UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dipartimento di Ingegneria e Architettura

Corso di misure meccaniche, termiche e collaudi

Prof. Lucia Parussini
Prof. Rodolfo Taccani

a.a.2021-2022

Outline

- Analisi dei dati
- Regressione dei dati

Analisi dei dati

Presentazione delle serie di dati

- ❑ Tabelle e grafici di frequenza
- ❑ Tabelle e grafici di frequenza relativa
- ❑ Istogramma
- ❑ Ogiva
- ❑ Grafico a stelo e foglie

Analisi dei dati

Tabelle e grafici di frequenza

Un insieme di dati che ha un numero relativamente piccolo di valori distinti può essere convenientemente presentato in una tabella di frequenza.

Esempio:

Starting salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

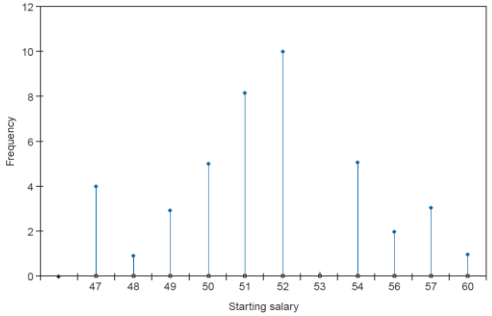
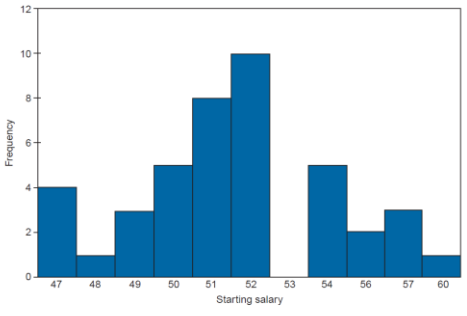
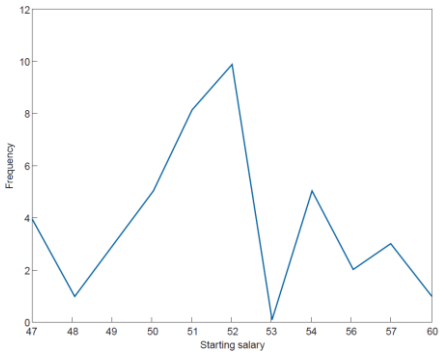


Grafico lineare.



Istogramma.



Poligono di frequenza.

Analisi dei dati

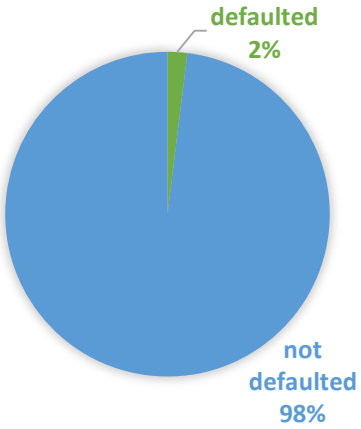
Tabelle e grafici di frequenza relativa

Si consideri un insieme di dati composto da n valori. Se f è la frequenza di un particolare valore, allora il rapporto f/n è chiamato la sua frequenza relativa.

Esempi:

Starting salary	Relative frequency
47	4/42 = 0.0952
48	1/42 = 0.0238
49	3/42 = 0.0714
50	5/42 = 0.119
51	8/42 = 0.1905
52	10/42 = 0.2381
53	0/42 = 0
54	5/42 = 0.119
56	2/42 = 0.0476
57	3/42 = 0.0714
60	1/42 = 0.0238

Type of result	Frequency	Relative frequency
defaulted	20	0.020
not defaulted	1000	0.980



Analisi dei dati

Istogramma

Per serie di dati in cui il numero di valori distinti è grande, è utile dividere i valori in raggruppamenti, o intervalli di classe, e poi tracciare il numero di valori di dati che rientrano in ogni intervallo di classe.

Gli **istogrammi** consentono di aggregare dati numerici in gruppi di intervalli uguali. Agli intervalli viene spesso dato il nome di bin. Ci sono due regole alternative per calcolare il miglior numero di bin da usare:

La **regola di Sturgis** calcola il numero di bande come:

$$Number\ of\ bins = 1 + 3.3log_{10}(n)$$

dove n è il numero di valori misurati.

La **regola di Rice** calcola il numero di bande come:

$$Number\ of\ bins = 2n^{1/3}$$

Ovviamente il risultato prodotto deve essere arrotondato al numero intero più vicino in entrambi i casi.

Analisi dei dati

Istogramma

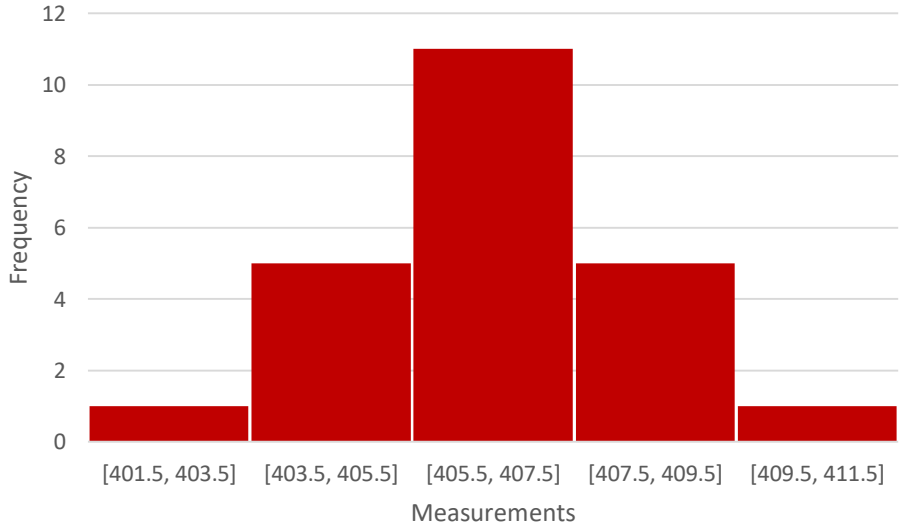
Number of measurements	Number of bins calculated by Sturgis rule	Number of bins by Sturgis (rounded)	Number of bins calculated by Rice rule	Number of bins by Rice (rounded)
10	4.30	4	4.31	4
15	4.88	5	4.93	5
20	5.29	5	5.43	5
25	5.61	6	5.85	6
30	5.87	6	6.21	6
50	6.61	7	7.37	7
100	7.60	8	9.28	9
200	8.59	9	11.70	12

Analisi dei dati

Istogramma

Esempio: Disegna un istogramma per le 23 misure di lunghezza [mm] del set di dati.

numero misura	*valore misura*	*numero misura*	*valore misura*	*numero misura*	*valore misura*
1	409	10	407	19	406
2	406	11	408	20	405
3	402	12	406	21	409
4	407	13	410	22	406
5	405	14	406	23	407
6	404	15	405		
7	407	16	408		
8	404	17	406		
9	407	18	409		

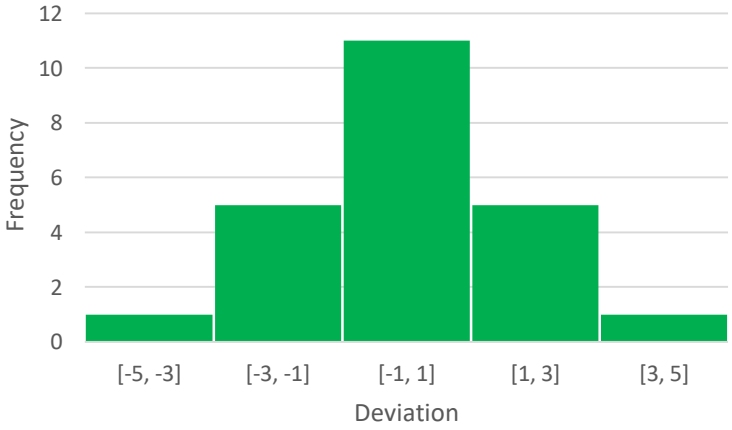


Analisi dei dati

Istogramma

Esempio: Disegna un istogramma per le 23 misure di lunghezza [mm] del set di dati.

E' spesso più utile disegnare un istogramma delle deviazioni delle misure dal valore medio piuttosto che disegnare un istogramma delle misure stesse.

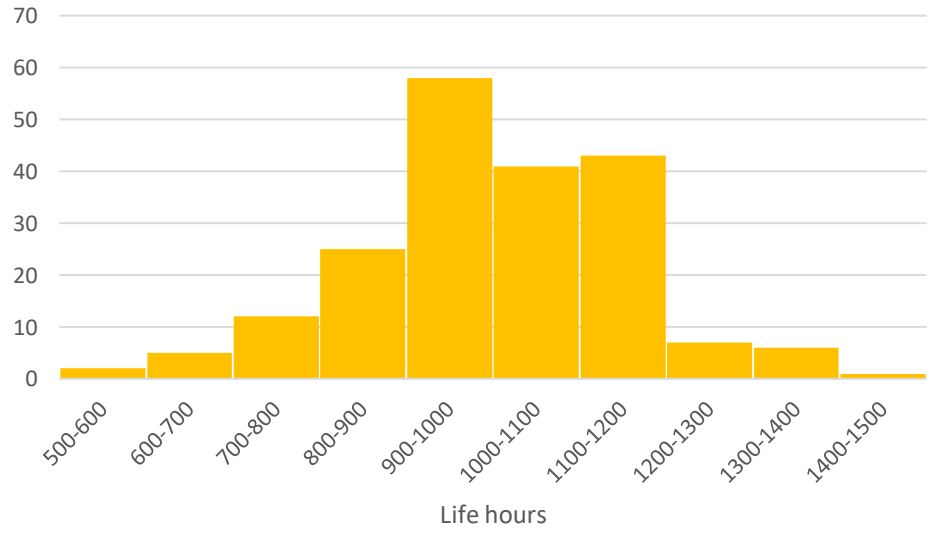


Analisi dei dati

Istogramma

Esempio: durata di vita di 200 lampade a incandescenza in ore.

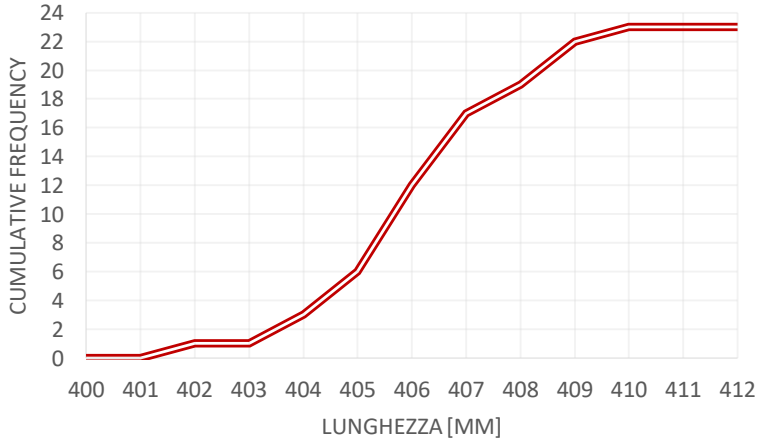
Class interval	Frequency
500-600	2
600-700	5
700-800	12
800-900	25
900-1000	58
1000-1100	41
1100-1200	43
1200-1300	7
1300-1400	6
1400-1500	1

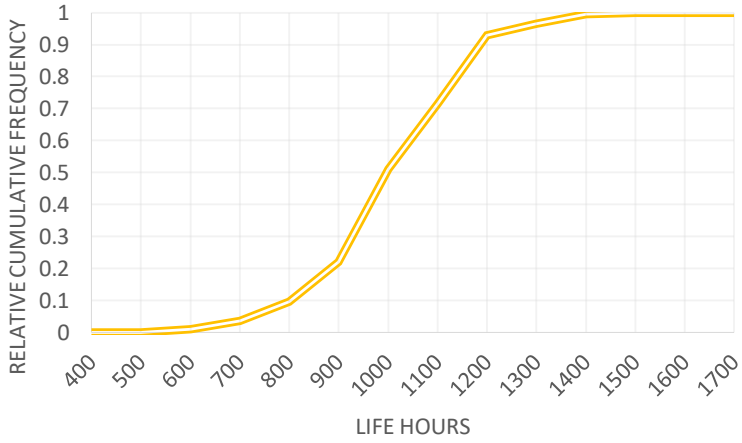


Analisi dei dati

Ogiva o grafico di frequenza cumulativa

Un punto sull'asse orizzontale di tale grafico rappresenta un possibile valore dei dati; il suo corrispondente valore sull'asse verticale fornisce il numero dei dati i cui valori sono inferiori o uguali ad esso.





