Il fit lineare di un insieme di dati: niente di più semplice, giusto?

(Una parte di) Quello che avrei voluto sapere la prima volta che ho incontrato dei dati astronomici

B. Trefoloni

Outline

- **Perchè un fit lineare?**
- **Stimare la correlazione**
- **Il fit dei dati**
- **Stimare la dispersione dei dati rispetto al modello**
- **Incertezze con metodi MC**
- **Aggiungere complessità al modello**

Perchè un fit lineare?

- **Modello in cui i parametri di interesse compaiono con potenza 1**
- **Più semplice modello da testare**
- **Molte forme funzionali possono essere riprodotte con modelli lineari**
- **e.g.**

$s = v*t + s0$ $\qquad$ $C = AT^{3/2} + B$

$L_X = \gamma L_{UV}^{\alpha} \longrightarrow \text{Log } L_X = \alpha \text{ Log } L_{UV} + \gamma$

Trattare dati astrofisici

Modello che vogliamo testare su dei dati per ottenere dei vincoli sui parametri

Vogliamo:

- **Stimare la correlazione dei dati**
- **I parametri di migliore regressione e le loro incertezze**
- **La dispersione della relazione**

Il modello

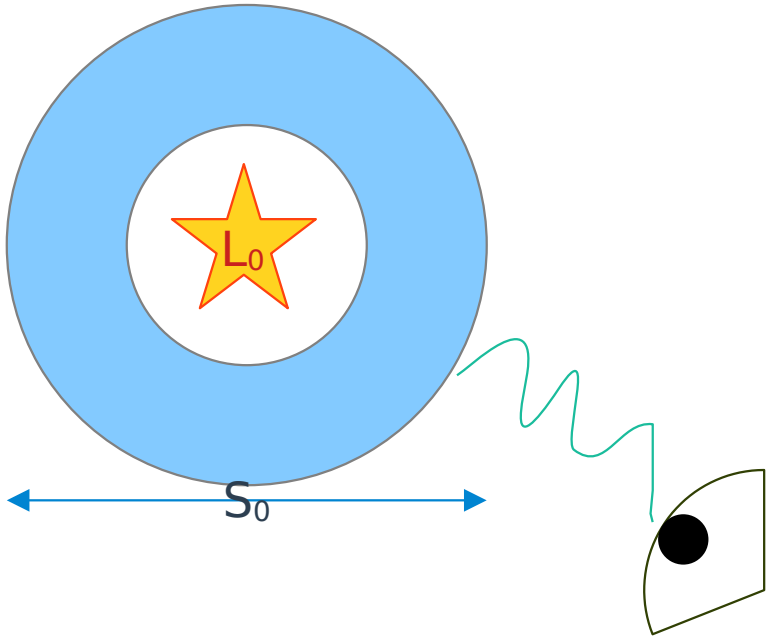
“I modelli sono tutti sbagliati, ma alcuni sono utili”

G. Box

Il modello

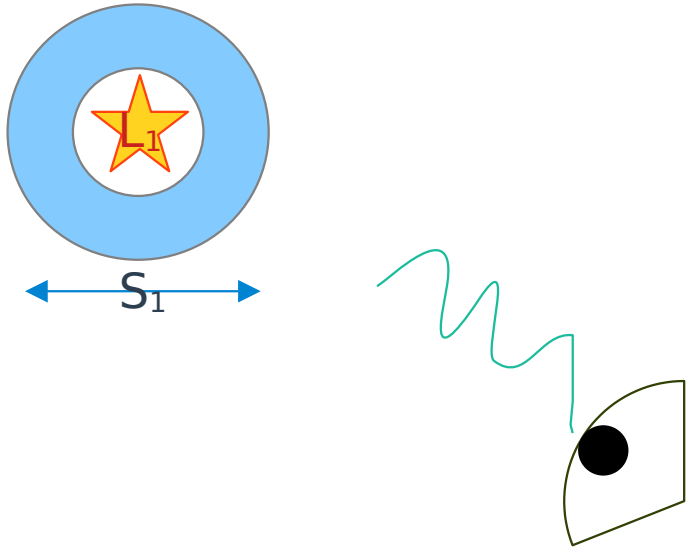
- **Astrazione matematica (equazione) del processo fisico investigato**
- **I modelli non sono utili in senso assoluto, ma in senso comparativo**

Il modello



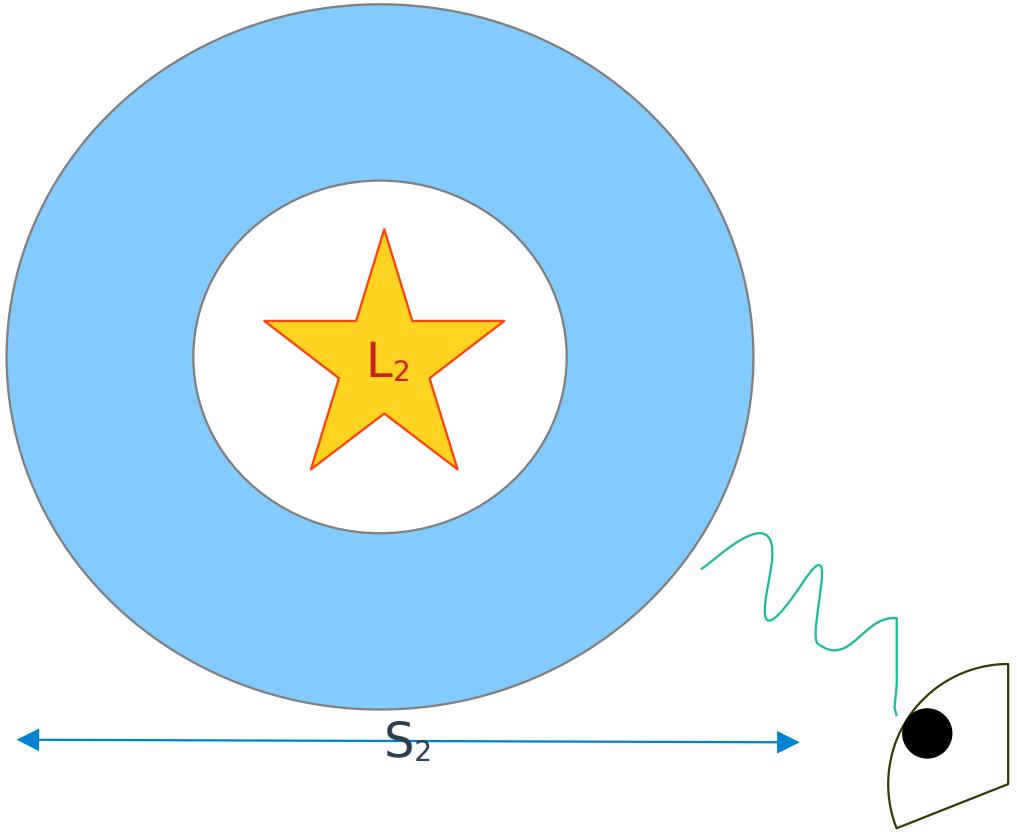
- **Una sorgente ionizzante circondata da gas**
- **Si produce una riga di emissione che viene osservata**

Il modello



- **Una variazione nel continuo ionizzante causa la variazione del flusso della riga di emissione**

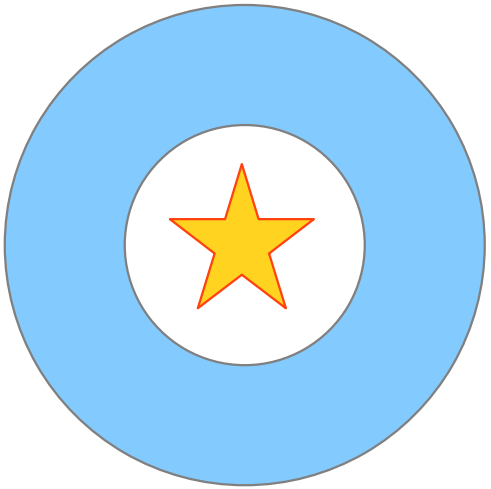
Il modello



$$S_1 < S_0 < S_2$$

$$L_1 < L_0 < L_2$$

Il modello



- **La produzione della riga di emissione richiede un flusso ionizzante minimo**

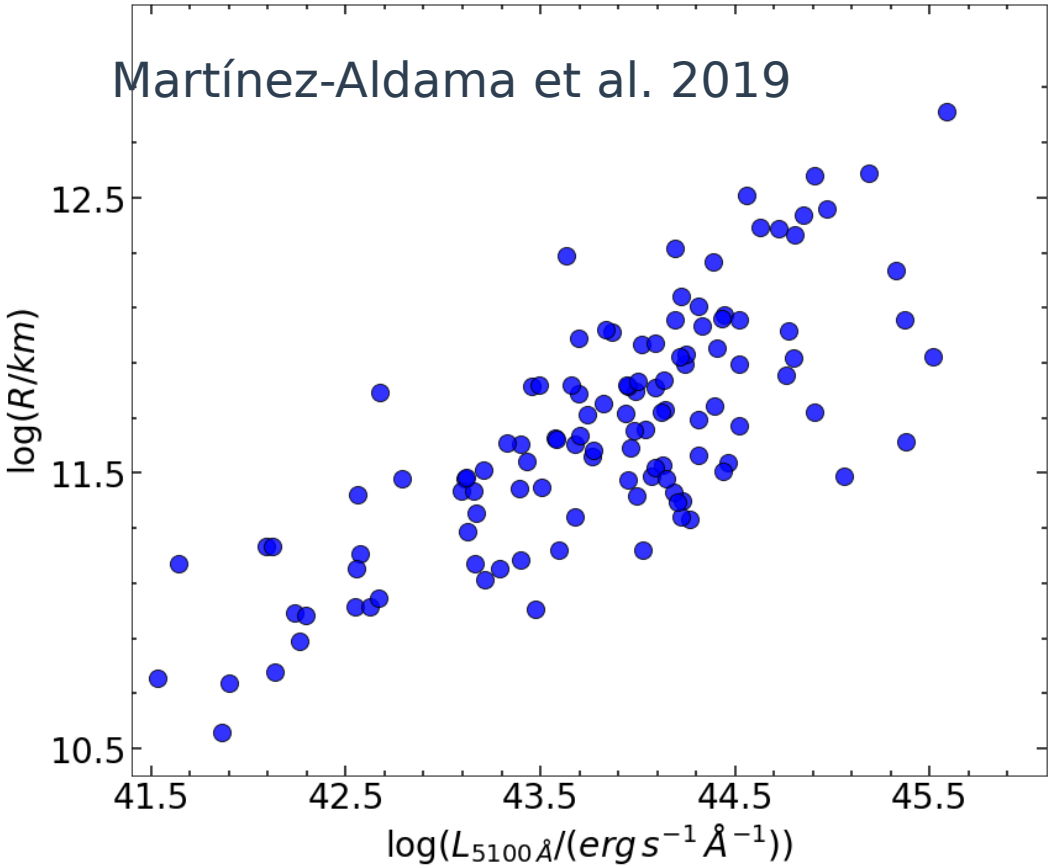
$$F_{ion} = \frac{L}{4\pi R^2} \longleftrightarrow R = \left(\frac{L}{4\pi F_{ion}}\right)^{0.5}$$

- **Quantità misurate R=cΔt e L**

$$R = \alpha\, L^{0.5} \longrightarrow \boxed{\mathrm{Log}(R) = m\,\mathrm{Log}(L) + q}$$

