Regressione

Vers. 1.0.2

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

Outline

La regressione I

Fino ad adesso abbiamo visto come si possa dire qualcosa su un parametro di uan distribuzione, ma abbiamo sempre avuto a che fare con una variabili (con qualche eccezione, tipo i sonniferi.)
Cn la regressione ci chiediamo che relazione c'è tra una **variabile dipendente** y_i, e una serie di **predittori**, o **variabili indipendenti** o **covariate** $\mathbf{x}_i = (x_{i,1}, \dots x_{i,p})$.
Partiamo dal caso in cui $\mathbf{x}_i = x_{i,1} \equiv x_i$, che è la regressione univariata.

L'ipotesi che faremo, sempre, è di considerare $\mathbf{x}_i$ come costanti (i.e., non random), mentre Y_i sono delle variabili aleatorie.

Attenzione: passando ai modelli, diventa complicato mantenere la notazione maiuscola per le variabili aleatorie, e minuscolo per realizzazioni, ma è sempre chiaro dal contensto, e quando non lo è lo specifico)

La regressione II

Se vogliamo stimare la relazione tra x_i e Y_i, dobbiamo assumere qualcosa sulla distribuzione delle Y, che, nel caso delle regressione lineare, è distribuita come una normale

$$Y_i \sim N(\mu_i, \sigma_i^2), \quad Y_i \perp Y_{i'}$$

dove, come vedete medie e varianze sono "potenzialmente" differenti per ogni i, quindi i dati **non sono iid**, ma solo indipendenti.
Se vogliamo che la distribuzione di Y_i dipenda dalla covariate, dobbiamo avere che i parametri μ_i e σ_i^2 devono essere funzioni di x_i. Questo è il caso più generale, ma nella regressione lineare si assume la cosa seguente

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad Y_i \perp Y_{i'}$$

La regressione III

cioè $\sigma_i^2 = \sigma^2$, mentre $\mu_i = \alpha + \beta x_i$. La forma specificata sopra non è la maniera classica di rappresentare un modello regressivo, ma lo si preferisce scrivere come

$$Y_i = \alpha + \beta x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2), \quad \epsilon_i \perp \epsilon_{i'}$$

che si dimostra facilmente essere la stessa cosa. Notate come la variabile aleatoria è ϵ_i, ma non si usa il maiuscolo quì.

Importante: il modello si chiama "lineare" perchè il β e α entrano in modo lineare, per esempio

$$y_i = \alpha + \beta log x_i + \epsilon_i$$

è un modelli lineare

La regressione IV

Le ipotesi che abbiamo fatto sono chiamate le ipotesi di Gauss, e, in maniera compatta sono

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = \sigma^2$, tutto uguali, e si di che sono **omoschedastiche**
- $Cov(\epsilon_i, \epsilon_j) = 0$, se $j \neq i$

a cui va aggiunta anche un'altra ipotesi, che è la normalità di ϵ. C'è un'altra richiesta, che derivano dall'assunzione di normalità, cioè

- $E(\epsilon_i|x_i) = 0$
- $Var(\epsilon_i|x_i) = \sigma^2$

La regressione V

cioè che le ϵ devono essere indipendenti da x. Questo perchè in ϵ ci deve essere tutto ciò che noi non riusciamo a spiegare con la x, ma se la media o varianza dipendono da x, allora c'è dell'informazione in x dentro ϵ.

Il modello ci da un'interpretazione interessante sulla relazione che stiamo assumendo tra x e y, e ci dice che

- c'è un effetto rettilineo $\alpha + \beta x_i$ deterministico;
- più un effetto "random" ϵ_i.

La retta

$$y^* = \alpha + \beta x$$

viene chiamata **retta di regressione**, α è l'**intercetta**, mentre β è il **coefficiente angolare**. Il termine ϵ_i può avere diverse interpretazioni, che dipendono da cosa stiamo analizzando, come

La regressione VI

- un termine di errore, cioè, $\alpha + \beta x$ è il valore vero che dovremmo osservare, ma lo strumento non è preciso e ci da y_i;
- variabilità non spiegate. Se per esempio x_i è l'altezza e y_i il peso, mi aspetto una relazione lineare (approssimativamente), ma avrò che persone con al stessa altezza hanno pesi diversi, perchè c'è della variabilità (informazione) di cui non sto tenendo conto.

IN generale si chiama ϵ_i errore, ma non è propiamente un errore (vedi sopra).
Abbiamo quindi la distribuzione dei dati, che dipende da dei parametri $(\alpha, \beta, \sigma^2)$, che non conosciamo e vogliamo stimare, ma abbiamo, adesso, tutti gli strumenti necessari per poter stimare i aprametri e/o fare test d'ipotesi.

La regressione VII

Riprendiamo l'esempio delle temperature. è chiaro che ci sian una "dipendenza" tra le temperatura e il tempo