Analisi dei dati

Grafico a stelo e foglie

E' un modo efficiente di organizzare una serie di dati di piccole e medie dimensioni. Si ottiene dividendo prima ogni valore di dati in due parti - il suo gambo e la sua foglia. Per esempio, se i dati sono tutti numeri a due cifre, allora potremmo lasciare che la parte del gambo di un valore di dati sia la sua cifra delle decine e che la foglia sia la sua cifra delle unità. Così, per esempio, il valore 62 è espresso come

Stelo	Foglia
6	2

Esempio: Le temperature minime medie giornaliere annuali in 35 città degli Stati Uniti possono essere rappresentate nel seguente grafico a stelo e foglie.

7	0.0
6	9.0
5	1.0,1.3,2.0,5.5,7.1,7.4,7.6,8.5,9.3
4	0.0,1.0,2.4,3.6,3.7,4.8,5.0,5.2,6.0,6.7,8.1,9.0,9.2
3	3.1,4.1,5.3,5.8,6.2,9.0,9.5,9.5
2	9.0,9.8

Analisi dei dati

Outlier

E’ un termine utilizzato in statistica per definire, in un insieme di osservazioni, un valore anomalo e aberrante, ossia un valore chiaramente distante dalle altre osservazioni disponibili. Nel campo dell’analisi sperimentale è frequente trovare, in una serie di misure, qualche dato che **non concorda con gli altri.**

In questi casi è prassi comune scartare queste misurazioni errate. Un livello di soglia spesso usato per determinare cosa dovrebbe essere scartato è $\pm 3\sigma$.

L'implementazione pratica di una tale procedura deve essere fatta con attenzione.

Analisi dei dati

Outlier

Esempio:
Viene effettuata una serie di misurazioni con un nuovo trasduttore di pressione. L'ispezione di un istogramma delle prime 20 misurazioni non mostra dati anomali. Non si osservano dati anomali.
$\mu = 4.41\,\text{bar} \quad \sigma = 0.05\,\text{bar}$

Viene effettuata un'ulteriore serie di misurazioni:

4.35	4.46	4.39	4.34	4.41	4.52	4.44	4.37	4.41	4.33	4.39	4.47	4.42	4.59	4.45	4.38	4.43	4.36	4.48	4.45

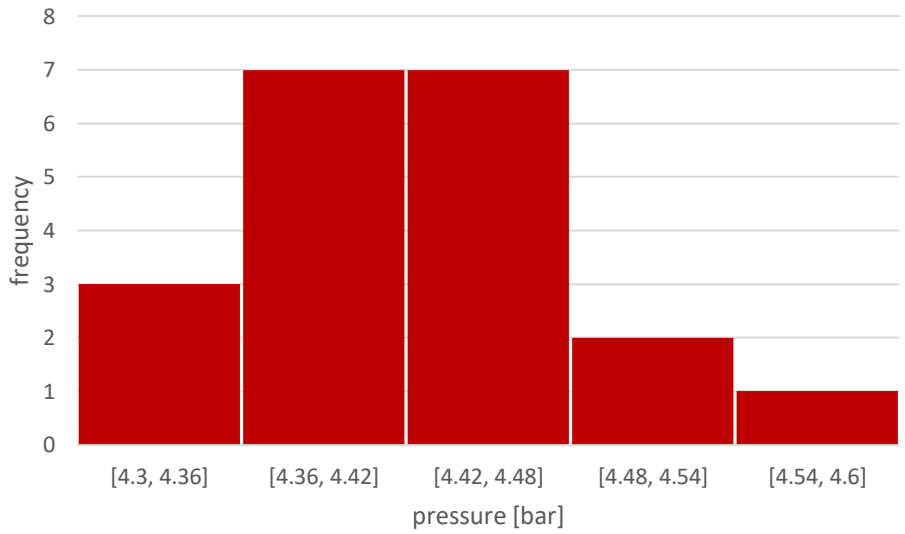
Usate la soglia di $\pm 3\sigma$ per determinare se ci sono punti di dati anomali nel set di misurazione.
Soluzione
Poiché il valore σ calcolato per un insieme di buone misurazioni è 0.05 bar, la soglia $\pm 3\sigma$ è ±0.15. Con un valore medio dei dati di 4.41, la soglia per i punti di dati anomali è costituita da valori inferiori a 4.26 ($\mu - 3\sigma$) o superiori a 4.56 ($\mu + 3\sigma$). Guardando l'insieme delle misurazioni, osserviamo che la misura di 4.59 è al di fuori della soglia di $\pm 3\sigma$, indicando che si tratta di un dato anomalo.

4.35	4.46	4.39	4.34	4.41	4.52	4.44	4.37	4.41	4.33	4.39	4.47	4.42	4.59	4.45	4.38	4.43	4.36	4.48	4.45

Analisi dei dati

Outlier

Esempio:



4.35 4.46 4.39 4.34 4.41 4.52 4.44 4.37 4.41 4.33 4.39 4.47 4.42 **4.59** 4.45 4.38 4.43 4.36 4.48 4.45

E' importante garantire che non ci siano outlier nell'insieme dei dati utilizzati per calcolare la deviazione standard dei dati e quindi la soglia per rifiutare outlier.

sui 19 valori escludendo la misura 4.59
$\mu = 4.41$ bar $\sigma = 0.052$ bar $\pm 3\sigma = \pm 0.16$ bar
$\mu - 3\sigma = 4.25$
$\mu + 3\sigma = 4.57$

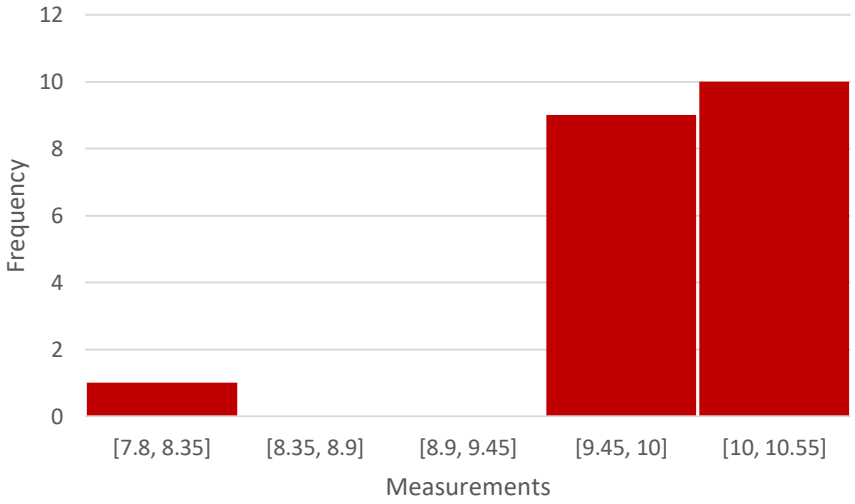
sui 20 valori **non escludendo** la misura 4.59
$\mu = 4.42$ bar $\sigma = 0.064$ bar $\pm 3\sigma = \pm 0.19$ bar
$\mu - 3\sigma = 4.23$
$\mu + 3\sigma = 4.61$

Analisi dei dati

Outlier

Esempio: Consideriamo il seguente set di dati di 20 misurazioni ed esaminiamolo alla ricerca di outlier:

10.4 9.9 9.7 9.6 10.1 10.3 9.8 10 10.2 9.5 9.8 10.1 10.3 8.1 9.7 10.2 10.3 9.7 9.9 10.2



La misura di 8.1 sembra chiaramente un outlier!

La media e la deviazione standard degli altri 19 punti (escluso il valore di 8.1) sono

$\mu = 9.98 \quad \sigma = 0.273 \quad \pm 3\sigma = \pm 0.819$ bar
$\mu - 3\sigma = 9.16$
$\mu + 3\sigma = 10.80$

Questo conferma che il dato 8.1 è ben al di fuori dei limiti di $\pm 3\sigma$ ed è quindi confermato come un outlier.

Analisi dei dati

Outlier

Criterio di Chauvenet

Consente di valutare se il dato anomalo è dovuto ad **errori grossolani** o, al contrario, se è rappresentativo di una **misura plausibile.**

In una serie di n dati sperimentali, se alcuni valori presentano uno scostamento dal valore medio che ha probabilità di verificarsi inferiore di $1/(2n)$, allora quei valori devono essere scartati.

Analisi dei dati

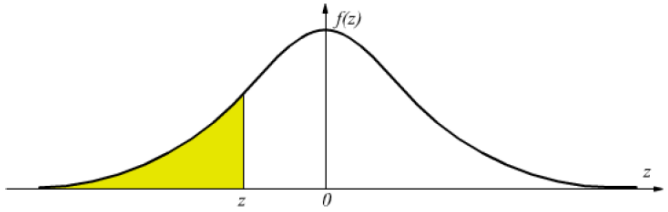
Outlier

Criterio di Chauvenet

Procedimento

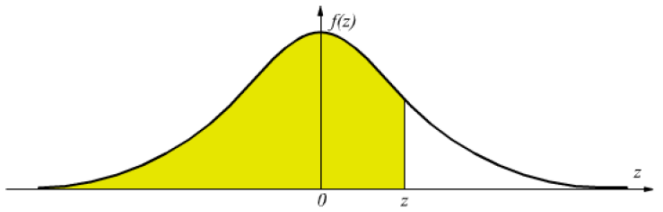
1) Si calcola la media, varianza e quindi lo scarto ridotto (o scarto standardizzato o scarto normale) dei dati $z_i = \frac{x_i - \mu}{\sigma}$
2) Si calcola la probabilità di Chauvenet $p_C = 1 - \frac{1}{2n}$
3) Si calcola la frequenza cumulata, ricordando che $p_C = 2F(z_{lim}) - 1$. Per cui $F(z_{lim}) = \frac{p_C + 1}{2}$
4) Si determina il valore dello scarto ridotto corrispondente al valore limite z_{lim}
5) Si escludono i valori per i quali il valore assoluto dello scarto ridotto è superiore a z_{lim}
6) Si ricalcola la media e lo scarto quadratico con i dati restanti, ma non è più lecito applicare di nuovo il criterio ai dati rimasti.

Analisi dei dati



Probabilità cumulativa per valori negativi di z

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,1	0,001	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,001	0,001
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,002	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,003	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,004	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,006	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,008	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,011
-2,1	0,0179	0,0174	0,017	0,0166	0,0162	0,0158	0,0154	0,015	0,0146	0,0143
-2	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,025	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,063	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,102	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,123	0,121	0,119	0,117
-1	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,166	0,1635	0,1611
-0,8	0,2119	0,209	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,242	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,305	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,281	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,33	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,352	0,3483
-0,2	0,4207	0,4168	0,4129	0,409	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0	0,5	0,496	0,492	0,488	0,484	0,4801	0,4761	0,4721	0,4681	0,4641



Probabilità cumulativa per valori positivi di z

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,648	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,67	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,719	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,791	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,834	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,877	0,879	0,881	0,883
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,898	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,937	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,975	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,983	0,9834	0,9838	0,9842	0,9846	0,985	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,989
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,992	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,994	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,996	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,997	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,998	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,999	0,999
3,1	0,999	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Analisi dei dati

Outlier

Criterio di Chauvenet

Esempio: Supponiamo di aver fatto 8 misure della grandezza X:

1	2	3	4	5	6	7	8
14.1	13.4	13.8	13.0	11.8	14.1	13.0	14.0

1) Calcoliamo $\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = 13.4$ e $\sigma = \sqrt{\frac{1}{7}\sum_{i=1}^{8}(x_i - \bar{x})^2} = 0.8$ e lo scarto normale:

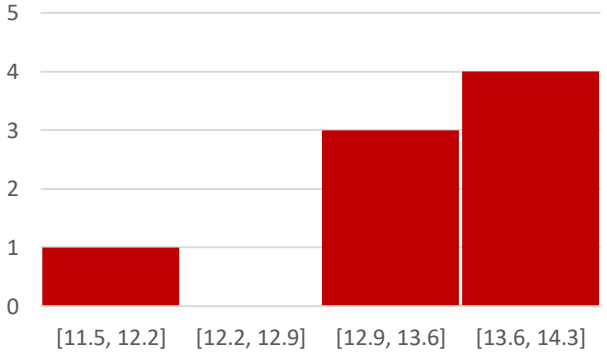
1	2	3	4	5	6	7	8
0.88	0.00	0.51	-0.51	-2.02	0.88	-0.51	0.76

2) $p_C = 1 - \frac{1}{2n} = 1 - \frac{1}{16} = 0.94$

3) $F(z_{lim}) = \frac{p_C + 1}{2} = 0.97$

4) $z_{lim} = 1.86$

5) Si **scarta** il dato $|x_5| > z_{lim}$



Analisi dei dati

Outlier

Criterio di Chauvenet

Esempio: Riprendiamo le misure di pressione precedenti.