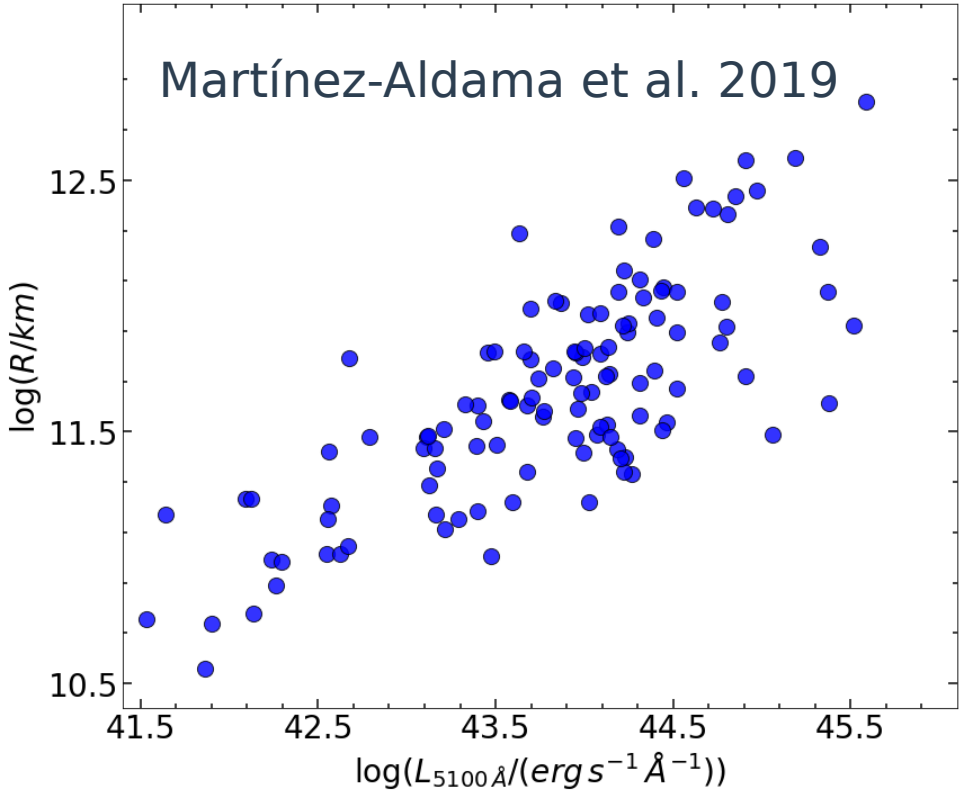
- **m = 0.5?**

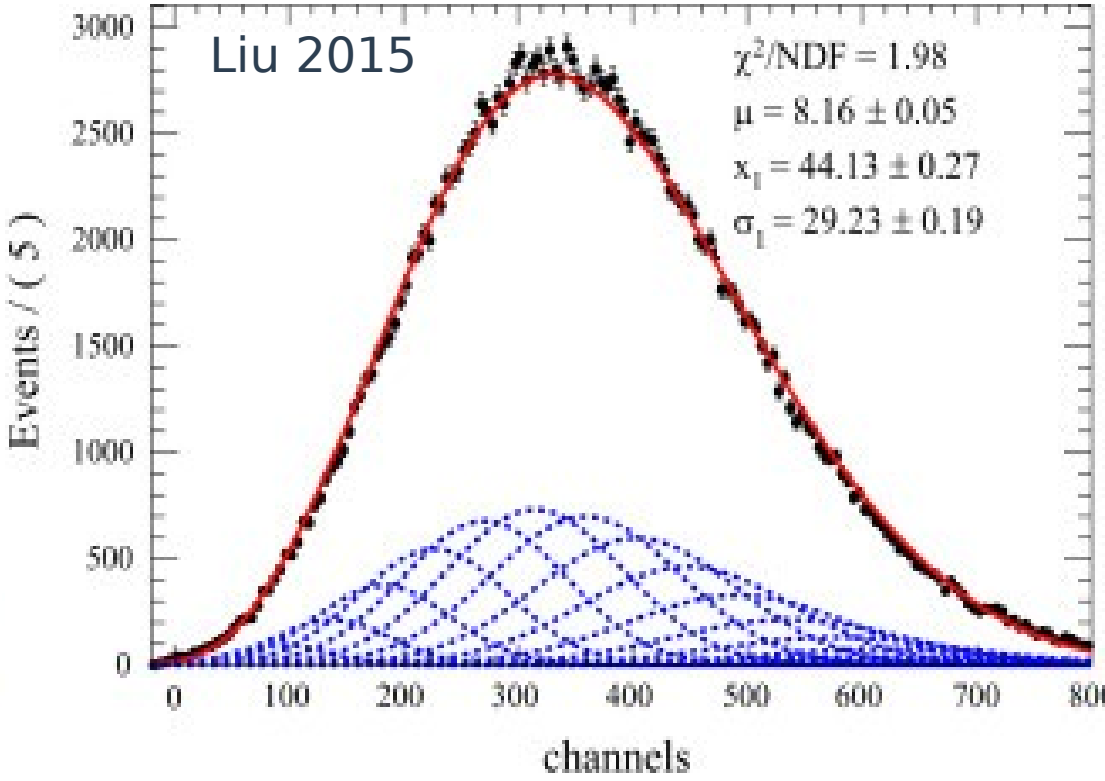
Un primo sguardo ai dati



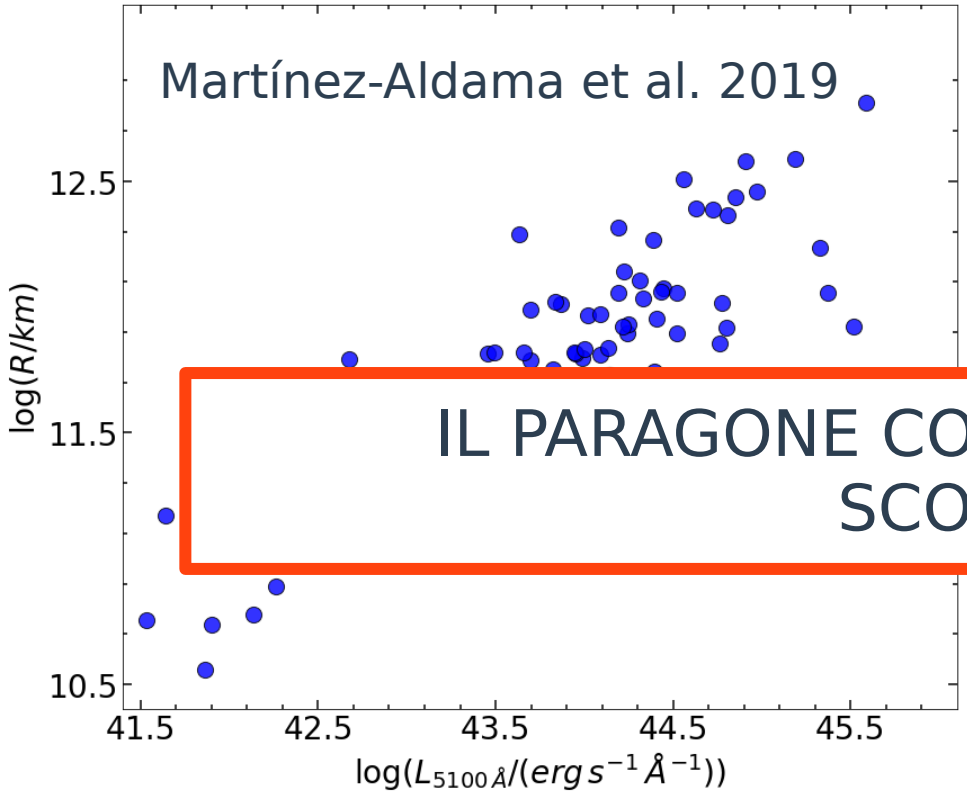
- **Misure astrofisiche generalmente molto più disperse rispetto a misure di laboratorio**
- **Quantità macroscopiche**
- **Incertezze sistematiche importanti**

Un primo sguardo ai dati



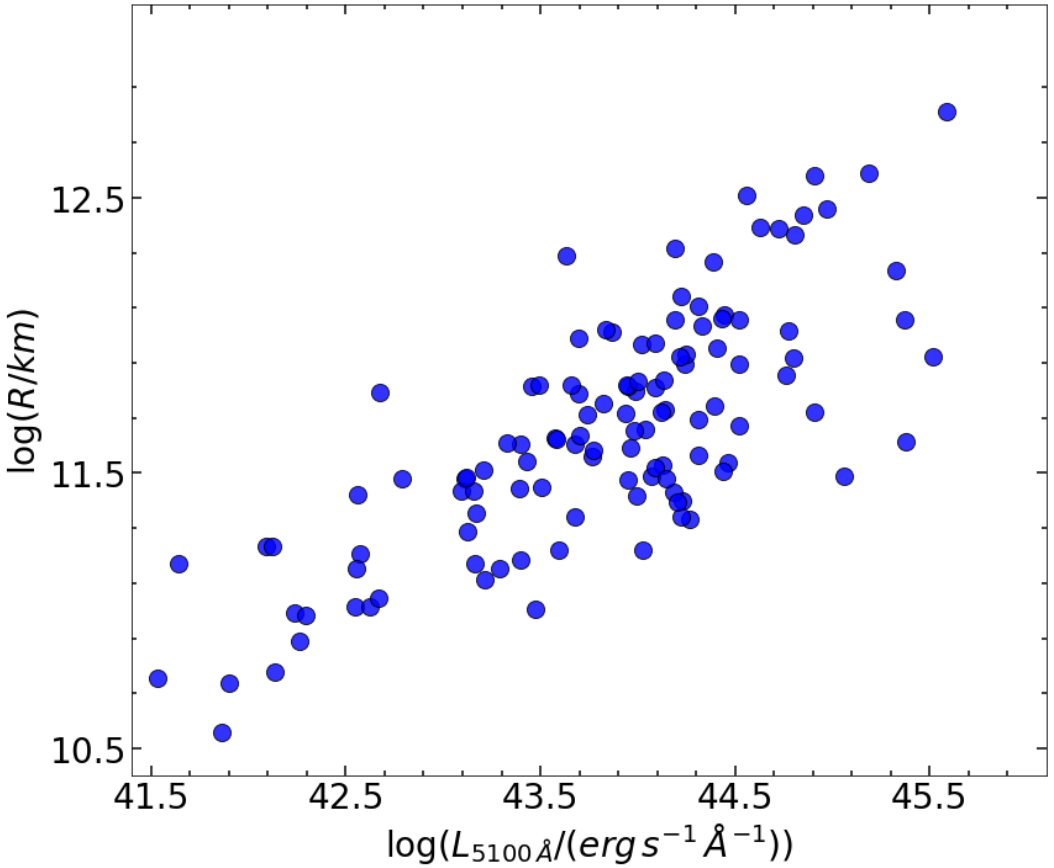


Un primo sguardo ai dati



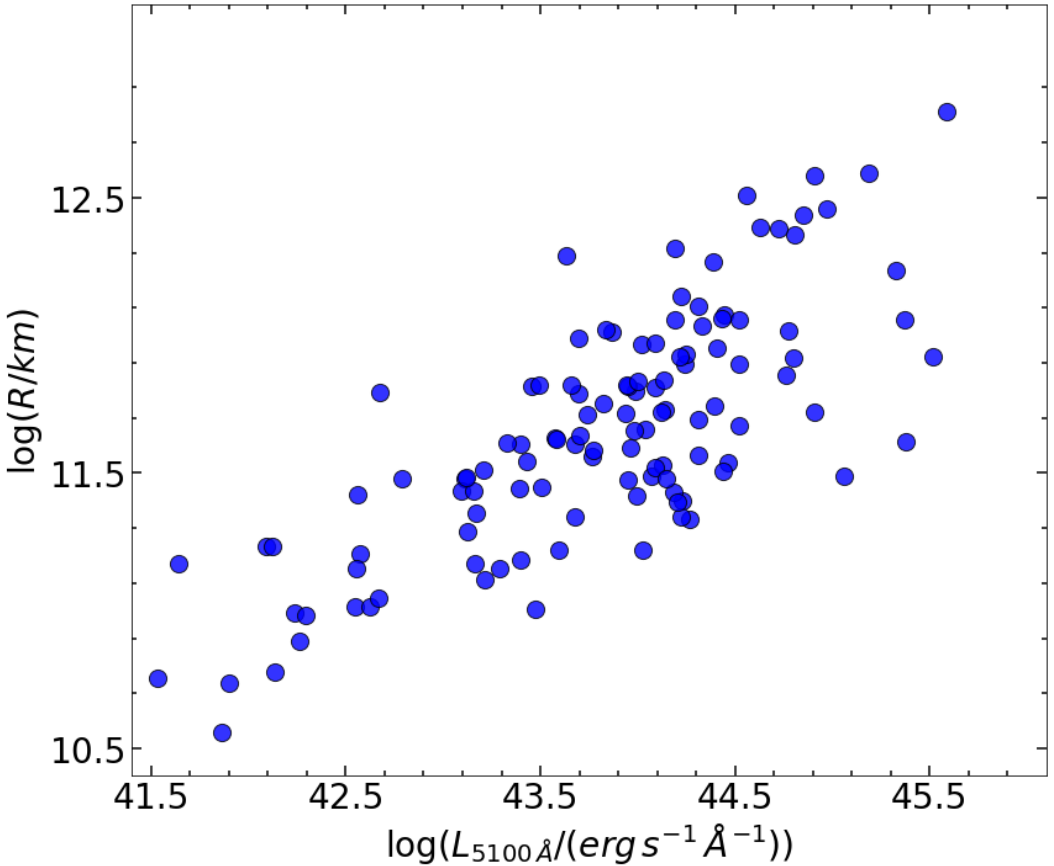
IL PARAGONE CON DATI DI LABORATORIO È SCONFORTANTE...

Stimare la correlazione



- **Le quantità in analisi sono correlate?**
- **Se lo sono, come possiamo stimare il grado di correlazione?**
- **Vari indici possibili (indice di correlazione di Pearson, Spearman rank,...)**

