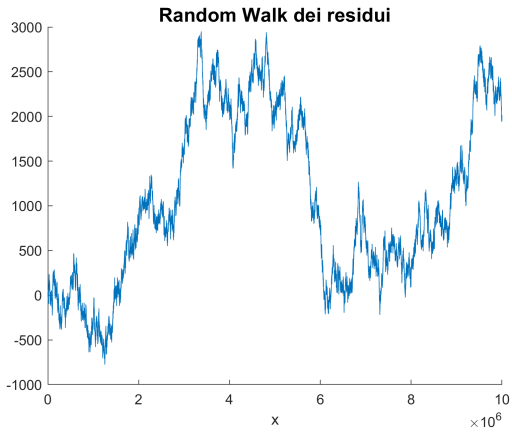
(a) $snr = 100$





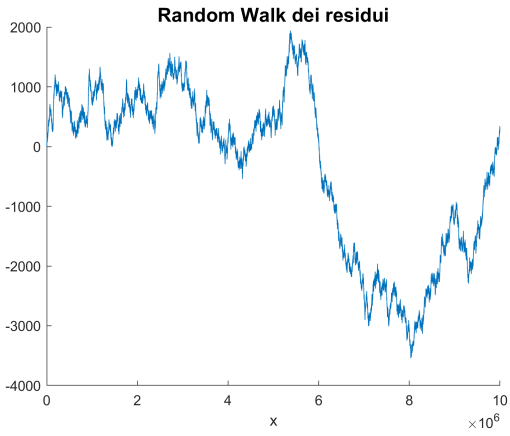
(b) $snr = 50$

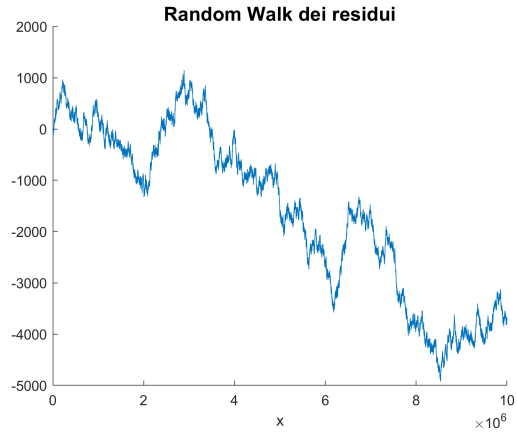




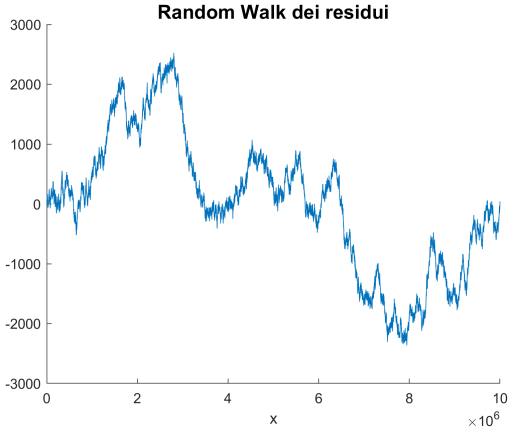
(c) $snr = 10$

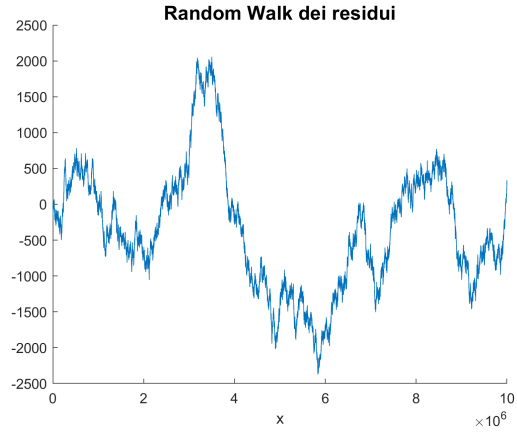
Figura 3.9: Random walk costruiti sui fit delle generazioni paraboliche (a sinistra), confrontati con i random walk costruiti con le generazioni lineari di Figura 3.3 (a destra).