4.35	4.46	4.39	4.34	4.41	4.52	4.44	4.37	4.41	4.33	4.39	4.47	4.42	4.59	4.45	4.38	4.43	4.36	4.48	4.45

1) Calcoliamo $\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i$= 4.42 e $\sigma = \sqrt{\frac{1}{19}\sum_{i=1}^{20}(x_i - \bar{x})^2}$ = 0.06 e lo scarto normale:

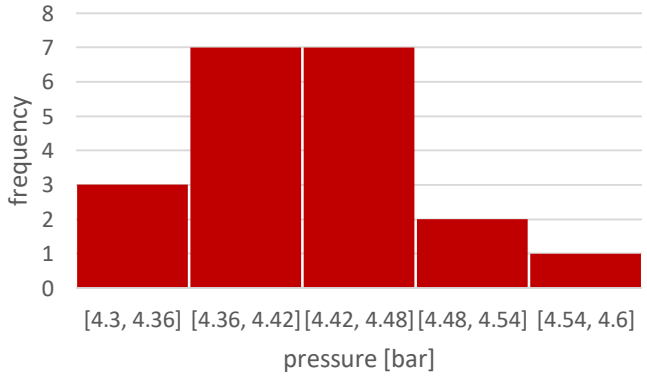
-1.12	0.59	-0.50	-1.28	-0.19	1.53	0.28	-0.81	-0.19	-1.44	-0.50	0.75	-0.03	2.62	0.44	-0.66	0.12	-0.97	0.91	0.44

2) $p_C = 1 - \frac{1}{2n} = 1 - \frac{1}{40} = 0.975$

3) $F(z_{lim}) = \frac{p_C+1}{2} = 0.9875$

4) z_{lim} =2.24

5) Si **scarta** il dato $|2.62| > z_{lim}$, quindi 4.59



Analisi dei dati

Come riassumere i dati

Media del campione, mediana del campione e moda del campione

Esempio:
La lunghezza di una barra d'acciaio viene misurata in millimetri da diversi osservatori (11 misure).
Measurement set A

398	420	394	416	404	408	400	420	396	413	430

Le misure sono fatte nuovamente usando una precedura di misura migliore e con gli osservatori che fanno più attenzione.
Measurement set B

409	406	402	407	405	404	407	404	407	407	408

Aumentiamo il numero di misurazioni estendendo il set di misurazioni B a 23 misurazioni.
Measurement set C

409	406	402	407	405	404	407	404	407	407	408	406	410	406	405	408	406	409	406	405	409	406	407

Analisi dei dati

Come riassumere i dati

Media del campione, mediana del campione e moda del campione

Measurement set A

$$x_{mean} = 409.0$$
$$x_{median} = 408$$
$$x_{mode} = 420$$

Measurement set B

$$x_{mean} = 406.0$$
$$x_{median} = 407$$
$$x_{mode} = 407$$

Measurement set C

$$x_{mean} = 406.5$$
$$x_{median} = 406$$
$$x_{mode} = 406$$

Come riassumere i dati

Varianza e deviazione standard del campione

deviazione (errore) $d_i = x_i - x_{mean}$

$$\sigma^2 = \frac{d_1^2 + d_2^2 + \cdots + d_n^2}{n}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{d_1^2 + d_2^2 + \cdots + d_n^2}{n}}$$

Le definizioni formali per la varianza e la deviazione standard dei dati sono fatte rispetto a una popolazione infinita di valori di dati, mentre in tutte le situazioni pratiche, possiamo avere solo un insieme finito di misure.

Come riassumere i dati

Varianza e deviazione standard del campione

Una migliore previsione della varianza della popolazione infinita può essere ottenuta applicando il fattore di correzione di Bessel $\frac{n}{n-1}$:

$$s^2 = \frac{n}{n-1}\sigma^2 = \frac{d_1^2 + d_2^2 + \cdots + d_n^2}{n-1}$$

$$s = \sqrt{s^2} = \sqrt{\frac{d_1^2 + d_2^2 + \cdots + d_n^2}{n-1}}$$

s^2 è la varianza dell'insieme finito di misure e σ^2 è la varianza della popolazione infinita di misure.

Analisi dei dati

Come riassumere i dati

Varianza e deviazione standard del campione

Esempio:

Set A		Measurement	Deviation	(Deviations)^2
	1	398	-11.0	121
	1	420	11.0	121
	1	394	-15.0	225
	1	416	7.0	49
	1	404	-5.0	25
	1	408	-1.0	1
	1	400	-9.0	81
	1	420	11.0	121
	1	396	-13.0	169
	1	413	4.0	16
	1	430	21.0	441
Sum	11	4499	0	1370

Mean	variance	standard deviation
409.0	137.0	11.7

Analisi dei dati

Come riassumere i dati

Varianza e deviazione standard del campione

Set B		Measurement	Deviation	(Deviations)^2
	1	409	3.0	9
	1	406	0.0	0
	1	402	-4.0	16
	1	407	1.0	1
	1	405	-1.0	1
	1	404	-2.0	4
	1	407	1.0	1
	1	404	-2.0	4
	1	407	1.0	1
	1	407	1.0	1
	1	408	2.0	4
Sum	11	4466	0	42

Mean	variance	standard deviation
406.0	4.20	2.05

Analisi dei dati

Come riassumere i dati

Varianza e deviazione standard del campione

Set C		Measurement	Deviation	(Deviations)^2
	1	409	2.5	6.4
	1	406	-0.5	0.2
	1	402	-4.5	20.1
	1	407	0.5	0.3
	1	405	-1.5	2.2
	1	404	-2.5	6.1
	1	407	0.5	0.3
	1	404	-2.5	6.1
	1	407	0.5	0.3
	1	407	0.5	0.3
	1	408	1.5	2.3
	1	406	-0.5	0.2
	1	410	3.5	12.4
	1	406	-0.5	0.2
	1	405	-1.5	2.2
	1	408	1.5	2.3
	1	406	-0.5	0.2
	1	409	2.5	6.4
	1	406	-0.5	0.2
	1	405	-1.5	2.2
	1	409	2.5	6.4
	1	406	-0.5	0.2
	1	407	0.5	0.3
Sum	23	9349	6.25E-13	77.7

mean	variance	standard deviation
406.5	3.53	1.88

Come riassumere i dati

Percentili del campione e box plot

Definizione

Il 100p percentile del campione è quel valore di dati tale che il 100p per cento dei dati è inferiore o uguale ad esso e il 100(1-p) per cento è maggiore o uguale ad esso. Se due valori di dati soddisfano questa condizione, allora il 100p percentile del campione è la media aritmetica di questi due valori.

Per determinare il 100p percentile del campione di un insieme di dati di dimensioni n, dobbiamo determinare i valori dei dati tali che

1. Almeno np dei valori siano inferiori o uguali ad esso.
2. Almeno n(1-p) dei valori siano maggiori o uguali ad esso.

Come riassumere i dati

Percentili del campione e box plot

Definizione

Il 25° percentile del campione è chiamato **primo quartile**; il 50° percentile del campione è chiamato mediana del campione o **secondo quartile**; il 75° percentile del campione è chiamato **terzo quartile**.
I quartili dividono una serie di dati in quattro parti, con circa il 25% dei dati che sono inferiori al primo quartile, il 25% tra il primo e il secondo quartile, il 25% tra il secondo e il terzo quartile e il 25% superiore al terzo quartile.

Analisi dei dati

Come riassumere i dati

Percentili del campione e box plot

Starting salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

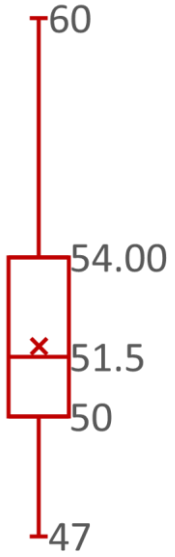
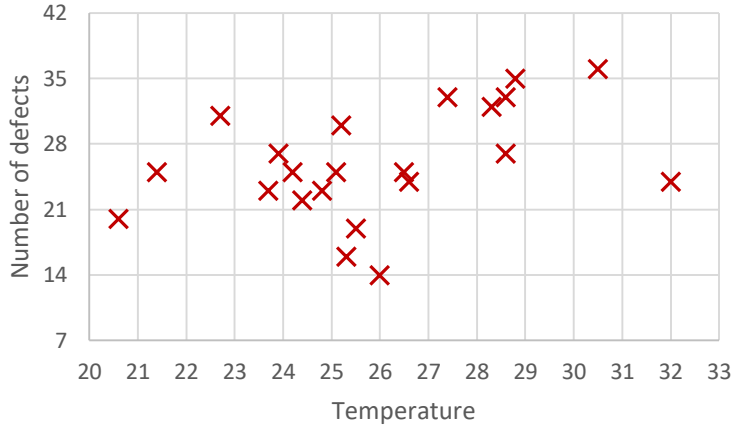


Diagramma a scatola e baffi o diagramma degli estremi e dei quartili o box and whiskers plot o box-plot.

Analisi dei dati

Serie di dati appaiati e coefficiente di correlazione

Esempio: nel tentativo di determinare la relazione tra la temperatura giornaliera di mezzogiorno (misurata in gradi Celsius) e il numero di pezzi difettosi prodotti durante quel giorno, un'azienda ha registrato i dati presentati nella tabella. Per questa serie di dati, x_i rappresenta la temperatura in gradi Celsius e y_i il numero di pezzi difettosi prodotti il giorno i.



Day	Temperature	Number of defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

Analisi dei dati

Serie di dati appaiati e coefficiente di correlazione

Supponiamo che l'insieme di dati sia costituito da valori appaiati $(x_i, y_i), i = 1, \dots n$.

Definizione
Lasciamo che $\bar{x}$ e $\bar{y}$ denotino le medie campionarie dei valori x e dei valori y, rispettivamente e s_x e s_y denotino, rispettivamente, le deviazioni standard del campione dei valori x e dei valori y. Il **coefficiente di correlazione** campionaria, detto r, delle coppie di dati $(x_i, y_i), i = 1, \dots n$ è definito da

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x\, s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Quando $r > 0$ diciamo che i dati del campione sono correlati positivamente e quando $r < 0$ diciamo che i dati del campione sono correlati negativamente.

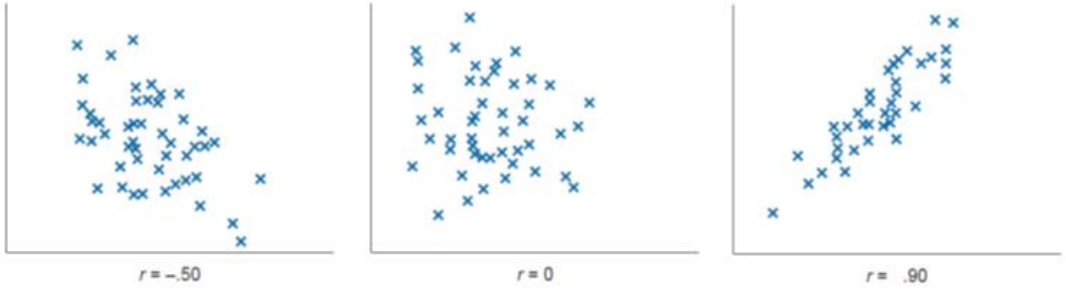
Serie di dati appaiati e coefficiente di correlazione

Proprietà di r

- $-1 \leq r \leq 1$
- Se per costanti a e b, con $b > 0$
  $y_i = a + bx_i,\ i = 1, \dots n$
  allora $r = 1$.
- Se per le costanti a e b, con $b < 0$
  $y_i = a + bx_i,\ i = 1, \dots n$
  allora $r = -1$.
- Se r è il coefficiente di correlazione campionaria per le coppie di dati $x_i\ , y_i\ , i = 1, \dots n$ allora esso è anche il coefficiente di correlazione campionaria per le coppie di dati
  $a + bx_i,\ \ c + dy_i,\ \ \ i = 1, \dots n$
  a condizione che b e d siano entrambi positivi o entrambi negativi.