Stimare la correlazione



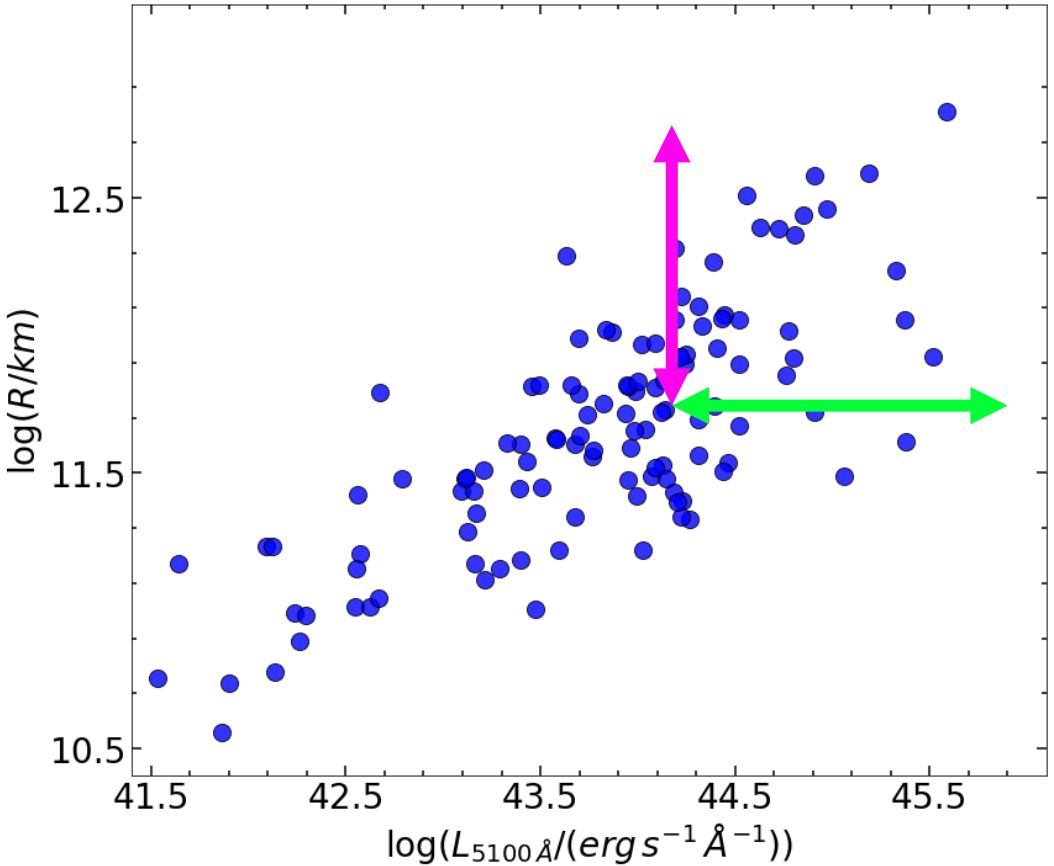
- **Indice di Pearson**

  -1 ≤ P ≤ +1

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

- **P = 0.76**

Stimare la correlazione



- **Indice di Pearson**

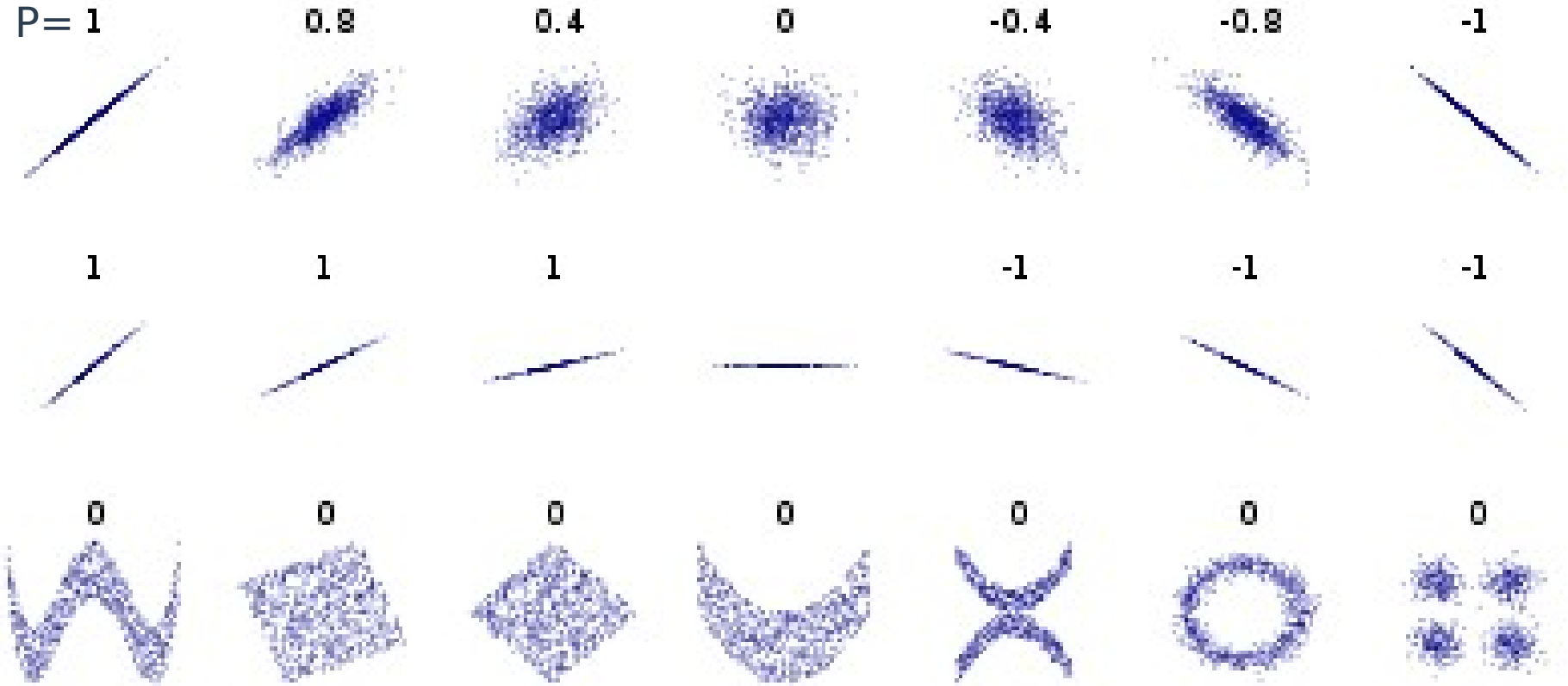
  -1 ≤ P ≤ +1

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

$\sim\sigma_x$ $\sim\sigma_y$

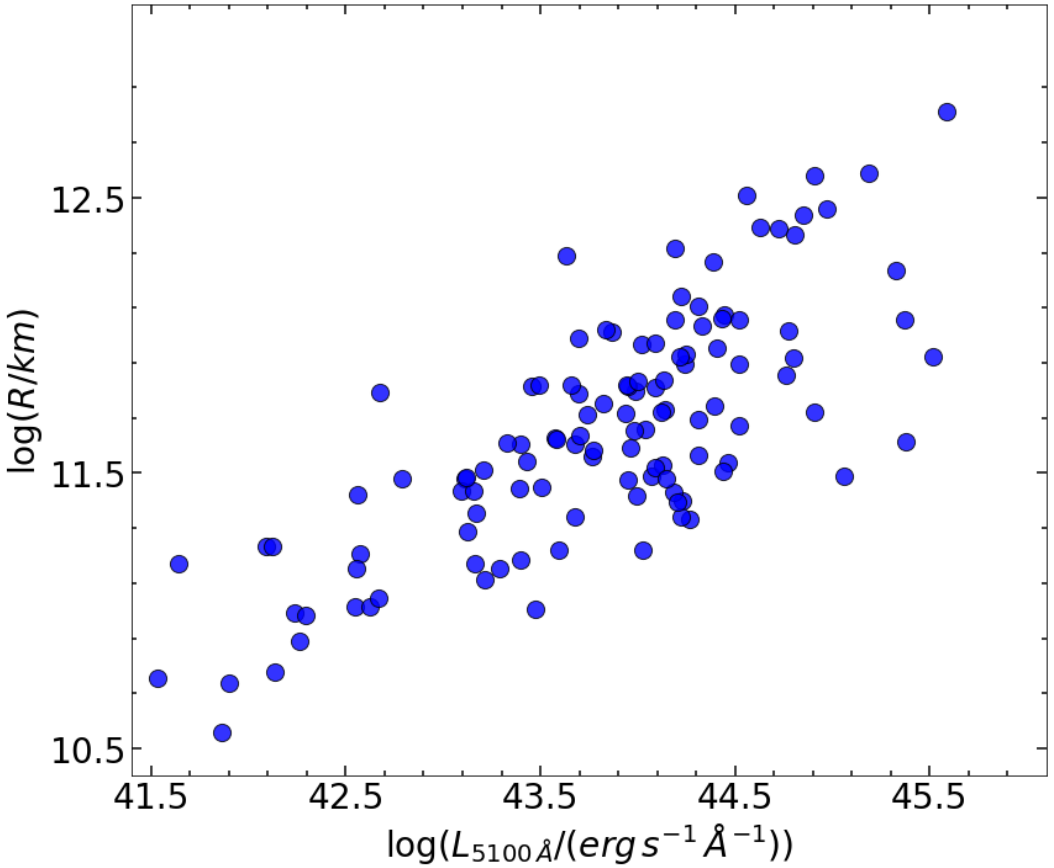
- **P = 0.76**

Stimare la correlazione



wikipedia

Stimare la correlazione



- **Pearson correlation**

  -1 ≤ P ≤ +1

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

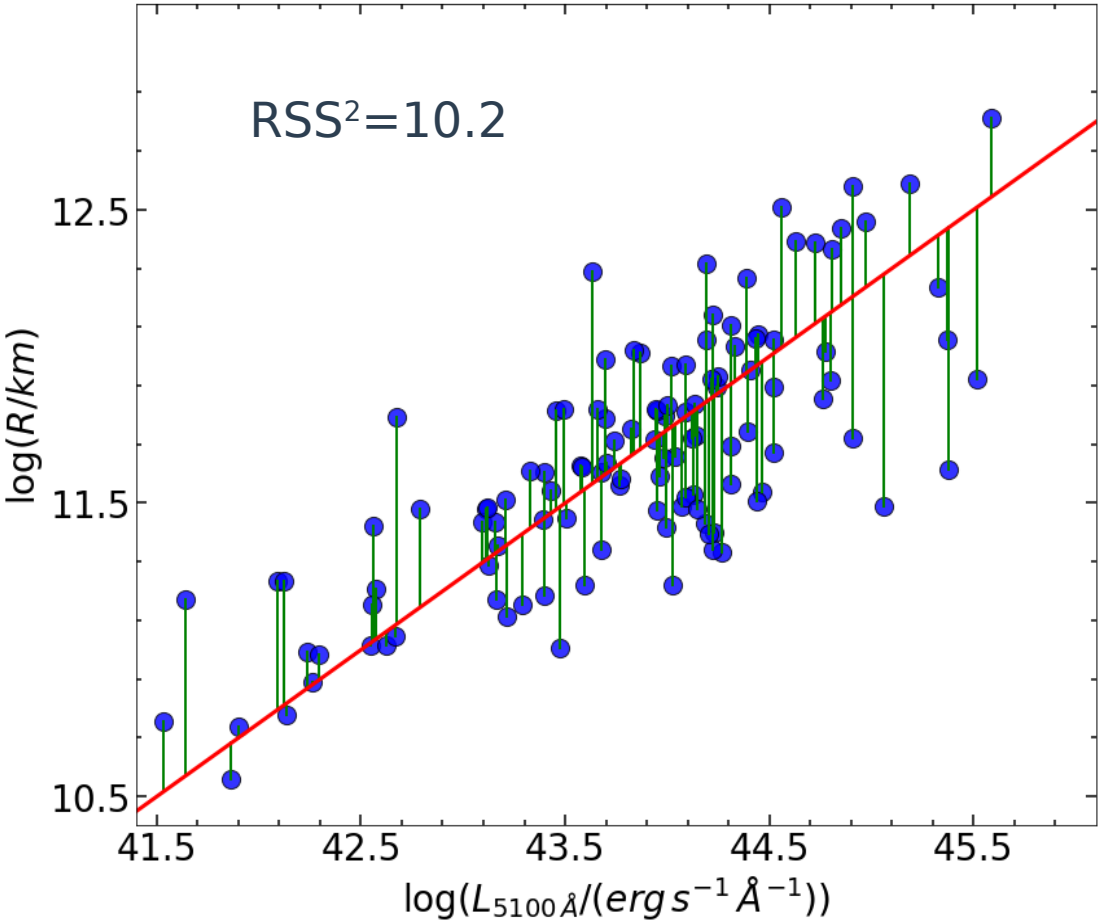
P = 0.76

Fit dei dati

- **Trovare i coefficienti di best fit è un problema di minimo**
- **Definire un indice di “badness of fit” che dipende dai parametri del modello**
- **Minimizzare rispetto ai parametri**
- **Nel contesto dei “Minimi Quadrati” l’indice ha la forma dei residui quadrati**

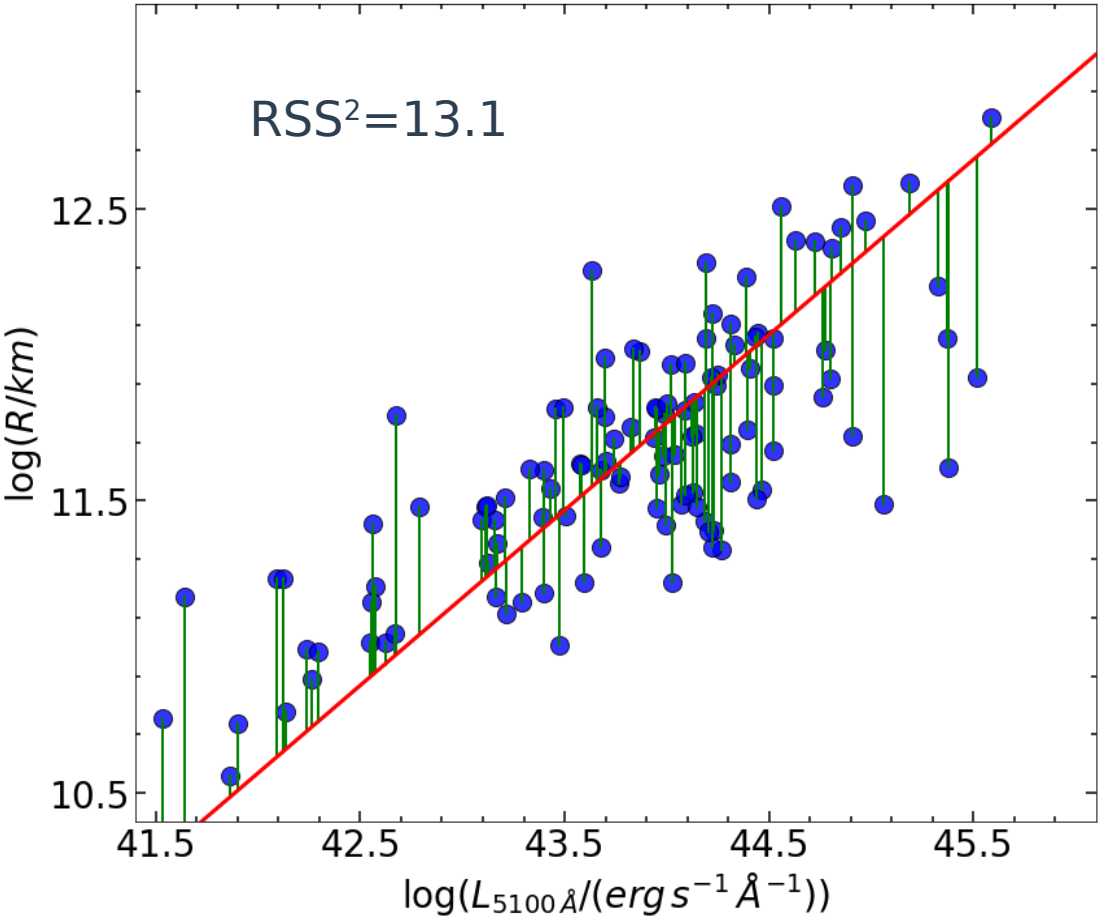
$$RSS^2 = \Sigma_i \, w_i[y_i - (Ax_i + B)]^2 \sim \chi^2$$

Fit dei dati



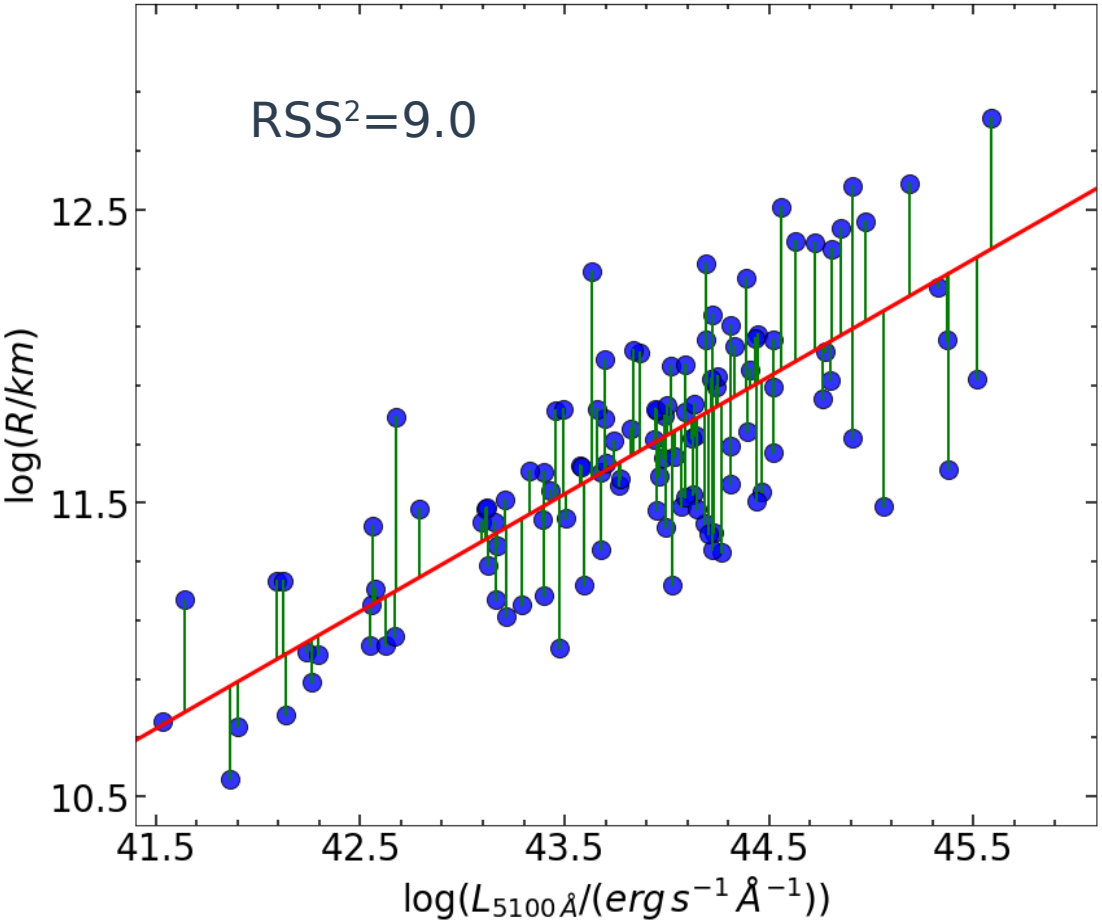
- **Cosa viene minimizzato in realtà?**
- **A,B di best fit (BF) sono i coefficienti che minimizzano la somma in quadratura delle lunghezze dei segmenti verdi**

Fit dei dati



- **Cosa viene minimizzato in realtà?**
- **A,B di best fit (BF) sono i coefficienti che minimizzano la somma in quadratura delle lunghezze dei segmenti verdi**

Fit dei dati



- **Cosa viene minimizzato in realtà?**
- **A,B di best fit (BF) sono i coefficienti che minimizzano la somma in quadratura delle lunghezze dei segmenti verdi**

Fit dei dati

$$\frac{\partial \chi^2}{\partial A} = \frac{-2}{\sigma_y^2} \sum_{i=1}^{N} (y_i - A - Bx_i) = 0$$

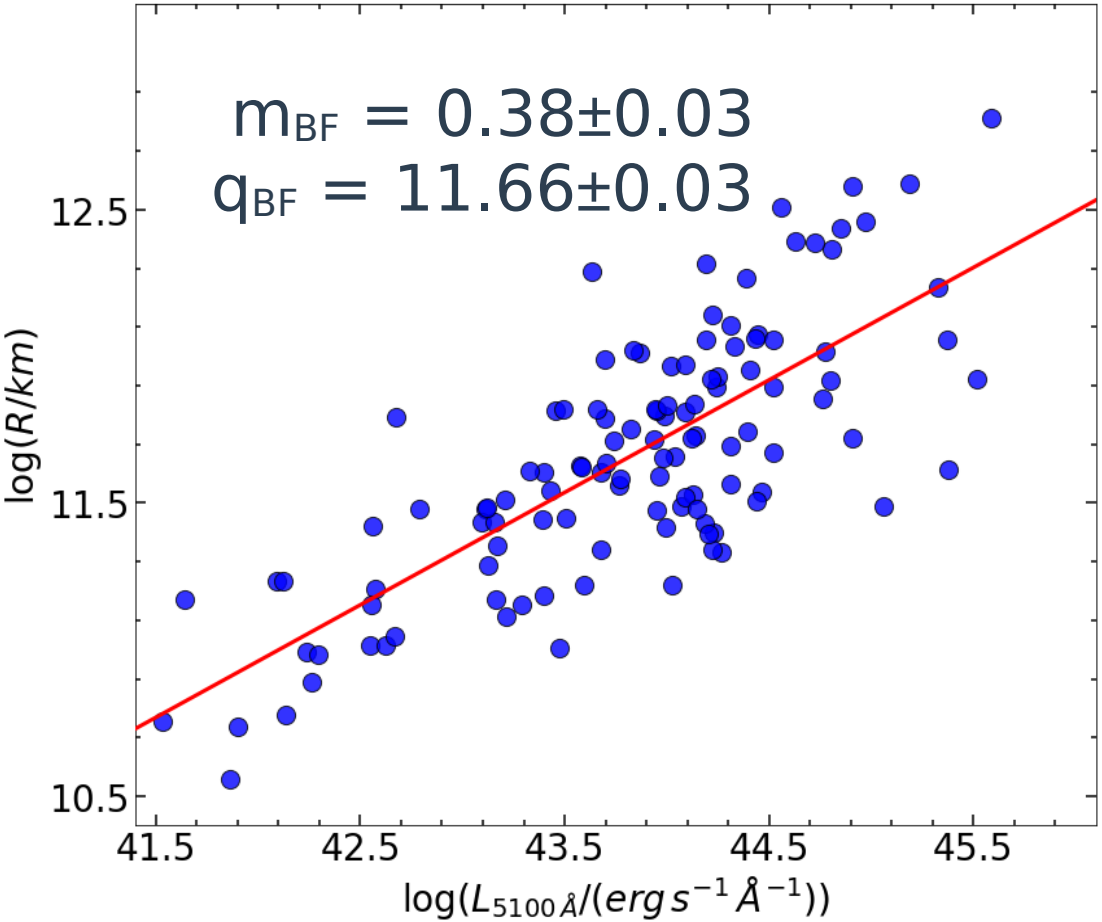
$$\frac{\partial \chi^2}{\partial B} = \frac{-2}{\sigma_y^2} \sum_{i=1}^{N} x_i (y_i - A - Bx_i) = 0$$

$$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{\Delta}$$

$$B = \frac{N \sum xy - \sum x \sum y}{\Delta}$$

- **Calcolare esplicitamente i coefficienti della regressione**
- **Utilizzare qualche programma di analisi dati (e.g. python/scipy) per trovare la retta di best fit (BF)**
- **Per ora non consideriamo le incertezze → w_i = cost**