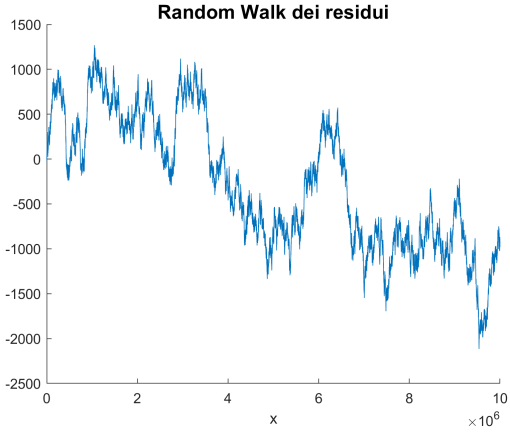


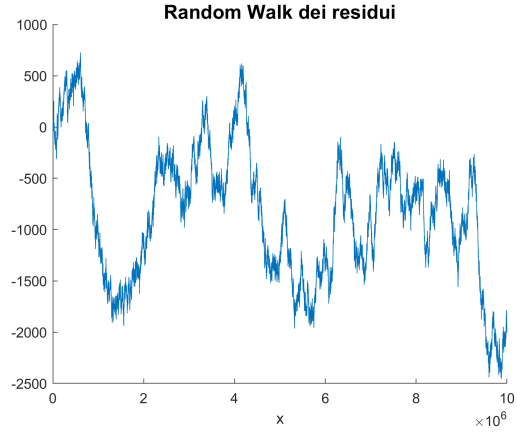
(d) $snr = 1$





(e) $snr = 0.1$





(f) $snr = 0.01$

Figura 3.9: Random walk costruiti sui fit delle generazioni paraboliche (a sinistra), confrontati con i random walk costruiti con le generazioni lineari di Figura 3.3 (a destra).

Capitolo 4

Conclusione

Dalle analisi riportate in Capitolo 3 emerge che l'utilizzo di parametri relativi ai random walk possono mostrare informazioni sulla bontà di alcuni fit mostrandone gli scostamenti dalla distribuzione fittata.

Dallo studio del modello lineare emerge che le escursioni sono distribuite secondo una distribuzione geometrica, come è atteso per l'aleatorietà del rumore gaussiano. Risulta che i parametri, in particolare la media, sono in linea con quelli attesi per una distribuzione di questo tipo. Dagli studi effettuati variando il numero di punti generati è emerso che i parametri rimangono invariati, eccetto l'escursione massima. Quest'ultima come è ovvio cresce poiché all'aumentare degli eventi si verificano risultati sempre meno probabili. Dalle generazioni lineari effettuate variando il rumore è emerso che se il fit è corretto i parametri rimangono invariati in ogni caso. Analogamente è stato osservato che essi non dipendono dall'intervallo di distribuzione dei dati. Infatti la variazione di quest'ultimo non provoca cambiamenti nella distribuzione.

Dallo studio del modello parabolico risulta che la variazione del rumore sui punti modifica la capacità dei nostri parametri di studiare la bontà del fit. Per valori di rumore bassi i parametri risultano scostarsi dai valori attesi. Le escursioni massime risultano molto maggiori, segno del fatto che per lunghi tratti il fit si trova separato dalla distribuzione. Questo fatto è chiaramente segnalato anche dall'escursione media, anch'essa infatti si discosta dai valori attesi. Anche l'escursione mediana subisce un'anomalia, assumendo solo il valore 0 e mai il valore 1. Questa variazione rispetto ai valori attesi decresce con l'aumentare del rumore, fino a non essere più apprezzabile per valori di rumore minori o uguali a 10. È interessante confrontare le variazioni nei parametri riguardanti i random walk con quelle nei parametri classici di misura della bontà di un fit r^2 e $rmse$. I parametri del random walk risultano più sensibili allo scostamento: essi mostrano anomalie per valori $snr = 50$ e $snr = 100$, mentre $rmse$ presenta una differenza dal valore atteso solo nel caso $rmse = 100$. r^2 invece rimane invariato in entrambi casi, ma scende leggermente per valori di rumore maggiori. Si può dedurre quindi che tale parametro non misuri lo scostamento del fit dalla distribuzione, ma solo la dispersione

dei dati.

Anche l'osservazione dei grafici dei random walk generati può essere un interessante parametro di valutazione della bontà del fit. È chiaramente visibile che fino a $snr = 10$, nonostante per questo valore i parametri non mostrino più alcuna anomalia, il grafico del random walk è chiaramente differente da quello atteso.

Dunque dagli studi fatti si può ritenere che lo studio del random walk generato dai residui possa essere un buon indicatore per studiare la bontà di un fit, poiché ne descrive lo scostamento dalla distribuzione originale in modo più sensibile rispetto ai classici parametri.

Bibliografia

[1] Johnston D., *An introduction to Random Walks*, 2011.

[2] Bernoulli trial, Wikipedia. URL:
`https://en.wikipedia.org/wiki/Bernoulli_trial`

[3] Geometric distribution, Wikipedia. URL:
`https://en.wikipedia.org/wiki/Geometric_distribution`

[4] Csáki E. e Hu Y., *Lengths and heights of random walk excursions*, DRW, 2003.

[5] Csaki E., Erdős P. e Révész P., *On the length of the longest excursion*, Probability Theory and Related Fields, 68, 365-382, 1985.