Serie di dati appaiati e coefficiente di correlazione

La figura mostra i diagrammi di dispersione per serie di dati con vari valori di r.



Regressione dei dati

Molti problemi ingegneristici e scientifici riguardano la determinazione di una relazione tra un insieme di variabili.

In molte situazioni, c'è una singola variabile di risposta Y, chiamata anche **variabile dipendente**, che dipende dal valore di un insieme di variabili di input, chiamate anche **variabili indipendenti**, $x_1, \dots, x_r$.

Il tipo più semplice di relazione tra la variabile dipendente Y e le variabili di input $x_1, \dots, x_r$ è una relazione lineare. Si potrebbe scrivere, per alcune costanti $\beta_0, \beta_1, \dots \beta_r$ l'equazione

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

Se questa fosse la relazione tra Y e gli $x_i, i = 1, \dots, r$, allora sarebbe possibile (una volta appresi i β_i) prevedere esattamente la risposta per qualsiasi insieme di valori di input.

Regressione dei dati

Tuttavia, in pratica, l'equazione precedente è valida introducendo un errore casuale.
La relazione esplicita è

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e$$

dove e, errore casuale, si assume essere una variabile casuale con media 0, o equivalentemente

$$E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

dove $\mathbf{x} = (x_1, \ldots, x_r)$ è l'insieme delle variabili indipendenti e $E[Y|\mathbf{x}]$ è la risposta attesa dati gli input $\mathbf{x}$.

L'equazione è chiamata **equazione di regressione lineare**.

Le quantità $\beta_0, \beta_1, \ldots \beta_r$ sono chiamate **coefficienti di regressione**.

Regressione dei dati

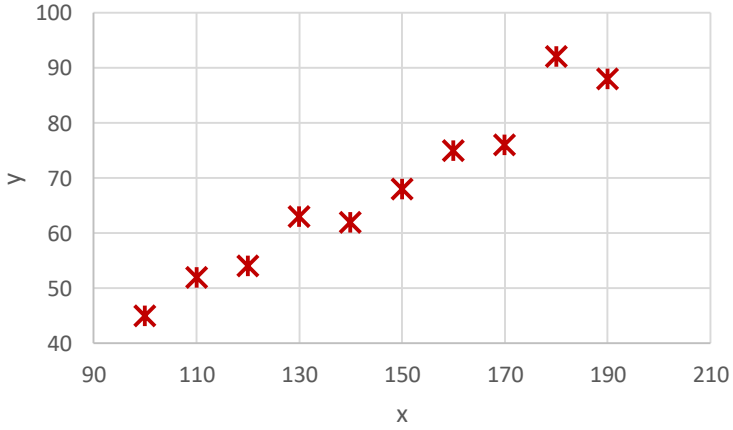
Regressione semplice

Un'equazione di regressione che contiene una sola variabile indipendente - cioè $r = 1$ - è chiamata equazione di regressione semplice:

$$Y = \alpha + \beta x + e$$

ESEMPIO: Consideriamo le seguenti 10 coppie di dati $(x_i, y_i), i = 1, \ldots 10$, che mettono in relazione y, il risultato percentuale di un esperimento di laboratorio, a x, la temperatura alla quale l'esperimento è stato eseguito.

i	x_i	y_i	i	x_i	y_i
1	100	45	6	150	68
2	110	52	7	160	75
3	120	54	8	170	76
4	130	63	9	180	92
5	140	62	10	190	88



Regressione dei dati

Stimatori ai minimi quadrati dei parametri di regressione

Per determinare gli stimatori di α e β ragioniamo come segue: Se A è lo stimatore di α e B di β, allora lo stimatore della risposta corrispondente alla variabile di ingresso x_i sarebbe $A + Bx_i$. Poiché la risposta effettiva è Y_i, la differenza al quadrato è $(Y_i - A - Bx_i)^2$, e quindi se A e B sono gli stimatori di α e β, allora la somma delle differenze al quadrato tra le risposte stimate e i valori della risposta effettiva - chiamiamola SS - è data da

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

Regressione dei dati

Stimatori ai minimi quadrati dei parametri di regressione

Per determinare gli stimatori di α e β, che minimizzano SS, differenziamo SS prima rispetto ad A e B:

$$\frac{\partial SS}{\partial A} = -2\sum_{i=1}^{n}(Y_i - A - Bx_i) \qquad \frac{\partial SS}{\partial B} = -2\sum_{i=1}^{n} x_i(Y_i - A - Bx_i)$$

ponendo le derivate pari a 0:

$$\sum_{i=1}^{n}(Y_i - A - Bx_i) = 0 \qquad \sum_{i=1}^{n} x_i(Y_i - A - Bx_i) = 0$$

Quindi otteniamo le equazioni:

$$\sum_{i=1}^{n} Y_i = nA + B\sum_{i=1}^{n} x_i \qquad \sum_{i=1}^{n} x_i Y_i = A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2$$

note come **equazioni normali.**

Regressione dei dati

Stimatori ai minimi quadrati dei parametri di regressione

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = A + B\frac{1}{n}\sum_{i=1}^{n} x_i \quad \rightarrow \quad \bar{Y} = A + B\bar{x} \rightarrow \quad \boxed{A = \bar{Y} - B\bar{x}}$$

$$\sum_{i=1}^{n} x_i Y_i = A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 \quad \rightarrow \quad \sum_{i=1}^{n} x_i Y_i = (\bar{Y} - B\bar{x})n\bar{x} + B\sum_{i=1}^{n} x_i^2$$

$$\rightarrow \quad B\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \quad \rightarrow \quad \boxed{B = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}$$

Regressione dei dati

Stimatori ai minimi quadrati dei parametri di regressione

Quindi gli stimatori ai minimi quadrati di α e β corrispondenti all'insieme dei dati $x_i, Y_i, i = 1, \ldots n$, sono, rispettivamente,

$$A = \bar{Y} - B\bar{x}$$

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

La linea retta $A + Bx$ è chiamata **linea di regressione stimata.**

Regressione dei dati

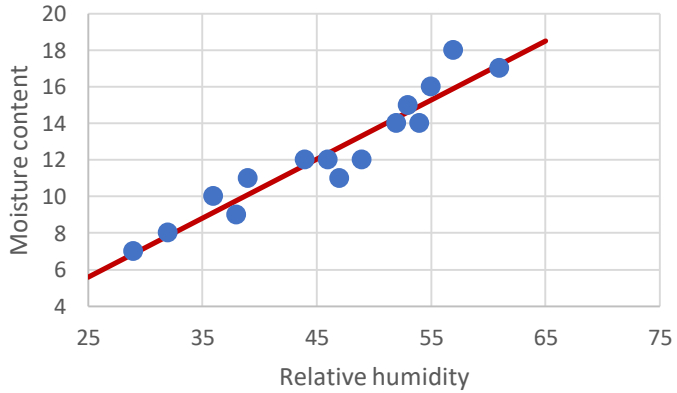
Stimatori ai minimi quadrati dei parametri di regressione

ESEMPIO: La materia prima utilizzata nella produzione di una certa fibra sintetica è conservata in un luogo senza controllo dell'umidità. Le misurazioni dell'umidità relativa nel luogo di stoccaggio e il contenuto di umidità di un campione della materia prima sono stati presi per 15 giorni con i seguenti dati (in percentuale) risultanti.

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

$$A = \bar{Y} - B\bar{x} = 12.4 - 46.1B = -2.510$$

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{15}(x_i - 46.1)Y_i}{\sum_{i=1}^{15} x_i^2 - 15\ 46.1^2} = \frac{\sum_{i=1}^{15}(x_i - 46.1)Y_i}{\sum_{i=1}^{15} x_i^2 - 15\ \cdot 46.1^2} = 0.323$$



Regressione dei dati

Distribuzione degli stimatori

Solitamente si assume che gli errori casuali siano variabili casuali normali indipendenti, con media 0 e varianza σ^2. Cioè, supponiamo che se Y_i è la risposta corrispondente al valore di ingresso x_i, allora $Y_1, \dots Y_n$, sono indipendenti e

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

Regressione dei dati

Distribuzione degli stimatori

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

è una combinazione lineare delle variabili casuali normali indipendenti $Y_i, i = 1, \dots, n$, e quindi è essa stessa normalmente distribuita. La media e la varianza di B sono calcolate come segue:

$$E[B] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E[Y_i]}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\alpha \sum_{i=1}^{n}(x_i - \bar{x}) + \beta \sum_{i=1}^{n} x_i(x_i - \bar{x})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Essendo $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

$$E[B] = \beta \frac{\sum_{i=1}^{n} x_i(x_i - \bar{x})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \beta \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \beta$$

Quindi B è uno stimatore non distorto (unbiased estimator) di β.

Regressione dei dati

Distribuzione degli stimatori

$$Var[B] = \frac{Var[\sum_{i=1}^{n}(x_i - \bar{x})Y_i]}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var[Y_i]}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2} = \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2}$$

$$= \sigma^2 \frac{\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2} = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2} = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2} =$$

$$= \sigma^2 \frac{1}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Regressione dei dati

Distribuzione degli stimatori

Calcoliamo media e varianza dello stimatore $A = \bar{Y} - B\bar{x} = \frac{1}{n}\sum_{i=1}^{n} Y_i - B\bar{x}$

$$E[A] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] - E[B]\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \beta\bar{x} = \frac{n\alpha}{n} + \beta\frac{1}{n}\sum_{i=1}^{n} x_i - \beta\bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha$$

Quindi A è uno stimatore non distorto (unbiased estimator) di α.

$$Var[A] = Var\left[\frac{1}{n}\sum_{i=1}^{n} Y_i - B\bar{x}\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[Y_i] + \bar{x}^2 Var[B] == \frac{\sigma^2}{n} + \sigma^2\frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right) = \sigma^2\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 - n\bar{x}^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)} = \sigma^2\frac{\sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

Regressione dei dati

Distribuzione degli stimatori

Le quantità $Y_i - A - Bx_i, i = 1, \dots, n$, che rappresentano le differenze tra le risposte effettive (cioè le Y_i) e i loro stimatori ai minimi quadrati (cioè $A + Bx_i$) sono chiamati **residui**.

La **somma dei quadrati dei residui**

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

può essere utilizzata per stimare la varianza dell'errore sconosciuta σ^2.

Regressione dei dati

Distribuzione degli stimatori

Si può dimostrare che

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

quindi

$$E\left[\frac{SS_R}{\sigma^2}\right] = n - 2 \qquad \text{o} \qquad E\left[\frac{SS_R}{n-2}\right] = \sigma^2$$

Quindi $SS_R/(n-2)$ è uno stimatore non distorto di σ^2.

Inoltre, si può dimostrare che SS_R è indipendente dalla coppia A e B.

Regressione dei dati

Distribuzione degli stimatori

Quando le Y_i sono variabili casuali normali, gli stimatori ai minimi quadrati sono anche gli **stimatori di massima verosimiglianza**. Per verificare questa osservazione, si noti che la densità congiunta di $Y_1, \ldots, Y_n$ è data da

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1(Y_i-\alpha-\beta x_i)^2}{2 \quad \sigma^2}} = \frac{1}{2\pi^{n/2}\sigma^n} e^{-\sum_{i=1}^{n} \frac{1(Y_i-\alpha-\beta x_i)^2}{2 \quad \sigma^2}}$$

Di conseguenza, gli stimatori di massima verosimiglianza di α e β sono precisamente i valori di α e β che minimizzano

$$\sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2$$

cioè, sono gli stimatori ai minimi quadrati.

Regressione dei dati

Distribuzione degli stimatori

Se indichiamo con

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

allora:

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xY}}{S_{xx}} \qquad A = \bar{Y} - B\bar{x} \qquad SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

Regressione dei dati

Distribuzione degli stimatori

Supponiamo che le risposte $Y_i, i = 1, \dots, n$ siano variabili casuali normali indipendenti con media $\alpha + \beta x_i$ e varianza comune σ^2. Gli stimatori dei minimi quadrati di α e β

$$A = \bar{Y} - B\bar{x} \qquad B = \frac{S_{xY}}{S_{xx}}$$

sono distribuiti come segue

$$A \sim N\left(\alpha, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}\right) \qquad B \sim N(\beta, \sigma^2/S_{xx})$$

La somma dei quadrati dei residui $SS_R = \sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2$ è allora

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

e SS_R è indipendente da A e B.