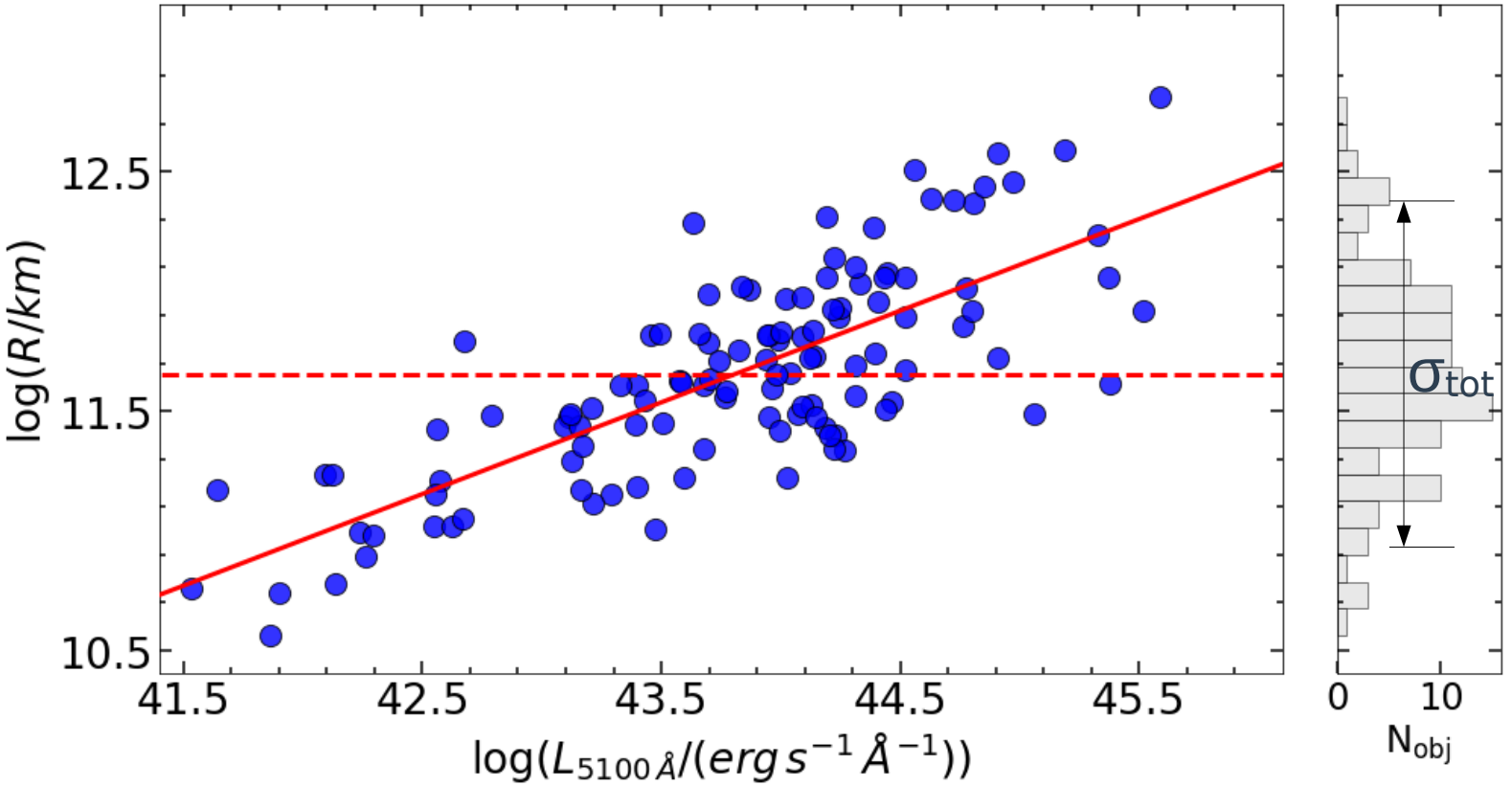
Fit dei dati



- **Calcolare esplicitamente i coefficienti della regressione**
- **Utilizzare qualche programma di analisi dati (e.g. python/scipy) per trovare la retta di best fit (BF)**

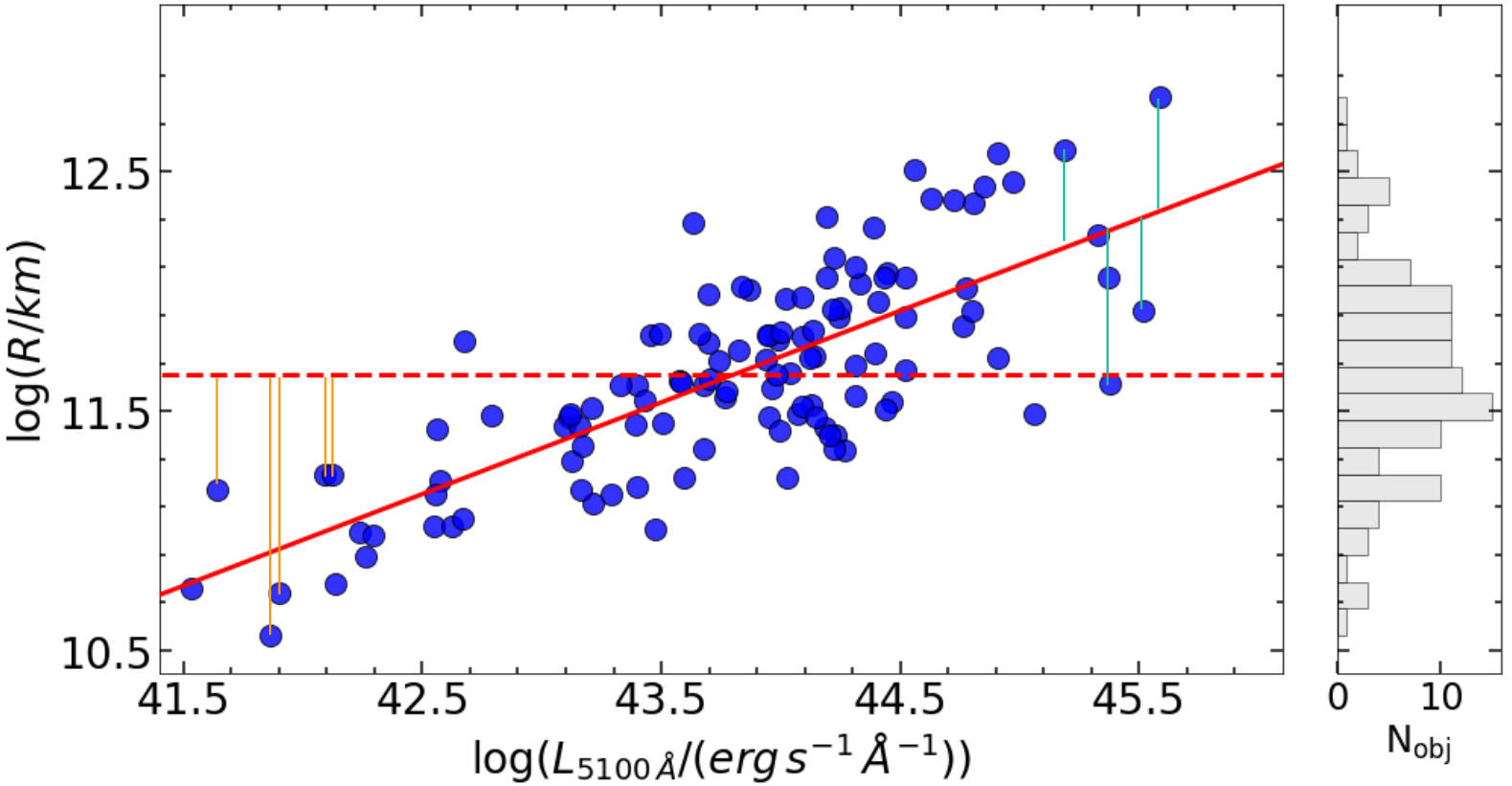
- **Per ora non consideriamo le incertezze → w_i = cost**

Varianza spiegata dal modello



- **Il modello riesce a spiegare la varianza dei dati?**

Explained variance



- **Il modello riesce a spiegare la varianza dei dati?**

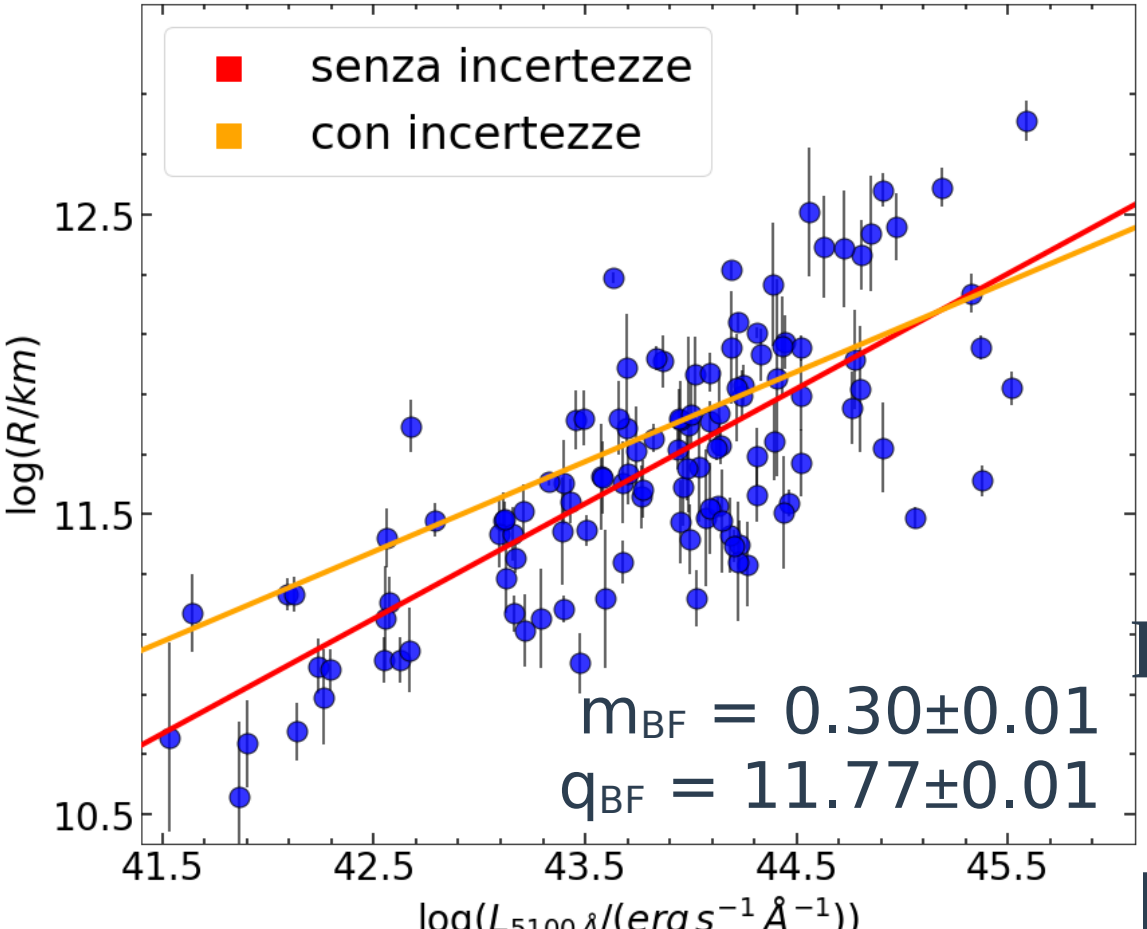
R^2 = 1-RSS/TSS

RSS = $(y_{exp}-y_i)^2$

TSS = $(<y>-y_i)^2$

R^2 = 0.6

Fit dei dati



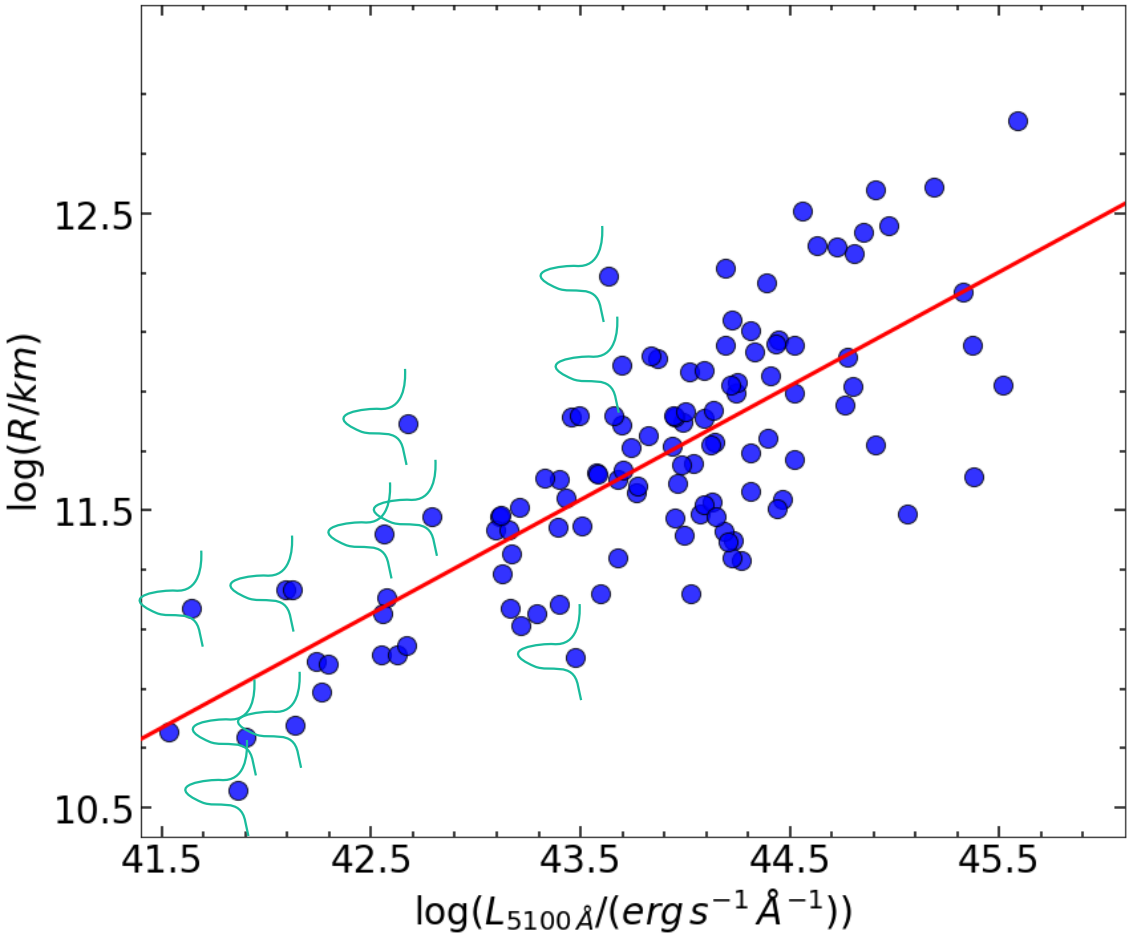
- **Prendendo in considerazione le incertezze sui dati σ_i si correggono i pesi → $w_i = 1/\sigma_i^2$**

$$R^2 = \Sigma_i\, w_i [y_i - (Ax_i + B)]^2$$

Incertezze MC

- **Modo alternativo molto potente per stimare le incertezze sui parametri di BF è il metodo MonteCarlo (MC)**
- **I dati osservati sono una realizzazione di un processo di misura**
- **I punti sono estrazioni da distribuzioni gaussiane**

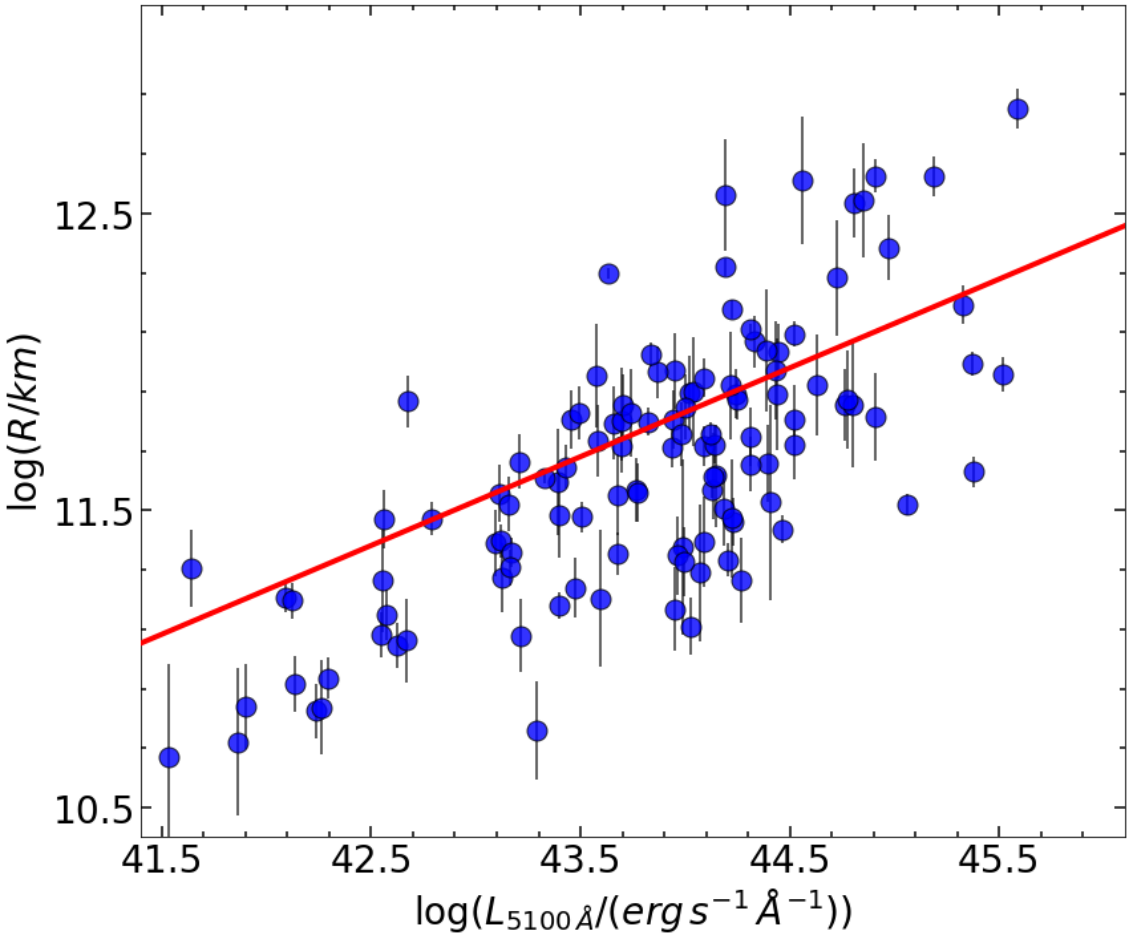
Incertezze MC



- **Si può rimpiazzare ogni dato y_j con un valore ottenuto come**

  $y_j' = y_j + \sigma_{yj}*N(0,1)$
- **Fittare di nuovo i dati ottenendo m', q'**
- **Costruire le distribuzioni $\{m\}_i$, $\{q\}_i$**
- **Calcolare le dispersioni (σ_m, σ_q) ed interpretarle come incertezze sui parametri di BF**

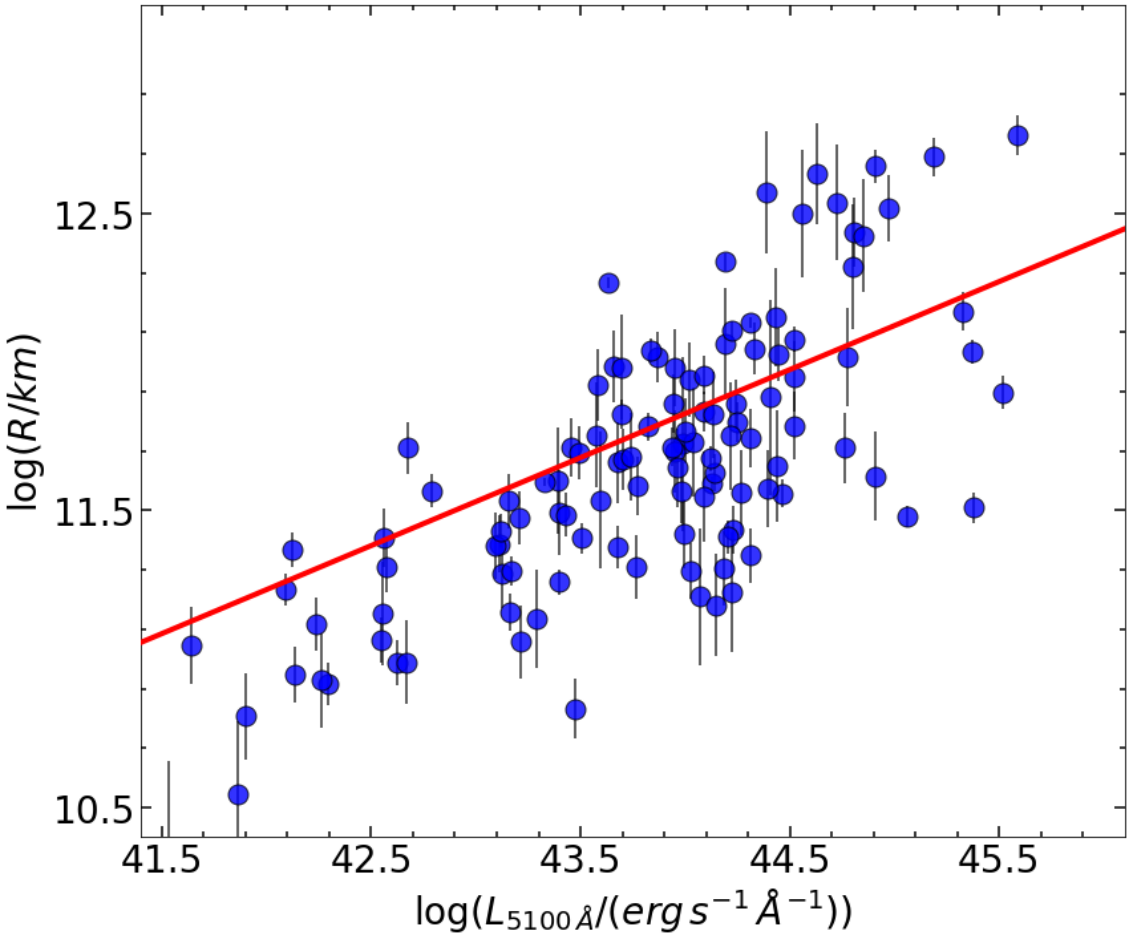
Incertezze MC



- **Si può rimpiazzare ogni dato y_j con un valore ottenuto come**

  $y_j' = yj + \sigma_{yj} * N(0,1)$
- **Fittare di nuovo i dati ottenendo m', q'**
- **Costruire le distribuzioni $\{m\}_i$, $\{q\}_i$**
- **Calcolare le dispersioni (σ_m, σ_q) ed interpretarle come incertezze sui parametri di BF**

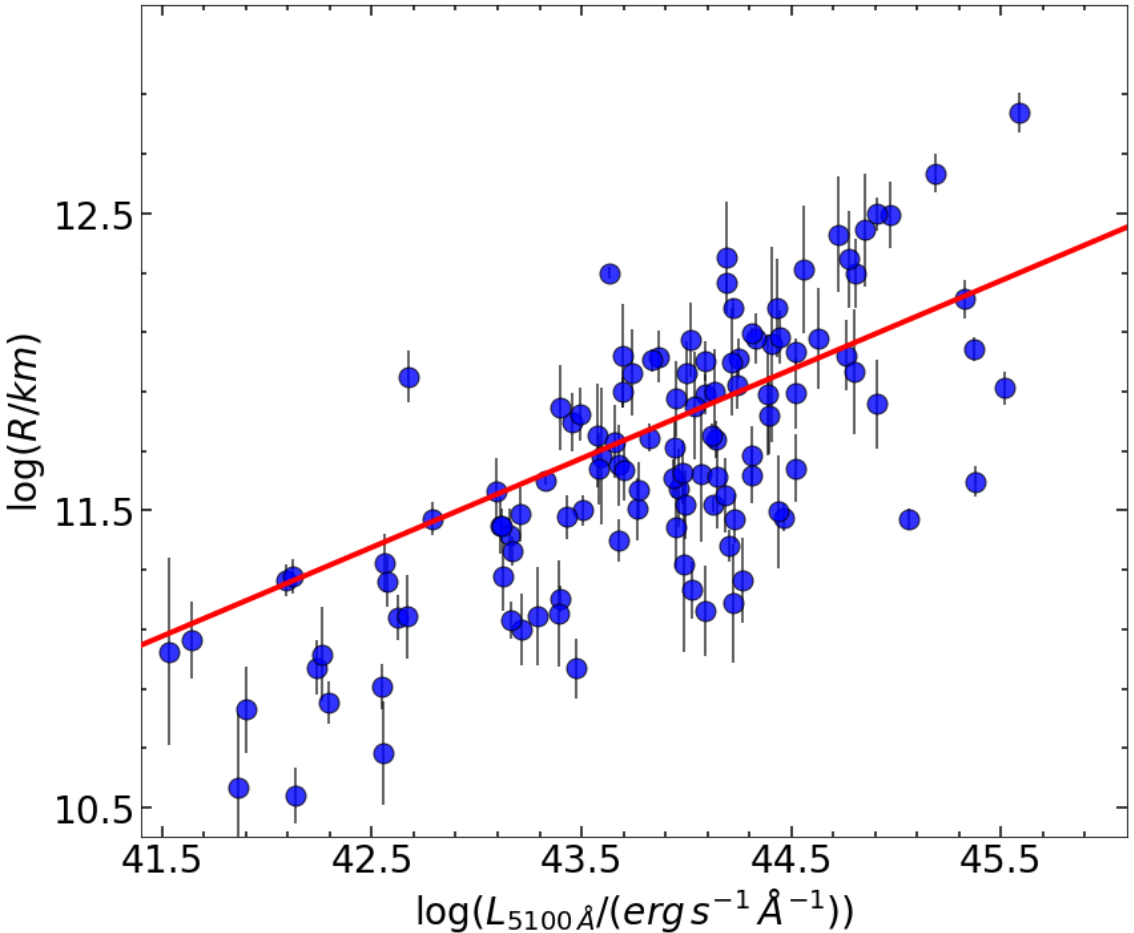
Incertezze MC



- **Si può rimpiazzare ogni dato y_j con un valore ottenuto come**

  $y_j' = y_j + \sigma_{yj}*N(0,1)$
- **Fittare di nuovo i dati ottenendo m', q'**
- **Costruire le distribuzioni $\{m\}_i$, $\{q\}_i$**
- **Calcolare le dispersioni (σ_m, σ_q) ed interpretarle come incertezze sui parametri di BF**

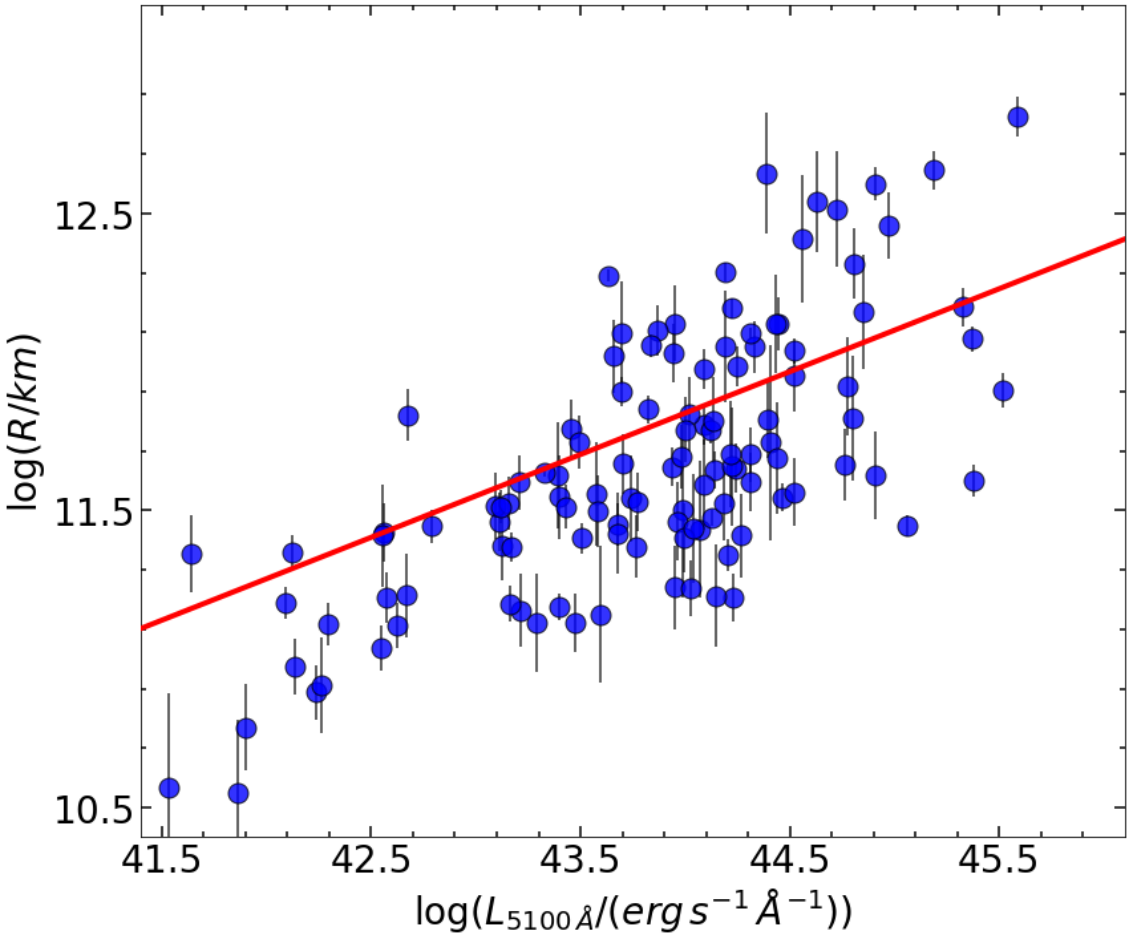
Incertezze MC



- **Si può rimpiazzare ogni dato y_j con un valore ottenuto come**

  $y_j' = y_j + \sigma_{yj}*N(0,1)$
- **Fittare di nuovo i dati ottenendo m', q'**
- **Costruire le distribuzioni $\{m\}_i$, $\{q\}_i$**
- **Calcolare le dispersioni (σ_m, σ_q) ed interpretarle come incertezze sui parametri di BF**