Regressione dei dati

Coefficiente di determinazione e coefficiente di correlazione campionaria

Supponiamo di voler misurare la quantità di variazione nell'insieme dei valori di risposta $Y_1, \dots, Y_n$ corrispondente all'insieme dei valori di ingresso $x_1, \dots, x_n$. Una misura standard in statistica della quantità di variazione in un insieme di valori $Y_1, \dots, Y_n$ è data dalla quantità:

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Per esempio, se tutte le Y_i sono uguali - e quindi sono tutte uguali a $\bar{Y}$ - allora S_{YY} uguale a 0.

La variazione nei valori di Y_i deriva da due fattori. Primo, perché i valori di input x_i sono diversi, le variabili di risposta Y_i hanno tutti valori medi diversi, il che una certa variazione nei loro valori. In secondo luogo, la variazione deriva anche dal fatto che anche quando si tiene conto delle differenze nei valori di input, ciascuna delle variabili di risposta Y_i ha varianza σ^2 e quindi non sarà esattamente uguale al valore previsto al suo ingresso x_i.

Regressione dei dati

Coefficiente di determinazione e coefficiente di correlazione campionaria

$SS_R = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2$ misura la quantità rimanente di variazione nei valori di risposta dopo che i diversi valori di input sono stati presi in considerazione.

$S_{YY} - SS_R$ rappresenta la quantità di variazione nelle variabili di risposta che è spiegata dai diversi valori di input.

Il **coefficiente di determinazione**, definito come

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}}$$

rappresenta la proporzione della variazione nelle variabili di risposta che è spiegata dai diversi valori di input.

Regressione dei dati

Coefficiente di determinazione e coefficiente di correlazione campionaria

R^2 è spesso usato come indicatore di quanto bene il modello di regressione si adatti ai dati.

$0 \leq R^2 \leq 1$

Un valore di R^2 vicino a 1 indica che la maggior parte della variazione dei dati di risposta è spiegata dai diversi valori di input

- buon adattamento del modello ai dati

Un valore di R^2 vicino a 0 indica che poco della variazione è spiegata dai diversi valori di input

- cattivo adattamento del modello ai dati

Regressione dei dati

Coefficiente di determinazione e coefficiente di correlazione campionaria

Il coefficiente di correlazione campionaria r dell'insieme di coppie di dati $(x_i, Y_i), i = 1, \dots, n$, è definito da

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}}$$

Essendo $SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$ vediamo che

$$r^2 = \frac{S_{xY}^2}{S_{xx}S_{YY}} = \frac{S_{xx}S_{YY} - S_{xx}SS_R}{S_{xx}S_{YY}} = 1 - \frac{SS_R}{S_{YY}} = R^2 \qquad \rightarrow \qquad |r| = \sqrt{R^2}$$

e quindi, a parte il suo segno che indica se è positivo o negativo, il coefficiente di correlazione campionaria è uguale alla radice quadrata del coefficiente di determinazione.

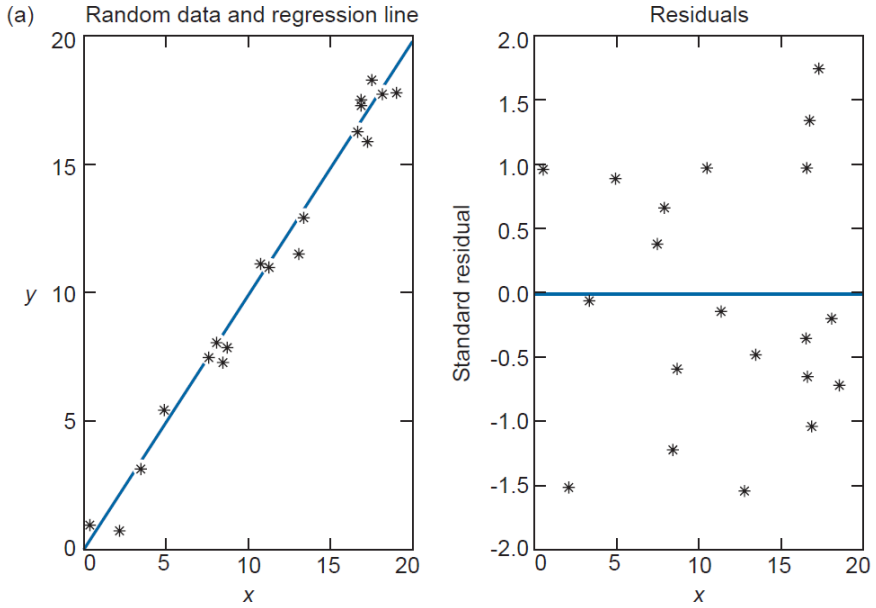
Regressione dei dati

Analisi dei residui: valutazione del modello

residui standardizzati

$$\frac{Y_i - (A + Bx_i)}{\sqrt{SS_R/(n-2)}}$$

Il modello, sia dal diagramma di dispersione che dalla natura casuale dei suoi residui standardizzati, sembra adattarsi abbastanza bene al modello a linea retta.



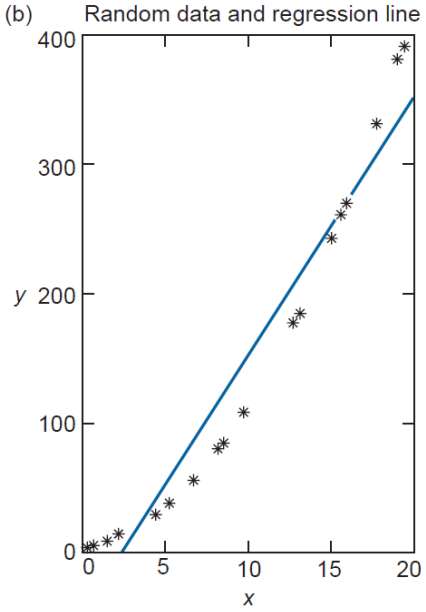
Regressione dei dati

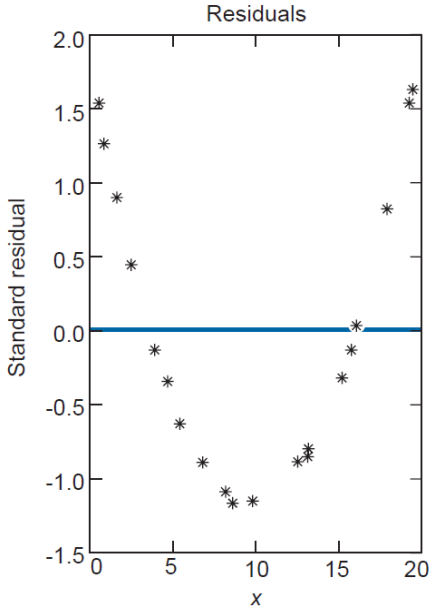
Analisi dei residui: valutazione del modello

residui standardizzati

$$\frac{Y_i - (A + Bx_i)}{\sqrt{SS_R/(n-2)}}$$

Il diagramma dei residui mostra che i residui prima diminuiscono e poi aumentano all'aumentare del livello di ingresso. Questo spesso significa che sono necessari termini di ordine superiore (non solo lineari) per descrivere la relazione tra l'input e la risposta. Si vede anche dal diagramma di dispersione in questo caso.





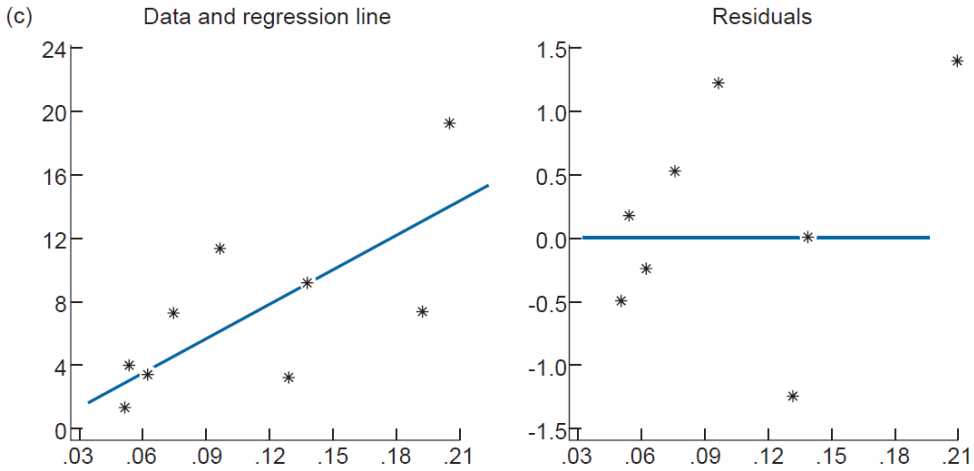
Regressione dei dati

Analisi dei residui: valutazione del modello

residui standardizzati

$$\frac{Y_i - (A + Bx_i)}{\sqrt{SS_R/(n-2)}}$$

Il diagramma dei residui standardizzati mostra che il valore assoluto dei residui, e quindi i loro quadrati, aumentano all'aumentare del livello di input. Questo spesso indica che la varianza della risposta non è costante ma, piuttosto, aumenta con il livello di input.

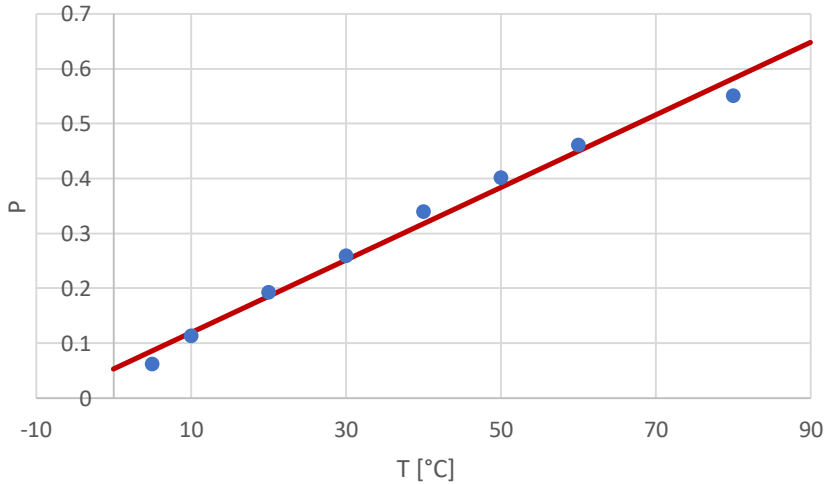


Regressione dei dati

Linearizzazione

ESEMPIO: La seguente tabella fornisce le percentuali di una sostanza chimica che sono state consumate quando un esperimento è stato eseguito a varie temperature (in gradi celsius). Usala per stimare la percentuale della sostanza chimica che verrebbe consumata se l'esperimento venisse eseguito a 350 gradi.

Temperature	Percentage
5°	0.061
10°	0.113
20°	0.192
30°	0.259
40°	0.339
50°	0.401
60°	0.461
80°	0.551



Regressione dei dati

Linearizzazione

ESEMPIO:
Sia $P(x)$ la percentuale della sostanza chimica che si consuma quando l'esperimento viene eseguito a $10x$ gradi. Anche se un grafico di $P(x)$ sembra approssimativamente lineare, possiamo migliorare l'adattamento considerando una relazione non lineare tra x e $P(x)$.

In particolare, consideriamo una relazione della forma

$$1 - P(x) \approx c(1 - d)^x$$

Cioè, supponiamo che la percentuale della sostanza chimica che sopravvive ad un esperimento condotto alla temperatura x diminuisce approssimativamente ad un tasso esponenziale quando x aumenta. Prendendo i logaritmi, il precedente può essere scritto come

$$\ln\big(1 - P(x)\big) \approx \ln(c) + x\ln(1 - d)$$

Così, impostando

$$Y = -\ln\big(1 - P(x)\big)$$
$$\alpha = -\ln c$$
$$\beta = -\ln(1 - d)$$

otteniamo

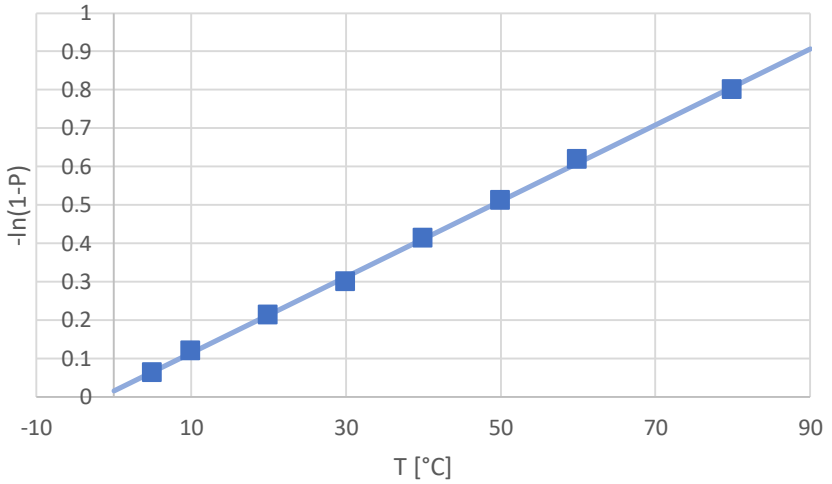
$$Y = \alpha + \beta x + e$$

Regressione dei dati

Linearizzazione

ESEMPIO:

Temperature	$-\ln(1-\text{Percentage})$
5°	0.063
10°	0.120
20°	0.213
30°	0.300
40°	0.414
50°	0.512
60°	0.618
80°	0.801



$A = 0.0154$

$B = 0.0099$

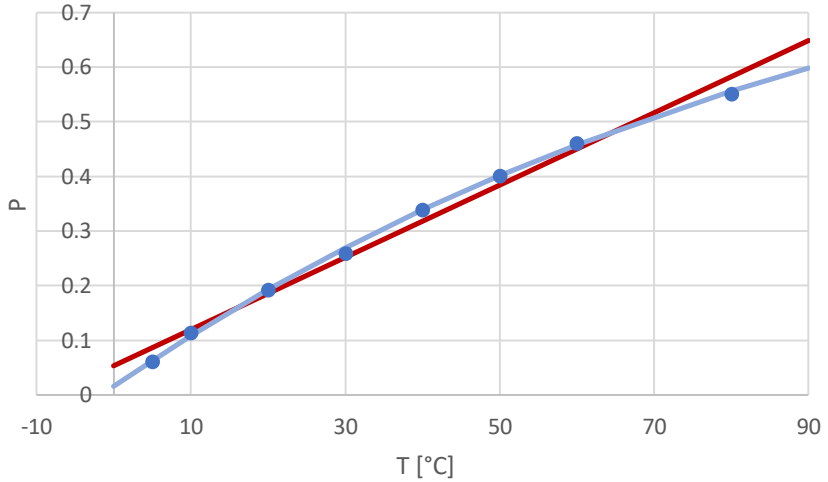
Regressione dei dati

Linearizzazione

ESEMPIO: Nella variabile originale

$$\hat{c} = e^{-A} = 0.9847$$
$$1 - \hat{d} = e^{-B} = 0.9901$$
$$\hat{P} = 1 - 0.9847(0.9901)^{x}$$

T	P	$\hat{P}$	$P - \hat{P}$
5	0.061	0.063	-0.002
10	0.113	0.109	0.004
20	0.192	0.193	-0.001
30	0.259	0.269	-0.010
40	0.339	0.339	0.000
50	0.401	0.401	0.000
60	0.461	0.458	0.003
80	0.551	0.556	-0.005



Regressione dei dati

Regressione polinomiale

In situazioni in cui la relazione funzionale tra la risposta Y e la variabile indipendente variabile x non può essere adeguatamente approssimata da una relazione lineare, potremmo provare ad adattare alla serie di dati una relazione funzionale della forma

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + e$$

dove $\beta_0, \beta_1, \ldots, \beta_r$ sono coefficienti di regressione che devono essere stimati.

Se l'insieme dei dati è costituito dalle n coppie $(x_i, Y_i), i = 1, \ldots, n$, allora gli stimatori ai minimi quadrati di $\beta_0, \beta_1, \ldots, \beta_r$ - chiamiamoli $B_0, B_1, \ldots, B_r$ - sono quei valori che minimizzano

$$SS_R = \sum_{i=1}^{n} \left(Y_i - B_0 - B_1 x_i - B_2 x_i^2 - \cdots - B_r x_i^r\right)^2$$

Regressione dei dati

Regressione polinomiale

Ponendo le derivate parziali di SS_R rispetto a $B_0, B_1, \dots, B_r$, uguale a 0, gli stimatori ai minimi quadrati $B_0, B_1, \dots, B_r$ soddisfano il seguente insieme di $r+1$ equazioni lineari chiamate **equazioni normali:**

$$\sum_{i=1}^{n} Y_i = B_0 n + B_1 \sum_{i=1}^{n} x_i + B_2 \sum_{i=1}^{n} x_i^2 + \cdots + B_r \sum_{i=1}^{n} x_i^r$$

$$\sum_{i=1}^{n} x_i Y_i = B_0 \sum_{i=1}^{n} x_i + B_1 \sum_{i=1}^{n} x_i^2 + B_2 \sum_{i=1}^{n} x_i^3 + \cdots + B_r \sum_{i=1}^{n} x_i^{r+1}$$

$$\sum_{i=1}^{n} x_i^2 Y_i = B_0 \sum_{i=1}^{n} x_i^2 + B_1 \sum_{i=1}^{n} x_i^3 + B_2 \sum_{i=1}^{n} x_i^4 + \cdots + B_r \sum_{i=1}^{n} x_i^{r+2}$$

$$\vdots \quad = \quad \vdots \quad + \quad \vdots \quad + \quad \vdots \quad + \cdots + \quad \vdots$$

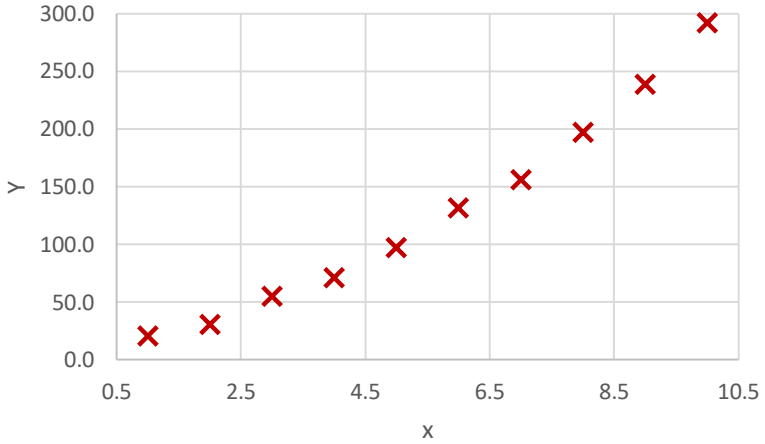
$$\sum_{i=1}^{n} x_i^r Y_i = B_0 \sum_{i=1}^{n} x_i^r + B_1 \sum_{i=1}^{n} x_i^{r+1} + B_2 \sum_{i=1}^{n} x_i^{r+2} + \cdots + B_r \sum_{i=1}^{n} x_i^{2r}$$

Regressione dei dati

Regressione polinomiale

ESEMPIO: Adatta un polinomio ai seguenti dati:

x	Y
1	20.6
2	30.8
3	55.0
4	71.4
5	97.3
6	131.8
7	156.3
8	197.3
9	238.7
10	291.7



Regressione dei dati

Regressione polinomiale

ESEMPIO:

Un grafico di questi dati indica che una relazione quadratica

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

è sufficiente. Siccome:

$$\sum_{i=1}^{n} x_i = 55 \qquad \sum_{i=1}^{n} x_i^2 = 385 \qquad \sum_{i=1}^{n} x_i^3 = 3025 \qquad \sum_{i=1}^{n} x_i^4 = 25333$$

$$\sum_{i=1}^{n} Y_i = 1291.1 \qquad \sum_{i=1}^{n} x_i Y_i = 9549.3 \qquad \sum_{i=1}^{n} x_i^2 Y_i = 77758.9$$

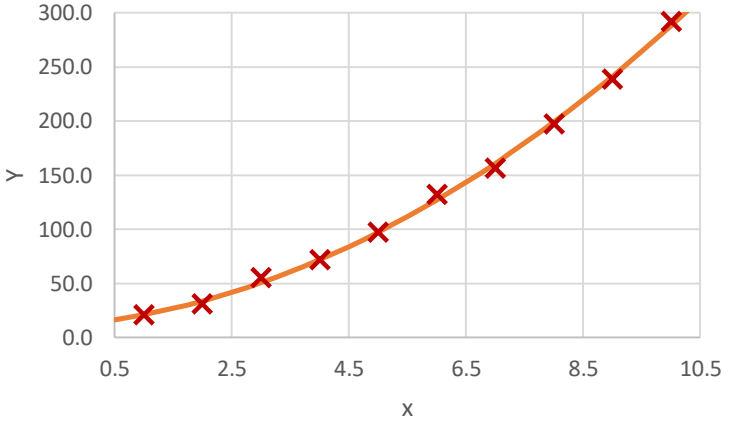
gli stimatori ai minimi quadrati sono le soluzioni del seguente sistema di equazioni:

$$1291.1 = 10B_0 + 55B_1 + 385B_2$$
$$9549.3 = 55B_0 + 385B_1 + 3025B_2$$
$$77758.9 = 385B_0 + 3025B_1 + 25333B_2$$

Da cui

B_0 =12.59326 B_1 =6.326172 B_2 =2.122818

$$Y = 12.59 + 6.33x + 2.12x^2$$



Regressione dei dati

Regressione lineare multipla

Nella maggior parte delle applicazioni, una situazione tipica è quella in cui c'è un insieme di k variabili di input e la risposta Y è legata ad esse dalla relazione

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$$

dove $x_j, j = 1, \ldots, k$, è il livello della j-esima variabile di ingresso ed e è un errore casuale che assumiamo sia normalmente distribuito con media 0 e varianza (costante) σ^2. Cioè, le Y_i sono collegate ai livelli di ingresso attraverso la:

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

Se $B_0, B_1, \ldots, B_k$ denotano gli stimatori di $\beta_0, \beta_1, \ldots, \beta_k$, allora

$$SS_R = \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2$$

Regressione dei dati

Regressione lineare multipla

Per determinare gli stimatori dei minimi quadrati, poniamo le derivate parziali di SS_R rispetto a $B_0, B_1, \dots, B_k$ uguale a 0. Otteniamo $k+1$ equazioni

$$\sum_{i=1}^{n}(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i1}(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik}) = 0$$

$$\sum_{i=1}^{n} x_{i2}(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik}) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik}(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik}) = 0$$

Regressione dei dati

Regressione lineare multipla

Le equazioni normali sono:

$$\sum_{i=1}^{n} Y_i = nB_0 + B_1 \sum_{i=1}^{n} x_{i1} + B_2 \sum_{i=1}^{n} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik}$$

$$\sum_{i=1}^{n} x_{i1} Y_i = B_0 \sum_{i=1}^{n} x_{i1} + B_1 \sum_{i=1}^{n} x_{i1}^2 + B_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{i1} x_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik} Y_i = B_0 \sum_{i=1}^{n} x_{ik} + B_1 \sum_{i=1}^{n} x_{ik} x_{i1} + B_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + B_k \sum_{i=1}^{n} x_{ik}^2$$

Regressione dei dati

Regressione lineare multipla

In notazione matriciale:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 x_{11} x_{12} \cdots x_{1k} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ 1 x_{n1} x_{n2} \cdots x_{nk} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

con $\mathbf{Y}$ matrice $n \times 1$, $\mathbf{X}$ matrice $n \times p$, $\boldsymbol{\beta}$ matrice $p \times 1$, ed $\mathbf{e}$ matrice $n \times 1$ dove $p \equiv k + 1$.

Il modello di regressione multipla è

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Regressione dei dati

Regressione lineare multipla

Sia

$$\mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix}$$

la matrice degli stimatori ai minimi quadrati, allora le equazioni normali possono essere scritte come:

$$\mathbf{X}^T\mathbf{X}\mathbf{B} = \mathbf{X}^T\mathbf{Y}$$

Da cui

$$\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Regressione dei dati

Regressione di processi gaussiani (GPR)

I modelli GPR sono una classe di modelli machine learning non parametrici comunemente utilizzati per modellare dati spaziali e serie temporali.

I modelli GPR hanno diversi vantaggi:

- ❑ funzionano bene su piccoli insiemi di dati
- ❑ forniscono misure di incertezza sulle previsioni.