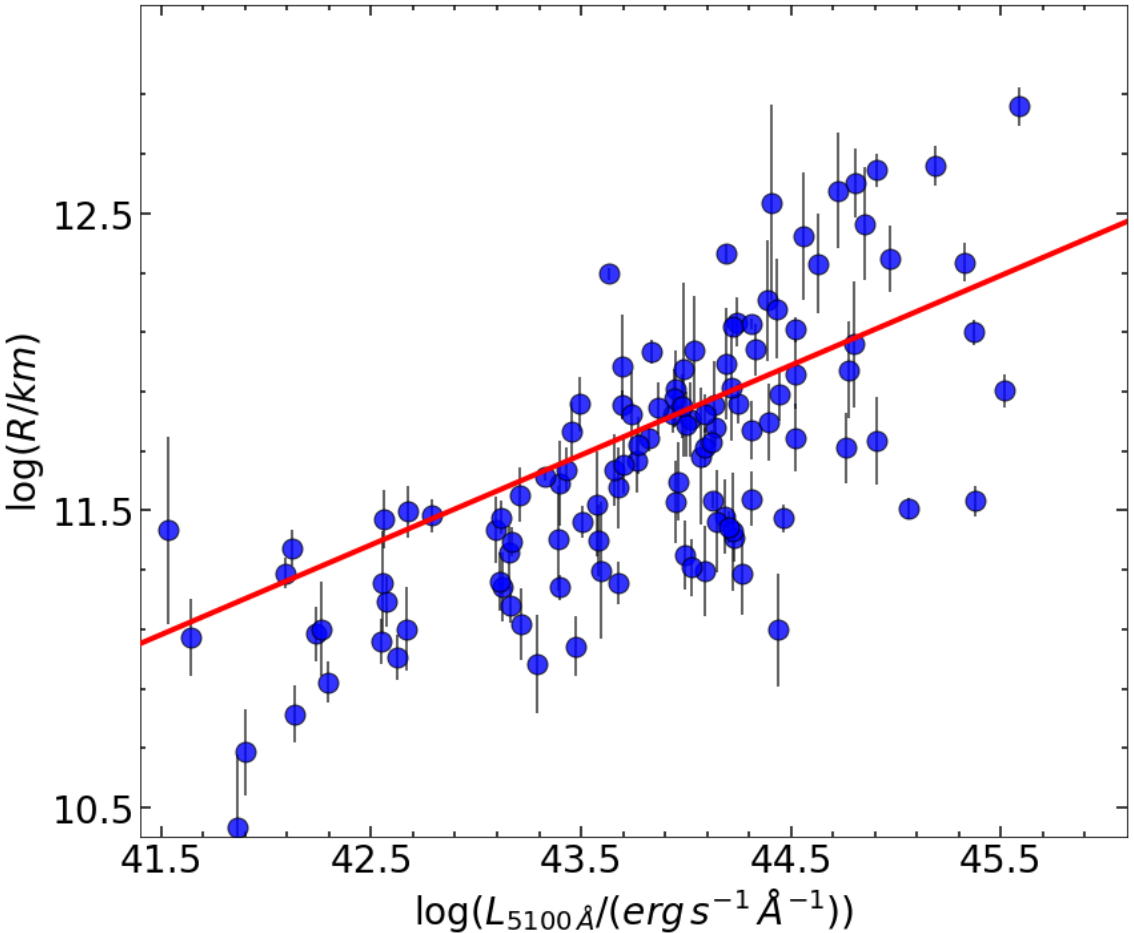
Incertezze MC



- **Si può rimpiazzare ogni dato y_j con un valore ottenuto come**

  $y_j' = y_j + \sigma_{yj} * N(0,1)$
- **Fittare di nuovo i dati ottenendo m', q'**
- **Costruire le distribuzioni $\{m\}_i$, $\{q\}_i$**
- **Calcolare le dispersioni (σ_m, σ_q) ed interpretarle come incertezze sui parametri di BF**

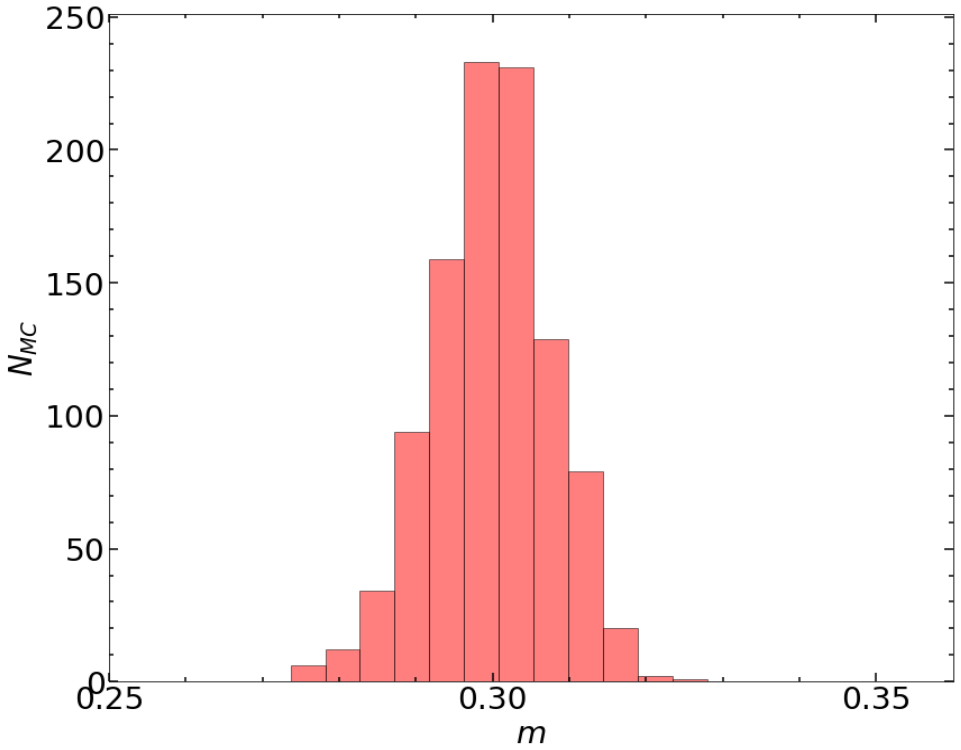
Incertezze MC



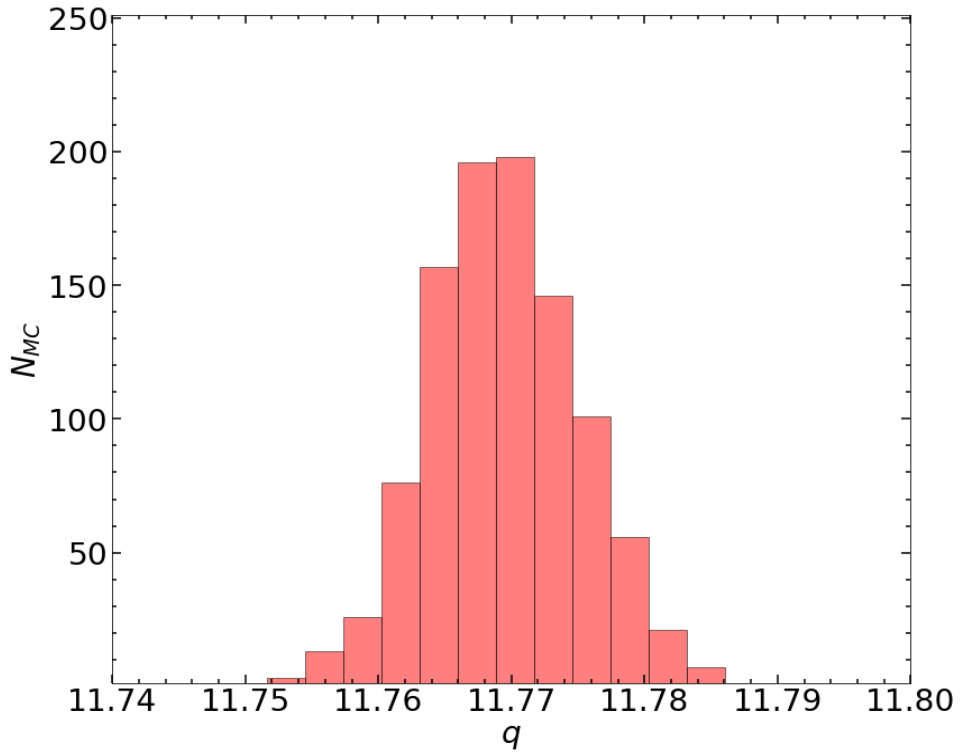
- **Si può rimpiazzare ogni dato y_j con un valore ottenuto come**

  $y_j' = y_j + \sigma_{yj}*N(0,1)$
- **Fittare di nuovo i dati ottenendo m', q'**
- **Costruire le distribuzioni $\{m\}_i$, $\{q\}_i$**
- **Calcolare le dispersioni (σ_m, σ_q) ed interpretarle come incertezze sui parametri di BF**

Incertezze MC

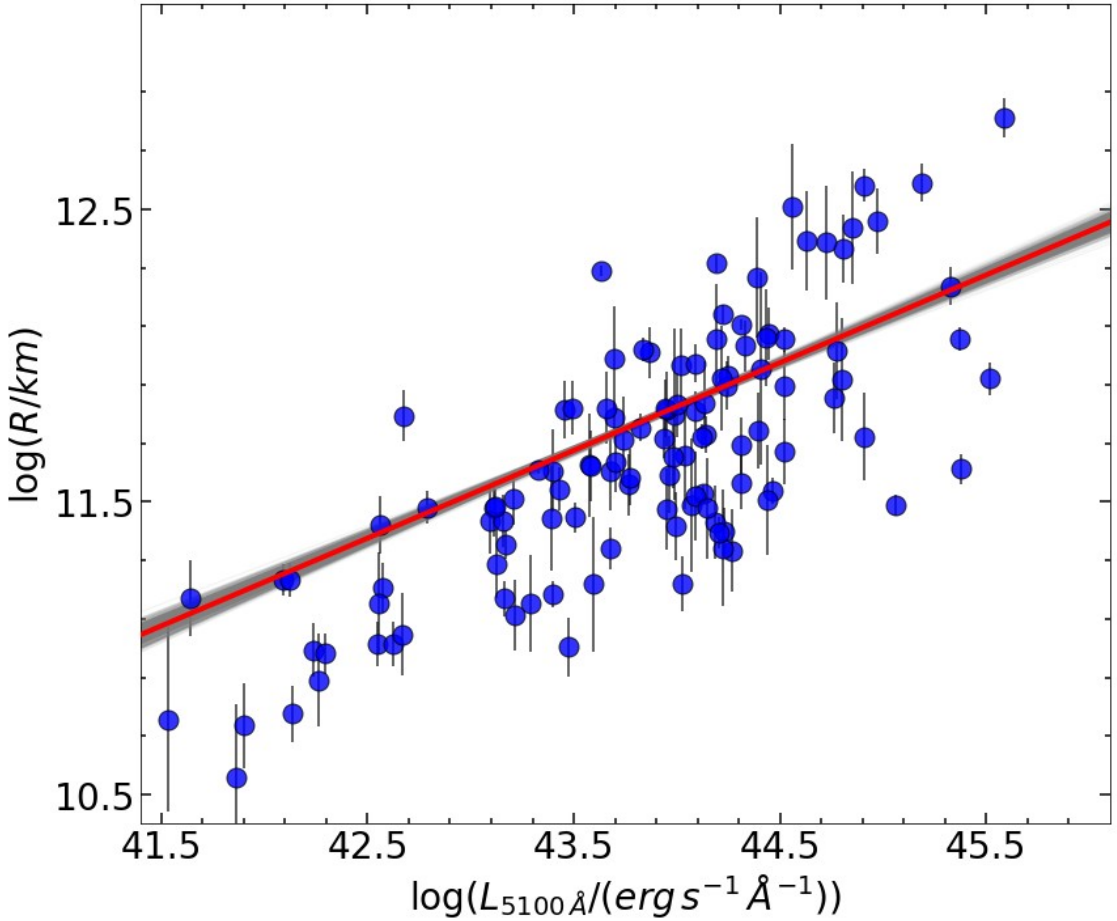


$\sigma_m = 0.008$



$\sigma_q = 0.006$

Incertezze MC

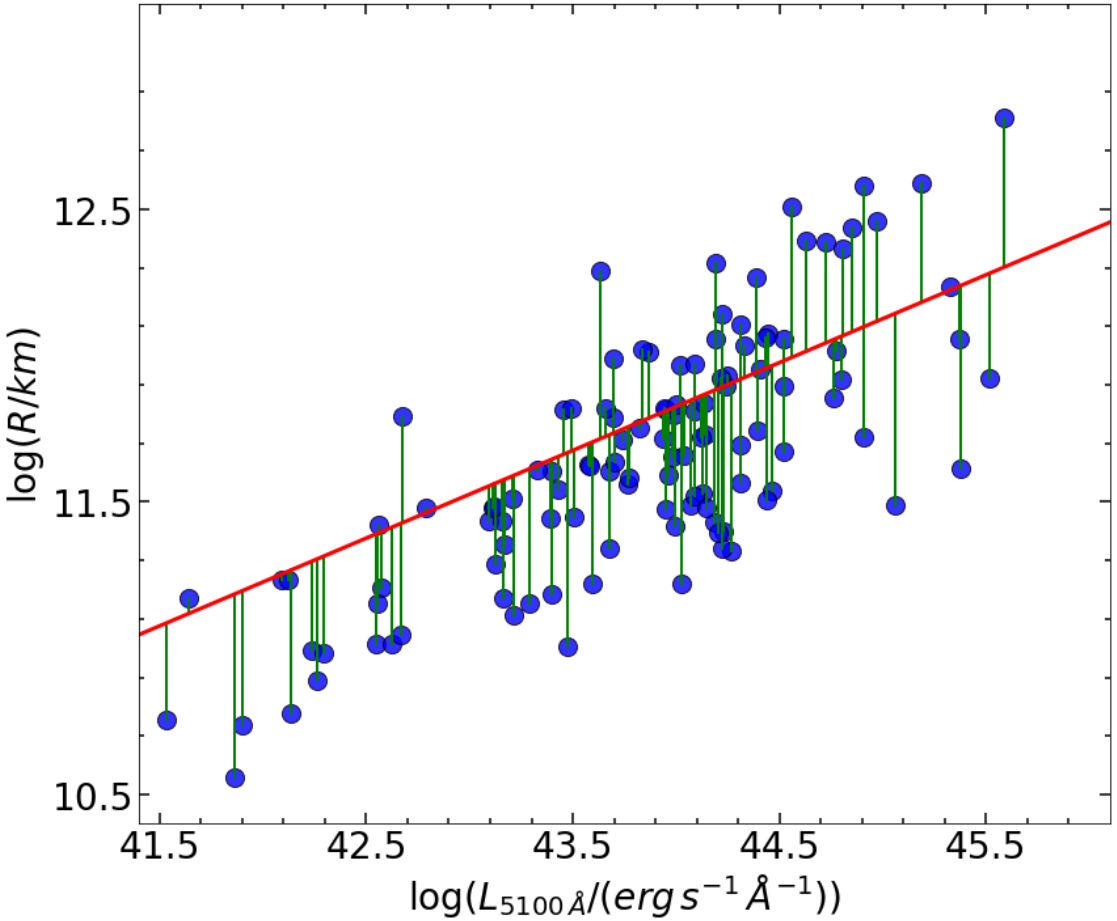


- **Molto versatile:**

  data un generico modello $f(x_i, y_i, \sigma_i, \theta_k)$ propaga le incertezze sui parametri di BF