I modelli GPR consentono una regressione non lineare dei dati.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Il modello GPR non è parametrico (cioè non è limitato da una forma funzionale).
Invece che calcolare la distribuzione di probabilità dei parametri di una funzione specifica, il modello GPR calcola la distribuzione di probabilità su tutte le funzioni ammissibili che si adattano ai dati.

Specifichiamo una prior (sullo spazio delle funzioni), calcoliamo la posterior usando i dati di allenamento, e calcoliamo la distribuzione predittiva a posteriori sui punti di interesse.

Ci sono diverse librerie che implementano efficientemente il modello (ad esempio scikit-learn, Gpytorch, GPy).

Regressione dei dati

Regressione di processi gaussiani (GPR)

Definizioni

Una **distribuzione Gaussiana multivariata** è definita da un vettore medio $\boldsymbol{\mu}$ e una matrice di covarianza $\boldsymbol{\Sigma}$. La media $\boldsymbol{\mu}$ si riferisce alla media di ciascuna variabile casuale X_i e la matrice di covarianza $\boldsymbol{\Sigma}$ fornisce la covarianza tra ciascuna delle variabili casuali X_i e X_j :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1j}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2j}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \sigma_{i1}^2 & \sigma_{i2}^2 & \cdots & \sigma_{ij}^2 & \cdots & \sigma_{in}^2 \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nj}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix}$$

Intuitivamente, la matrice di covarianza generalizza la varianza a più dimensioni e la diagonale è costituita dalla varianza di ciascuna variabile casuale X_i.

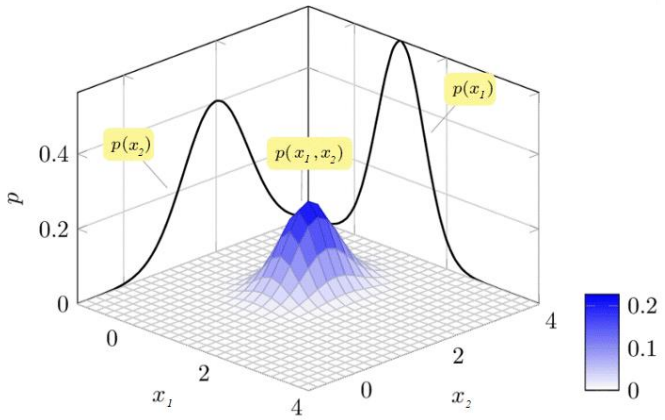
Regressione dei dati

Regressione di processi gaussiani (GPR)

Definizioni

Sia $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ una coppia di variabili aleatorie con distribuzione di probabilità congiunta $p(x_1, x_2)$ gaussiana con media $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ e convarianza $\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$.

$$p(x_1, x_2) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$



Esempio di una distribuzione gaussiana bivariata.

Regressione di processi gaussiani (GPR)

Definizioni

Si possono derivare le seguenti utili formule in forma chiusa per le probabilità marginali e condizionali:

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_2 = N(\mu_1, \Sigma_{11})$$

$$p(x_2) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_1 = N(\mu_2, \Sigma_{22})$$

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} = N\left(\mu_1 + \Sigma_{12}{\Sigma_{22}}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{21}\right)$$

Regressione dei dati

Regressione di processi gaussiani (GPR)

Definizioni

I processi aleatori sono un'estensione del concetto di variabile aleatoria.

Si definisce processo stocastico una famiglia di variabili aleatorie $X(\mathbf{s}, \omega)$, $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d$, definite su uno spazio campione Ω con $\omega \in \Omega$, e che assumono valori in un insieme definito *spazio degli stati del processo*. Un processo stocastico è quindi un insieme di funzioni che evolvono nello spazio $\mathcal{D}$ (le cosiddette *funzioni campione* o *realizzazioni*), ognuna delle quali è associata ad un determinato elemento dello spazio campione, così che il risultato di un esperimento casuale corrisponde di fatto all'estrazione di una di queste funzioni.

Il processo stocastico è una funzione stocastica specificata dalle sue distribuzioni congiunte dimensionalmente finite:

$$F(y_1, y_2, \dots y_n; \mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n) = P(X(\mathbf{s}_1) \leq y_1, \dots, X(\mathbf{s}_n) \leq y_n)$$

per ogni n finito e ogni insieme di punti $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$ in $\mathcal{D}$.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Definizioni

In teoria delle probabilità un **processo gaussiano** è un processo stocastico $X(\mathbf{s})$ tale che prendendo un qualsiasi numero finito di variabili aleatorie, dalla collezione che forma il processo aleatorio stesso, esse hanno una distribuzione di probabilità congiunta gaussiana.

Un processo gaussiano è specificato interamente dalla sua media $\mu(\mathbf{s})$ e dalla covarianza $k(\mathbf{s}, \mathbf{s}') = Cov\big(X(\mathbf{s}); X(\mathbf{s}')\big)$ e viene indicato nel modo seguente:

$$X(\mathbf{s}) \sim N\big(\mu(\mathbf{s}), k(\mathbf{s}, \mathbf{s}')\big)$$

Ha la proprietà che per ogni insieme finito di punti $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$

$$\mathbf{x} \equiv \big(X(\mathbf{s}_1), X(\mathbf{s}_2), \dots, X(\mathbf{s}_n)\big)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

con $\Sigma_{ij} = k\big(\mathbf{s}_i, \mathbf{s}_j\big)$.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Definizioni

Per l'esistenza di un **processo gaussiano** con media e covarianza prescritte è sufficiente assicurare che k sia definita positiva. In questo caso la distribuzione ha densità:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Una funzione k è definita positiva se per ogni insieme finito di punti $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ in $\mathcal{D}$ la matrice di covarianza

$$\boldsymbol{\Sigma} = \begin{bmatrix} k(\mathbf{s}_1,\mathbf{s}_1) & k(\mathbf{s}_1,\mathbf{s}_2) & \cdots & k(\mathbf{s}_1,\mathbf{s}_n) \\ k(\mathbf{s}_2,\mathbf{s}_1) & k(\mathbf{s}_2,\mathbf{s}_2) & \cdots & k(\mathbf{s}_2,\mathbf{s}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{s}_n,\mathbf{s}_1) & k(\mathbf{s}_n,\mathbf{s}_2) & \cdots & k(\mathbf{s}_n,\mathbf{s}_n) \end{bmatrix}$$

è semi-definita positiva: $\mathbf{z}^T\boldsymbol{\Sigma}\mathbf{z} \geq 0$ per ogni vettore a valori reali $\mathbf{z}$.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Introduciamo l'approccio bayesiano:

Con una regressione lineare standard, assumiamo che gli output della funzione possano essere calcolati come una combinazione lineare degli input: $y = wx + \epsilon$.
L'approccio bayesiano specifica una distribuzione a priori, $p(w)$, sul parametro, w, e ridefinisce le probabilità in base all'evidenza (cioè i dati osservati (y, x)) usando la formula di Bayes:

$$p(w|y,x) = \frac{p(y|x,w)p(w)}{p(y|w)} \qquad posterior = \frac{likelihood \times prior}{marginal\ likelihood}$$

La distribuzione $p(w|y,X)$, chiamata distribuzione a posteriori, incorpora quindi informazioni sia dalla distribuzione a priori che dal dataset.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Introduciamo l'approccio bayesiano:

Per ottenere previsioni in punti di interesse non osservati, x^*, la distribuzione predittiva può essere calcolata ponderando tutte le possibili previsioni per la loro distribuzione a posteriori calcolata come:

$$p(f^*|x^*, y, x) = \int_w p(f^*|x^*, w)p(w|y, x)dw$$

La prior e la verosimiglianza sono di solito assunte come gaussiane per rendere l'integrazione trattabile.

Usando questa assunzione e risolvendo la distribuzione predittiva, otteniamo una distribuzione gaussiana, dalla quale possiamo ottenere una predizione puntuale usando la sua media e una quantificazione dell'incertezza usando la sua varianza.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Nel modello GPR, la risposta $z(\mathbf{s})$ è considerata una realizzazione del processo Gaussiano multivariato $Z(\mathbf{s})$:

$$Z(\mathbf{s}) = m(\mathbf{s}) + Y(\mathbf{s})$$

con $m(\mathbf{s}) = \mathbf{h}(\mathbf{s})\boldsymbol{\beta}$ è una funzione di regression deterministica, costruita dai dati osservati, e $Y(\mathbf{s})$ è un processo Gaussiano, costruito sui residui, con media nulla e funzione di covarianza $k(\mathbf{s}_i, \mathbf{s}_j) = \sigma^2 r(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})$.

σ^2 è un fattore di scala, chiamato VARIANZA del PROCESSO

$r(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})$ è una funzione positiva con parametric $\boldsymbol{\theta}$, chiamata funzione di CORRELAZIONE

$\boldsymbol{\theta}$ sono detti IPER- PARAMETRI

Regressione dei dati

Regressione di processi gaussiani (GPR)

Se $m(\mathbf{s}) = 0$, il modello GPR è detto SIMPLE KRIGING.

Se $m(\mathbf{s}) = \beta$ con β costante, il modello GPR è detto ORDINARY KRIGING.

Se $m(\mathbf{s}) = \mathbf{h}(\mathbf{s})\boldsymbol{\beta}$ con $\mathbf{h}(\mathbf{s}) = \left(h_1(\mathbf{s}), \dots h_p(\mathbf{s})\right)^T$ p funzioni scelte e $\boldsymbol{\beta}$ vettore di p coefficienti incogniti, il modello GPR è detto UNIVERSAL KRIGING.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Indichiamo con $\mathbf{z}^{(n)}$ i valori osservati di $z(\mathbf{s})$ in n punti noti $\widehat{\mathcal{D}} = (\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n)^T \subset \mathcal{D}$.

In molti casi, non abbiamo accesso diretto alla funzione da approssimare ma solo a una sua versione rumorosa. Consideriamo questo caso più generale, assumendo un rumore di osservazione gaussiano indipendente con media zero e varianza $\sigma_\epsilon^2(\mathbf{s})$. Questo è solitamente indicato come nugget effect. Quindi, $\mathbf{z}^{(n)}$ sono realizzazioni del vettore gaussiano $\boldsymbol{Z}^{(n)} = Z(\widehat{\mathcal{D}}) + \mathbf{E}^{(n)}$, dove $Z(\widehat{\mathcal{D}})$ è il processo gaussiano $Z(\mathbf{s})$ nei punti $\widehat{\mathcal{D}}$ e $\mathbf{E}^{(n)} = (\sigma_\epsilon(\mathbf{s}_1)E_1, \dots, \sigma_\epsilon(\mathbf{s}_n)E_n)^T$ è il rumore bianco con $E_{i=1,\dots,n}$ indipendenti e identicamente distribuite rispetto a una distribuzione gaussiana con media zero e varianza uno.