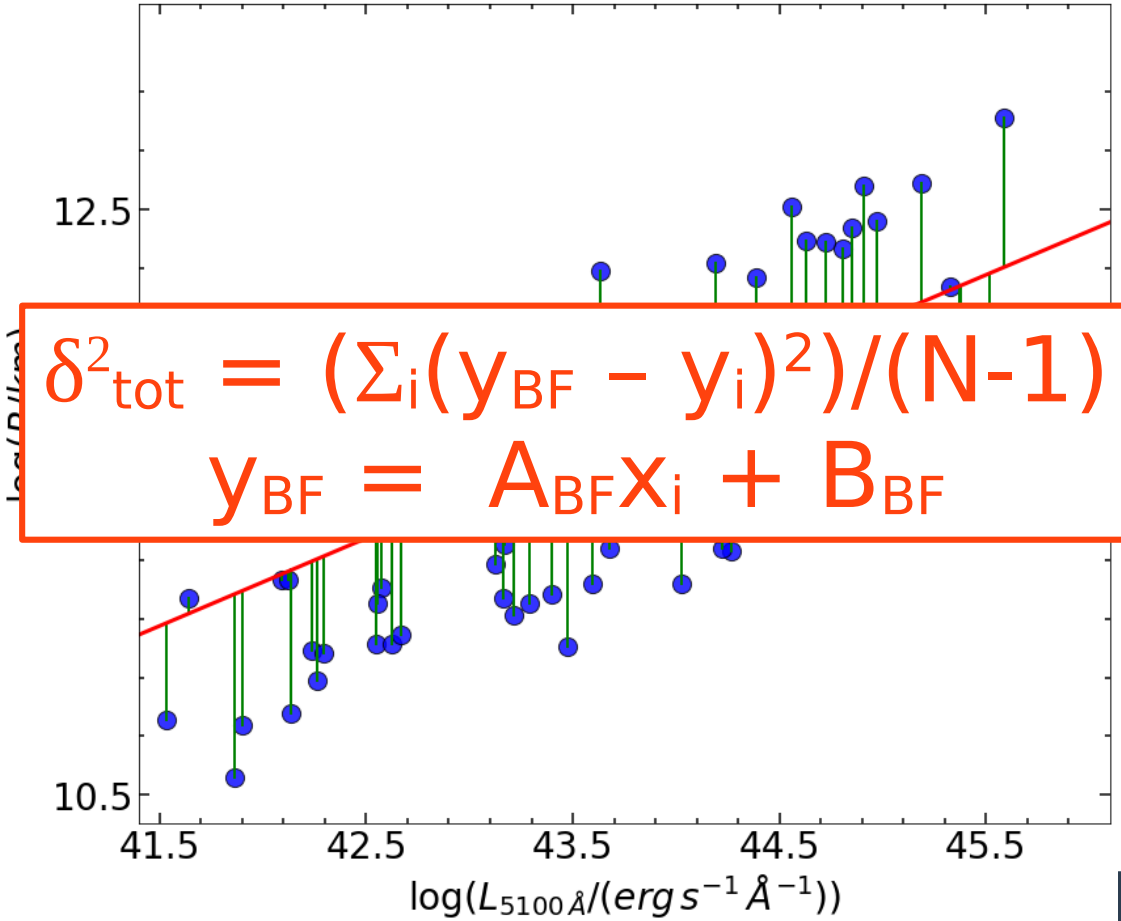
- **Tiene conto soltanto delle variazioni indipendenti fra i dati (no covarianza)**

Dispersione della relazione



- **Quanto dispersi rispetto alla retta BF sono i dati?**
- **La dispersione δ^2_{tot} quantifica quanto i dati del campione sono distanti in media dalla retta BF**
- **Forma ~ $(RSS^2/(N-1))^{0.5}$**

Dispersione della relazione

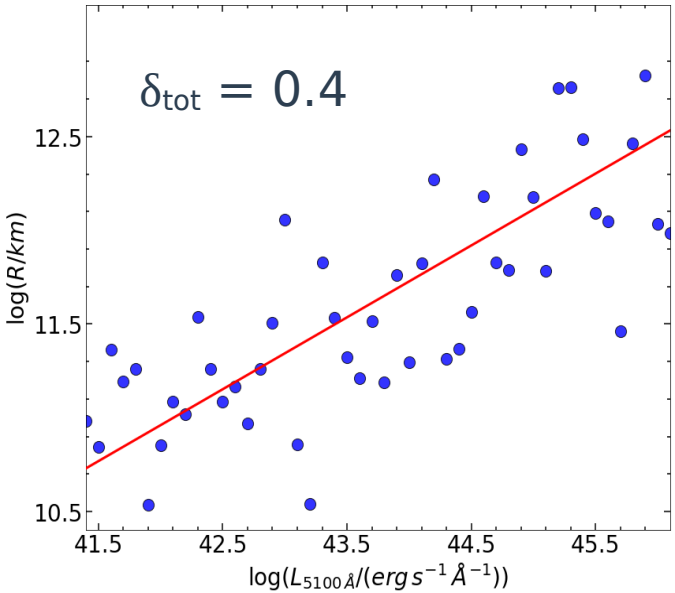


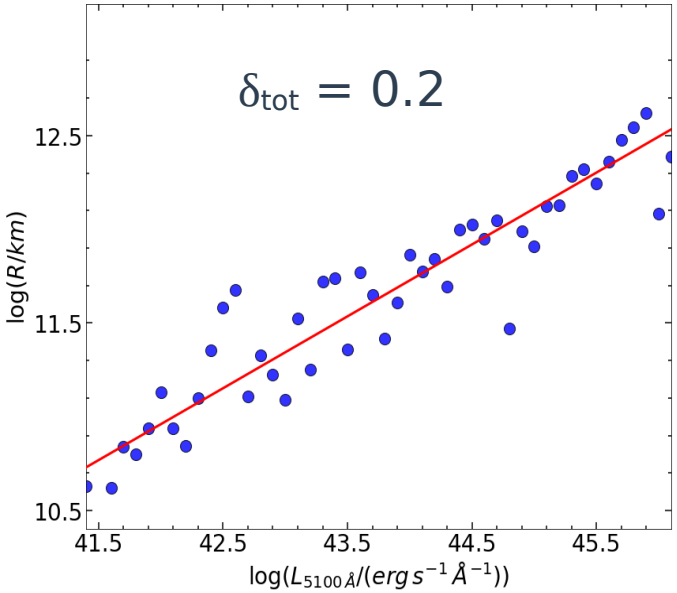
$$\delta^2_{tot} = (\Sigma_i(y_{BF} - y_i)^2)/(N-1)$$

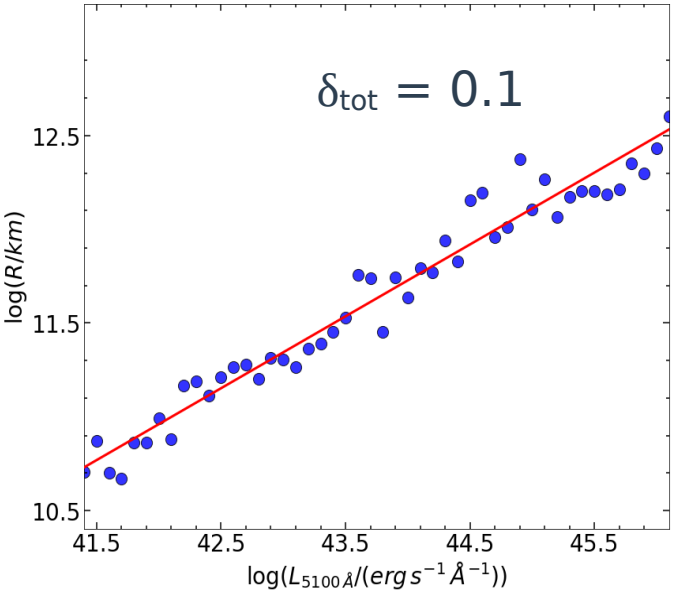
$$y_{BF} = A_{BF}x_i + B_{BF}$$

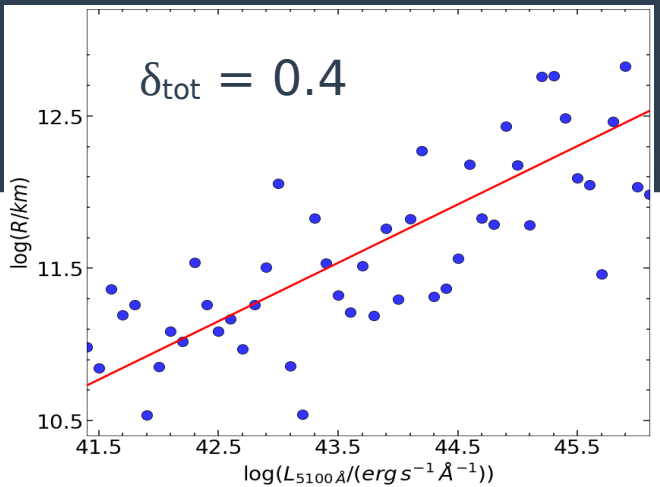
- **Quanto dispersi rispetto alla retta BF sono i dati?**
- **La dispersione δ^2_{tot} quantifica quanto i dati del campione sono distanti in media dalla retta BF**
- **Forma ~ $(RSS^2/(N-1))^{0.5}$**

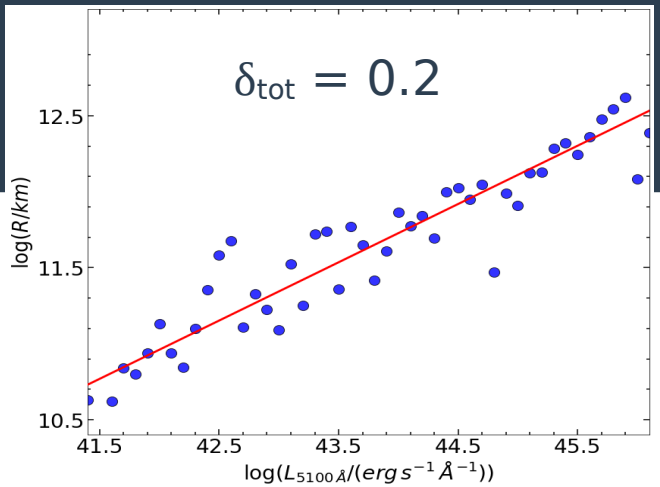
Dispersione della relazione

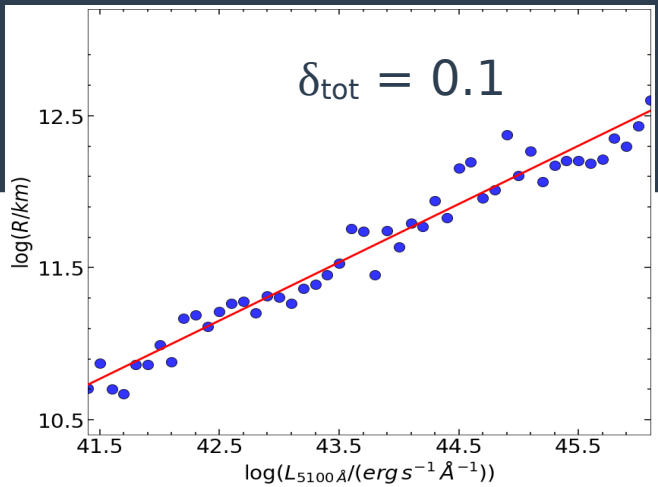


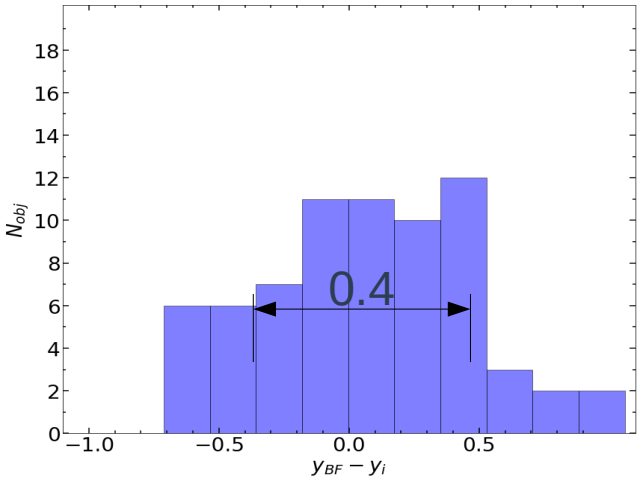


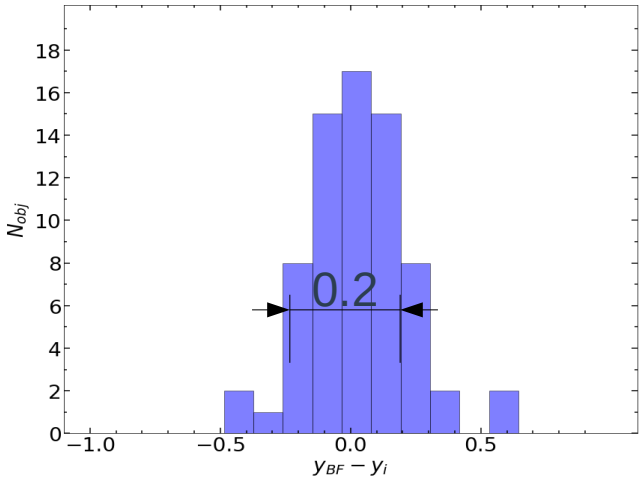


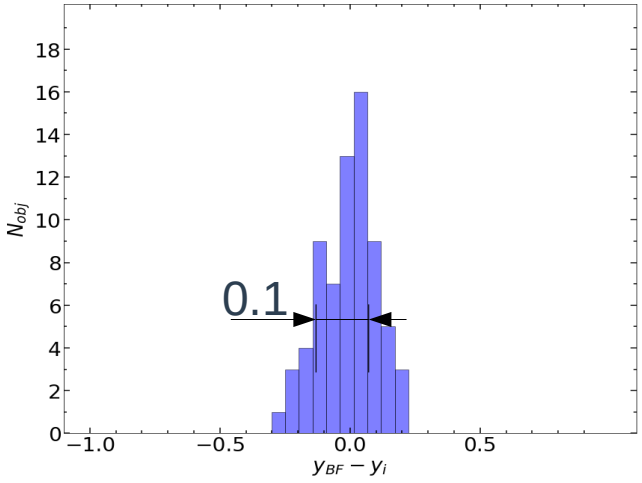

δ_{tot} = 0.4
log(R/km)
$\log(L_{5100\,Å}/(erg\,s^{-1}\,Å^{-1}))$


δ_{tot} = 0.2
log(R/km)
$\log(L_{5100\,Å}/(erg\,s^{-1}\,Å^{-1}))$


δ_{tot} = 0.1
log(R/km)
$\log(L_{5100\,Å}/(erg\,s^{-1}\,Å^{-1}))$


0.4
N_{obj}
$y_{BF} - y_i$


0.2
N_{obj}
$y_{BF} - y_i$

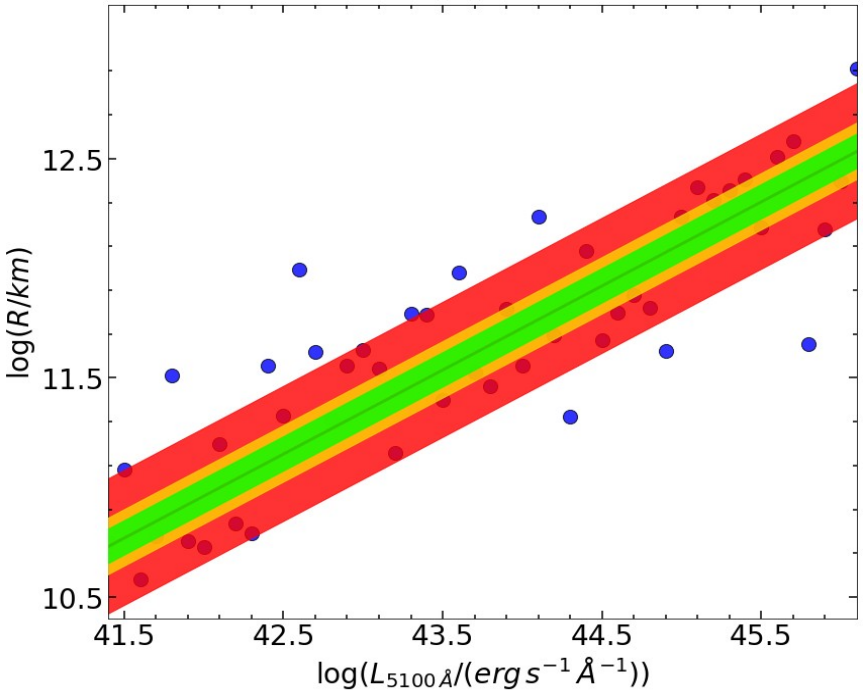

0.1
N_{obj}
$y_{BF} - y_i$

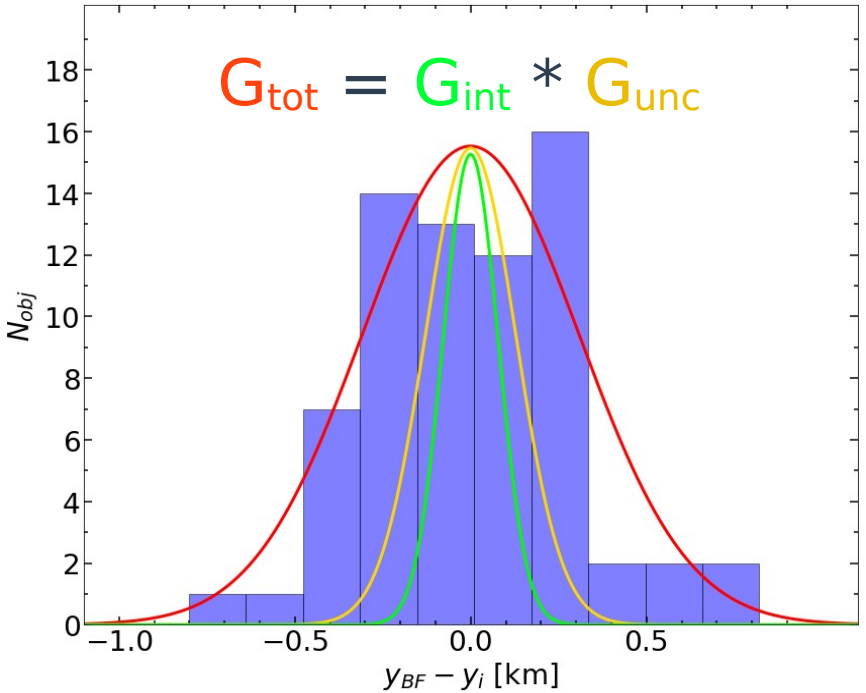
Dispersione ed incertezza di misura

- **Si può pensare che la dispersione totale sia formata da due contributi**
- **Una dispersione intrinseca δ_{int} che è dovuta al fatto che la relazione fra le grandezze in analisi può essere influenzata da fattori non presi in esame (e.g. s = vt con aria)**
- **Una dispersione dovuta alle incertezze di misura δ_{unc}**

$$\delta_{tot}^2 = \delta_{int}^2 + \delta_{unc}^2$$

Dispersione ed incertezza di misura





$\delta_{tot} = 0.30$ $\delta_{int} = 0.08$ $\delta_{unc} = 0.29$

Aggiunta di un parametro libero

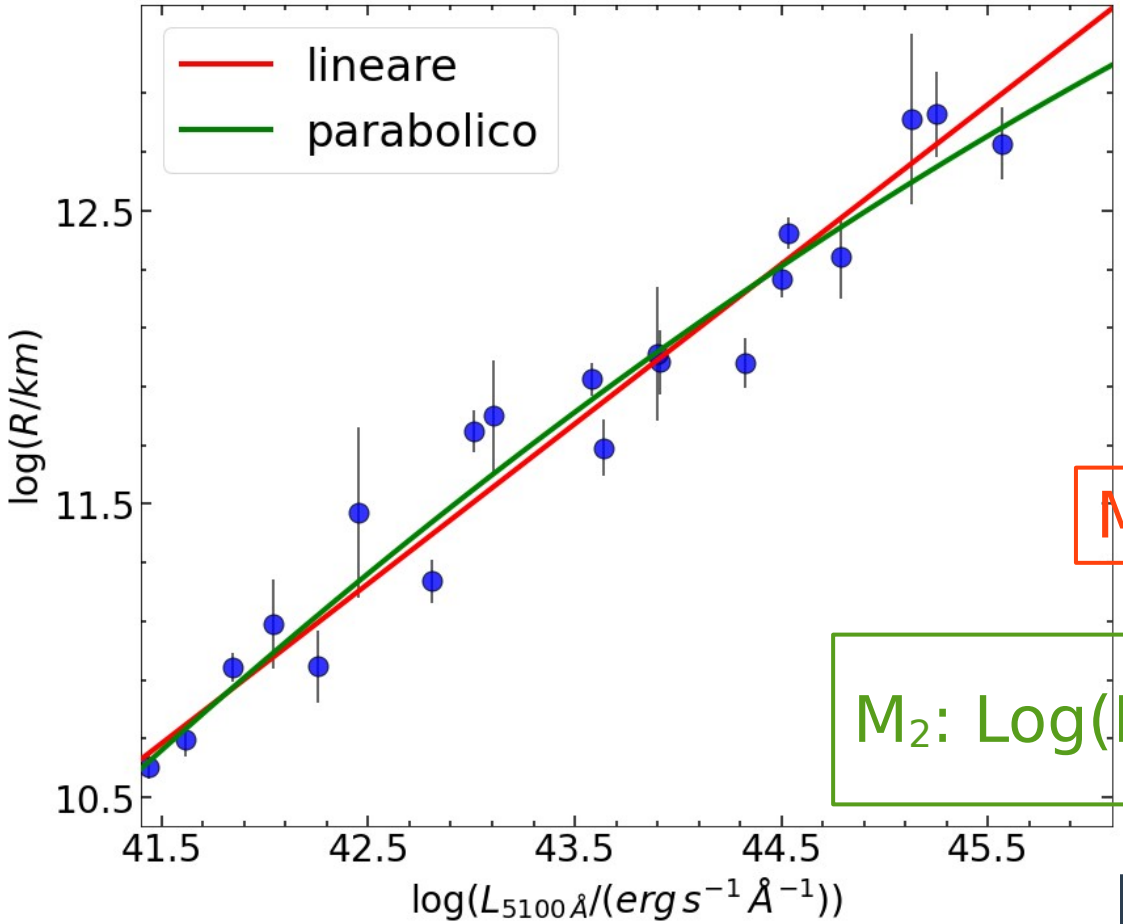
“Con quattro parametri posso fittare un elefante e con cinque posso fargli dondolare la proboscide”

J. Von Neumann

Aggiunta di un parametro libero

- **Avete condotto delle osservazioni molto lunghe e dispendiose (sigh)**
- **Al costo di ridurre significativamente il vostro campione (sigh x2) avete raccolto dei dati su un sotto-campione notevolmente meno disperso e con incertezze di misura più piccole.**
- **Trade-off *studi di popolazione vs. Rosetta stones***

Aggiunta di un parametro libero

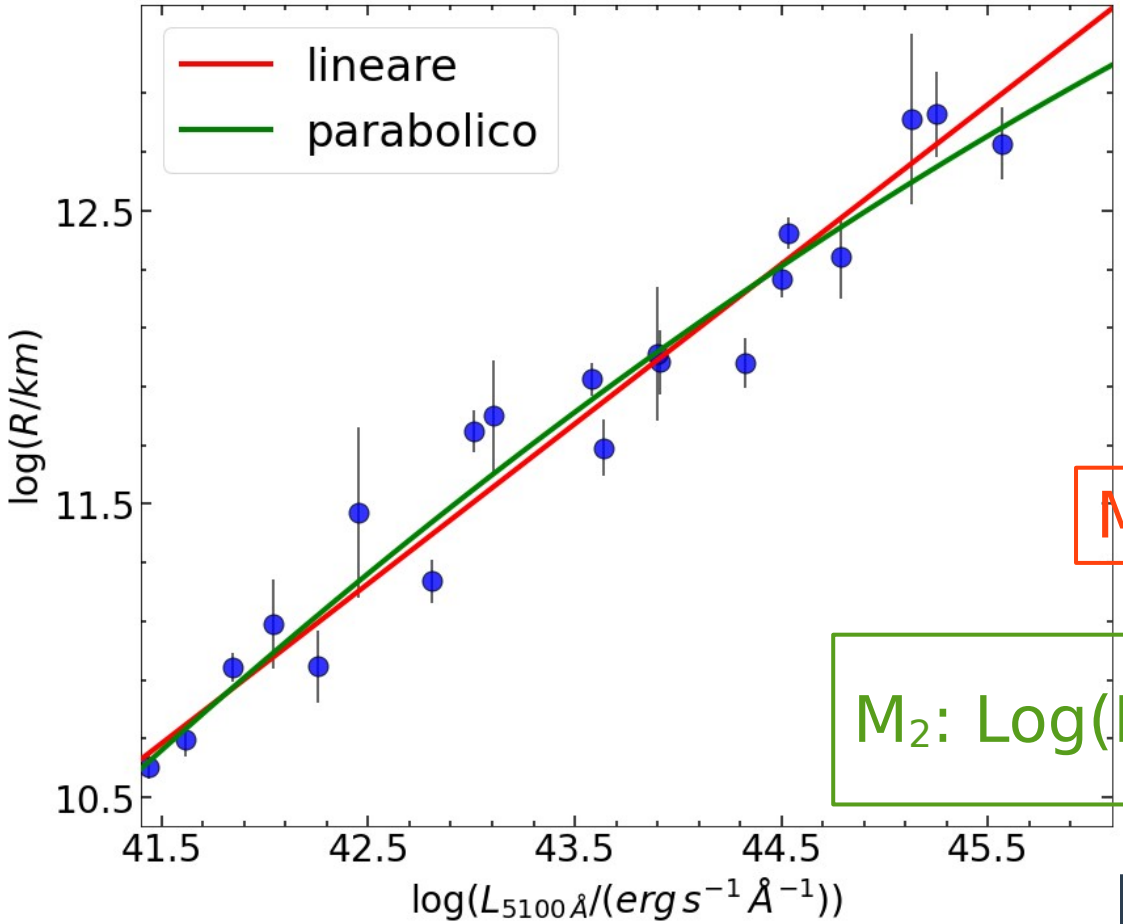


- **Testare un modello parabolico di letteratura (comunque lineare in a0, a1, a2)**
- **Vogliamo verificare quale dei due modelli descrive meglio la relazione fra i dati in analisi**

M_1: $Log(R) = a_0 + a_1 Log(L)$

M_2: $Log(R) = a_0 + a_1 Log(L) + a_2 Log(L)^2$

Aggiunta di un parametro libero

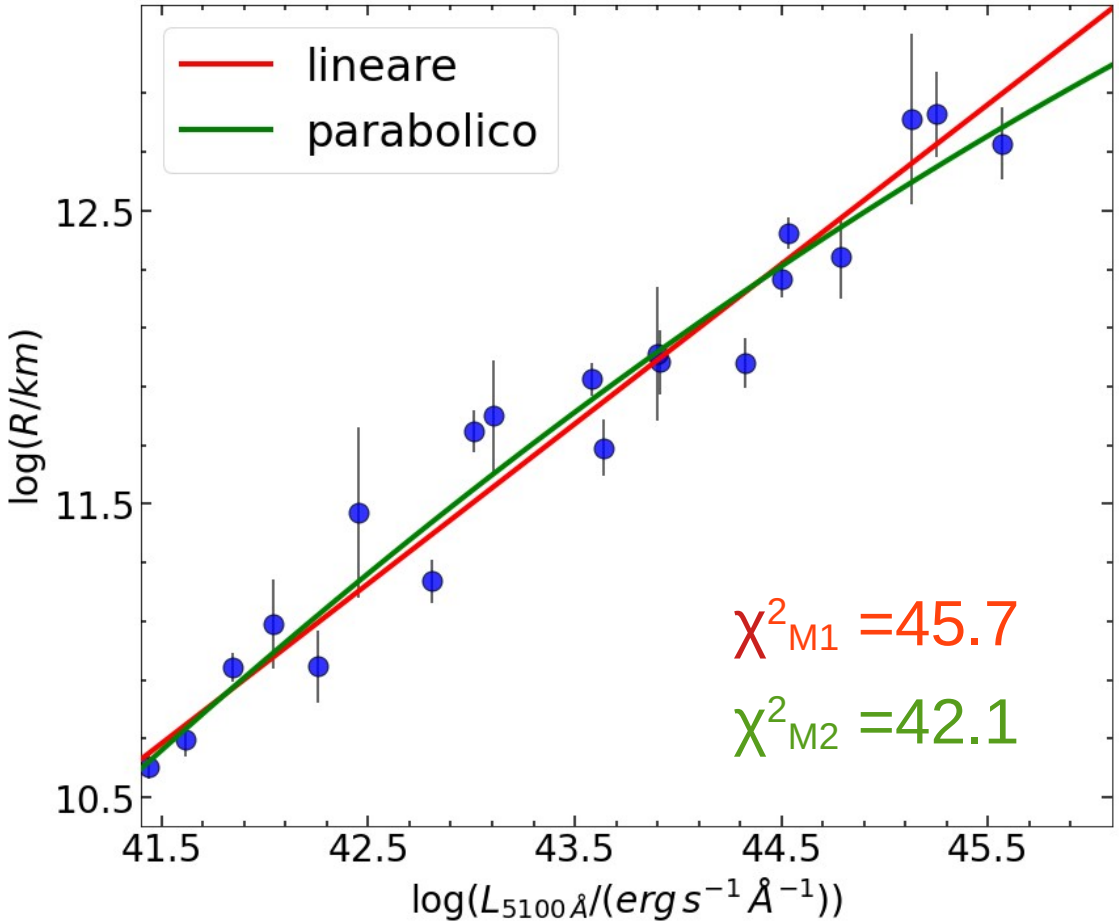


- **Nested model: M_1 può essere definito eliminando uno/alcuni parametri liberi di M_2**

M_1: $Log(R) = a_0 + a_1 Log(L)$

M_2: $Log(R) = a_0 + a_1 Log(L) + a_2 Log(L)^2$

Aggiunta di un parametro libero



- **Vari stimatori per determinare quale modello descriva meglio i dati (*L-ratio*, AKAIKE, BIC)**
- **$\Delta\chi^2$, F-test**
- **Intuitivamente: più la variazione di χ^2 è *grande* più è probabile che M_2 descriva I dati meglio di M_1**

Aggiunta di un parametro libero

- **Test di ipotesi su $\Delta\chi^2/\Delta\nu = (\chi^2_{M1} - \chi^2_{M2})/(\nu_2 - \nu_1) = 3.3/1 = 3.3$**
- **Calcolare il p-value per una distribuzione di χ^2 con $(\nu_2 - \nu_1)$ gradi di libertà**

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838

$\chi^2 < \chi^2_{crit} \rightarrow$ reject M_2 at 95% CL

Aggiunta di un parametro libero

Fisher F test $F = \dfrac{(\chi^2_{M1} - \chi^2_{M2})\ \ (p_2 - p_1)}{\chi^2_{M2}\ \ (n-p_2)} = 1.3$

14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.97	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38

$F < F_{crit}$ → rifiutare M_2 al 95% CL

Aggiunta di un parametro libero

- **ATTENZIONE! aggiungere parametri liberi va contro il rasoio di Ockham**
- **Le componenti aggiuntive devono essere fisicamente giustificate**
- **Altrimenti si può fittare rumore (*overfitting*)**

Cosa abbiamo visto

- **Perchè un fit lineare?**
- **Stimare la correlazione**
- **Il fit dei dati**
- **Stimare la dispersione dei dati rispetto al modello**
- **Incertezze con metodi MC**
- **Aggiungere complessità al modello**

Cosa NON abbiamo visto

- **Bande di confidenza**
- **Minimise the orthogonal distance to the best fit line (ODR)**
- **Sigma clipping**
- **Incertezze su entrambe le variabili (Deming regression)**
- **Fit usando la dispersione intrinseca come parametro libero (EMCEE)**
- **Argh! Le mie variabili hanno incertezze asimmetriche: cosa dovrei fare?**
- **(Bayesian) Likelihood**
- **...**

Buone feste a tutti!
Anche al Ballini