Per semplicità assumiamo un modello ordinary kriging, dove $m(\mathbf{s}) = h(\mathbf{s})\beta$ con $h(\mathbf{s}) = 1$ e $p = 1$, e σ_ϵ costante.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Usiamo le informazioni contenute in $\boldsymbol{Z}^{(n)}$ per prevedere $Z(\mathbf{s})$ considerando la distribuzione congiunta di $Z(\mathbf{s})$ e $\boldsymbol{Z}^{(n)}$:

$$\begin{pmatrix} Z(\mathbf{s}) \\ \boldsymbol{Z}^{(n)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{h}(\mathbf{s})\boldsymbol{\beta} \\ \mathbf{H}\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} k(\boldsymbol{s},\boldsymbol{s}) & \mathbf{k}^T(\boldsymbol{s}) \\ \mathbf{k}(\boldsymbol{s}) & \boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I} \end{pmatrix} \right)$$

dove

$\mathbf{H} = \mathbf{h}(\widehat{\mathcal{D}})$ matrice del modello $n \times p$

$\boldsymbol{\Sigma}$ matrice di covarianza $n \times n$ tra i punti osservati $\widehat{\mathcal{D}}$

$\mathbf{k}(\boldsymbol{s})$ vettore di covarianza di dimensione n tra i punti da predire $\boldsymbol{s}$ e i punti osservati $\widehat{\mathcal{D}}$

Regressione dei dati

Regressione di processi gaussiani (GPR)

La distribuzione condizionale

$$p\left(Z(\mathbf{s})|\boldsymbol{Z}^{(n)};\boldsymbol{\beta},\sigma^2,\sigma_\epsilon^2,\boldsymbol{\theta}\right)=\frac{p\left(Z(\mathbf{s}),\boldsymbol{Z}^{(n)};\boldsymbol{\beta},\sigma^2,\sigma_\epsilon^2,\boldsymbol{\theta}\right)}{p(\boldsymbol{Z}^{(n)};\boldsymbol{\beta},\sigma^2,\sigma_\epsilon^2,\boldsymbol{\theta})}$$

è gaussiana con media e varianza:

$$\widehat{m}_Z(\mathbf{s})=\mathbf{h}(\mathbf{s})\boldsymbol{\beta}+\mathbf{k}^T(\boldsymbol{s})(\boldsymbol{\Sigma}+\sigma_\epsilon^2\mathbf{I})^{-1}\left(\boldsymbol{z}^{(n)}-\mathbf{H}\boldsymbol{\beta}\right)$$

$$\hat{s}_Z^2(\mathbf{s})=k(\boldsymbol{s},\boldsymbol{s})-\mathbf{k}^T(\boldsymbol{s})(\boldsymbol{\Sigma}+\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{k}(\boldsymbol{s})$$

Regressione dei dati

Regressione di processi gaussiani (GPR)

Per stimare $\boldsymbol{\beta}, \sigma^2, \sigma_\epsilon^2, \boldsymbol{\theta}$ un metodo molto popolare è la Stima della Massima Verosimiglianza (Maximum Likelihood Estimastion MLE).
L'assunzione di distribuzione normale multivariata per $\boldsymbol{Z}^{(n)}$ porta alla seguente marginal likelihood:

$$p\left(\boldsymbol{Z}^{(n)}; \boldsymbol{\beta}, \sigma^2, \sigma_\epsilon^2, \boldsymbol{\theta}\right) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I}|^{1/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{z}^{(n)} - \mathbf{H}\boldsymbol{\beta}\right)^T (\boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I})^{-1}\left(\boldsymbol{z}^{(n)} - \mathbf{H}\boldsymbol{\beta}\right)\right)$$

Dato

$$\widehat{\boldsymbol{\beta}} = (\mathbf{H}^T(\boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I})^{-1}\mathbf{H})^{-1}\mathbf{H}^T(\boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I})^{-1}\boldsymbol{z}^{(n)}$$

che è la MLE di $\boldsymbol{\beta}$ corrispondente alla sua stima generalizzata ai minimi quadrati, la stima di $\sigma^2, \sigma_\epsilon^2, \boldsymbol{\theta}$ si ottiene minimizzando

$$\mathcal{L}_{LME}(\sigma^2, \sigma_\epsilon^2, \boldsymbol{\theta}) = \left(\boldsymbol{z}^{(n)} - \mathbf{H}\boldsymbol{\beta}\right)^T (\boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I})^{-1}\left(\boldsymbol{z}^{(n)} - \mathbf{H}\boldsymbol{\beta}\right) + \log(|\boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I}|)$$

che è l'opposto della log-likelihood a meno di una costante.

Regressione dei dati

Regressione di processi gaussiani (GPR)

Come scegliere la $prior\ Z = GP\big(m(\cdot), k(\cdot,\cdot)\big)$:

La scelta del kernel o covarianza è molto importante e specifica come *credi* che *sia* la funzione latente Z.

Il kernel maggiormente utilizzato è il Gaussian Kernel o RBF o exponential quadratic, definito dalla funzione:

$$k(x, x') = \sigma^2 \exp\left(-\frac{1}{2\theta^2}\|x - x'\|^2\right)$$

θ è la scala di lunghezza
σ^2 è la varianza del GP

Regressione dei dati

Regressione di processi gaussiani (GPR)

Generazione dei campioni dato $m(\cdot)$ e $k(\cdot,\cdot)$ (esempio 1d):

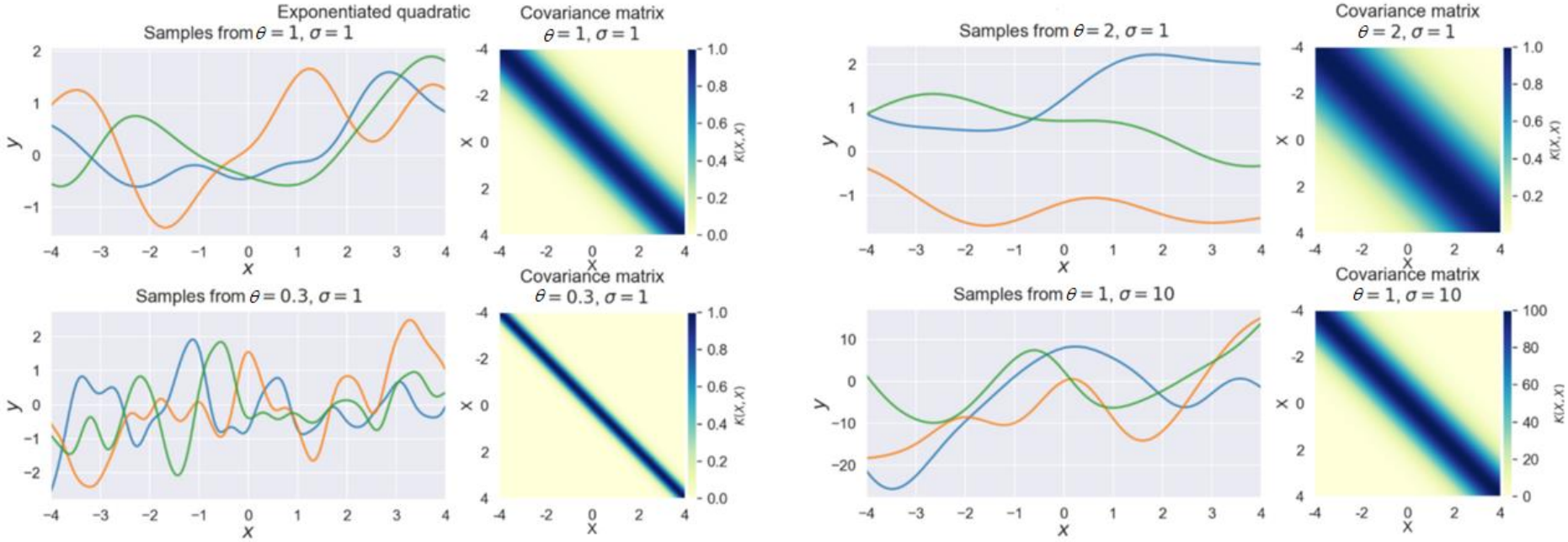
1) Consideriamo l'insieme di n punti $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$
2) Calcoliamo la matrice di covarianza $\boldsymbol{\Sigma}$ nei punti
3) Calcoliamo la decomposizione di Cholesky di $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$
4) Generiamo il campione

$$z(\mathbf{x}) = m(\mathbf{x}) + \mathbf{L}^T \cdot randn(n)$$

dove $randn(n)$ restituisce n scalari tratti da una distribuzione normale standard

Regressione dei dati

Regressione di processi gaussiani (GPR)



Regressione dei dati

Regressione di processi gaussiani (GPR)

Exponential quadratic kernel

$$k(x, x') = \sigma^2 \exp\left(-\frac{1}{2\theta^2}\|x - x'\|^2\right)$$

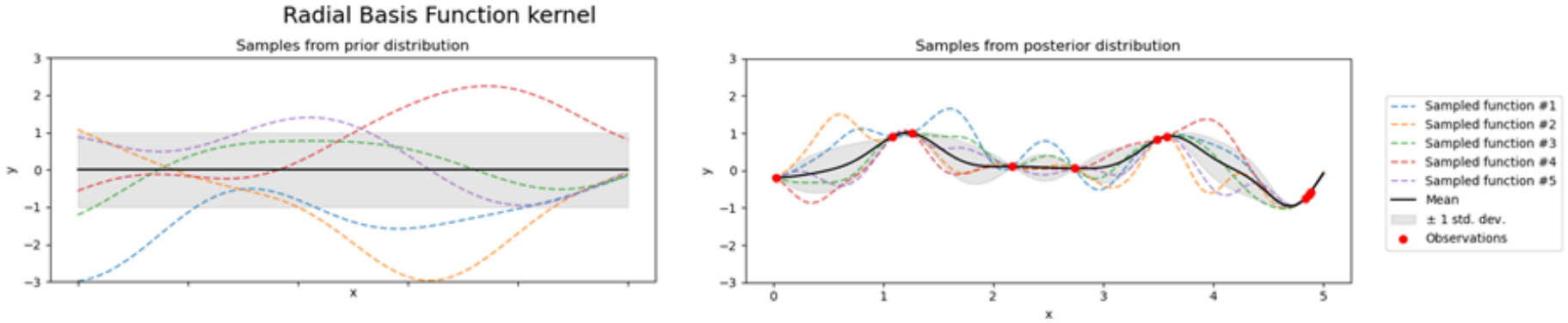
θ è la scala di lunghezza: descrive quanto velocemente la correlazione tra due osservazioni x e x' si riduce quanto più sono lontane.
Un alto θ risulta in una funzione regolare.
Un bassoθ risulta in una funzione irregolare.

Il parametro σ^2 , varianza del GP, controlla l'ampiezza verticale della funzione.

Sono stati sviluppati un gran numero di kernel sia stazionari che instazionari: Matern 3/2, Matern 5/2, exponential, cubic, periodic, Gibbs, neural network...

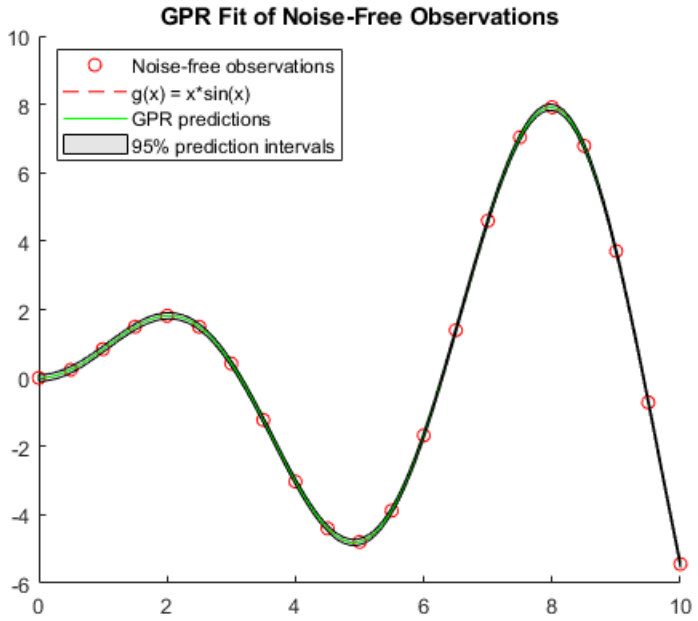
Regressione dei dati

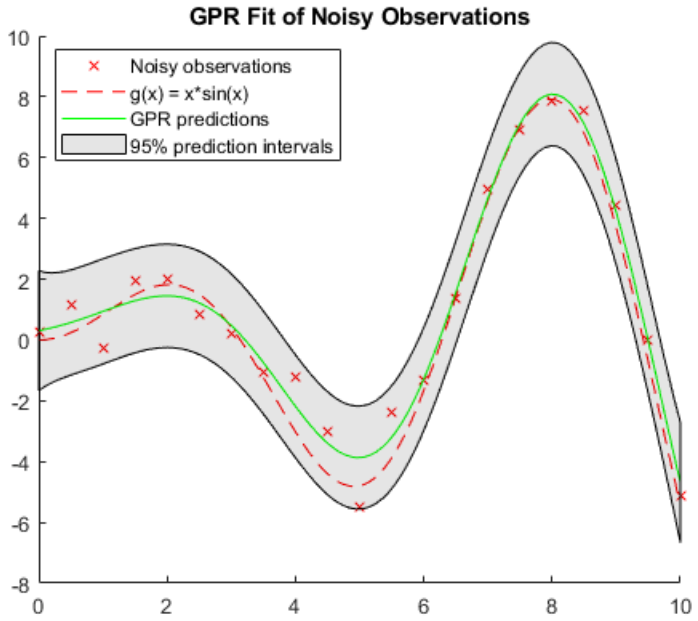
Regressione di processi gaussiani (GPR)



Regressione dei dati

Regressione di processi gaussiani (GPR)





UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Dipartimento di
Ingegneria
e Architettura