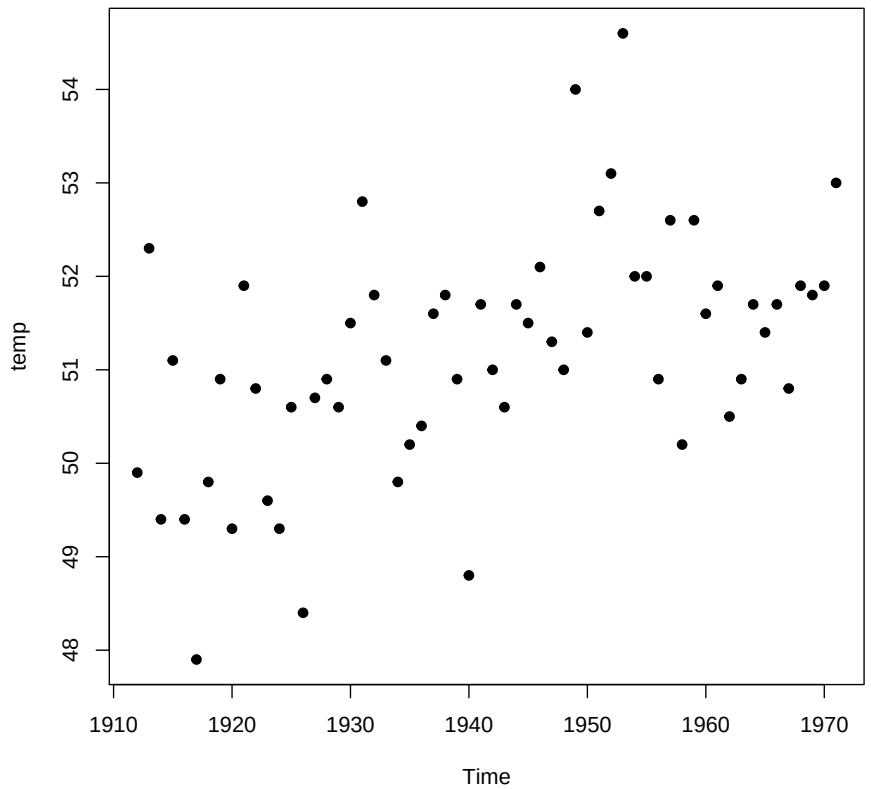

Figure: Temperature musurate

La regressione VIII

E proviamo a vedere i risultati di una regressione (delle stime dei parametri ce ne occuperemo dopo), del tipo $\text{temp}_i = \alpha + \beta \text{anno}_i + \epsilon_i$

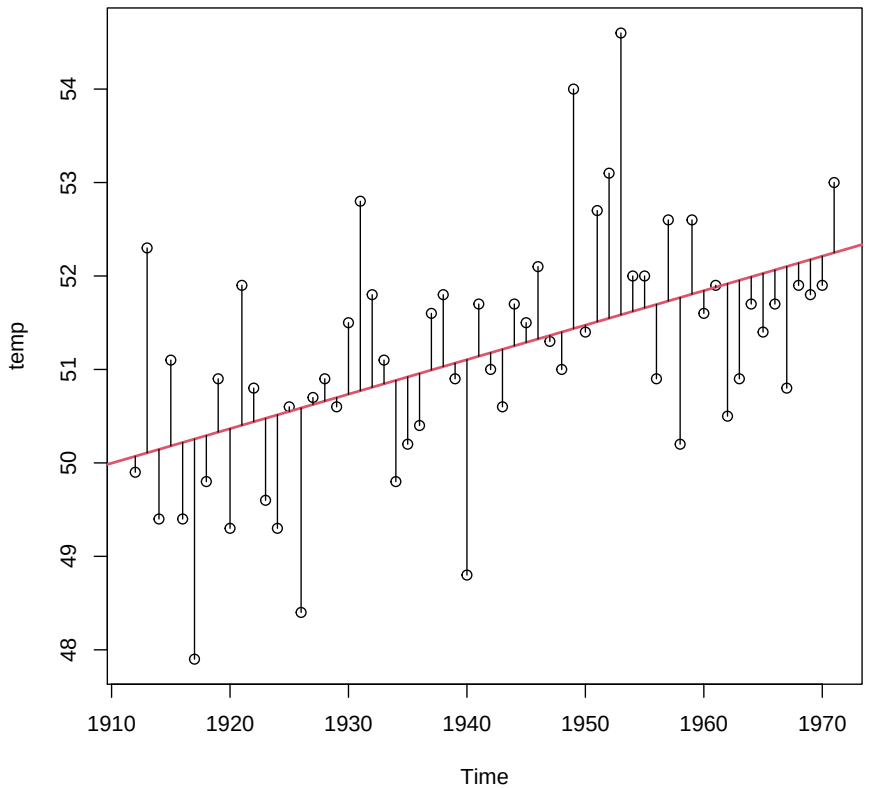

Figure: Regressione

La regressione IX

dove la linea rossa è la retta di regresione $y = \hat{\alpha} + \hat{\beta}x$, e le linee verticali sono i valori di ϵ_i.

In una regressione, soprattutto multivariata, è importante saper interpretare i parametri, cioè cose ci dicono i valori α e β (σ^2 è un parametro di disturbo, e generalmente poco interessante).

- α, da $y = \alpha + \beta x$, vediamo che α è il valore di y sulla retta, quando $x = 0$. L'interpretazione dipende da i dati che si analizzano. Per esempio, nel caso delle temperature ha poco senso, ma avrebbe senso se la regressione fosse $\text{temp}_i = \alpha + \beta \text{anno}_i^* + \epsilon_i$ dove anno_i^* è un indicatore dell'anno a partire dal 1910, o dal 1900.

La regressione X

- per interpretare β, prendiamo due valori sulla retta y_1^* e y_2^*, uno con valore x, e l'altro con valore $x+1$ come covariata. Avremmo che

$$y_1^* = \alpha + \beta x$$

mentre per il secondo valore

$$y_2^* = \alpha + \beta(x+1) = \alpha + \beta x + \beta = y_1^* + \beta$$

quindi β mi dice quanto aumenta y se aumento di 1 x.

è immediato vedere come per $\beta = 0$ non ci sia relazione tra x e y, $\beta > 0$ la relazione è di proporzionalità, mentre $\beta < 0$ ha una relazione inversamente proporzionale.

Stima dei parametri I

Passimao adesso alla stima dei parametri. Scordiamoci per un attimo di essere in un corso di statistica e definiamo il modello come

$$y_i = \alpha + \beta x_i + \epsilon_i$$

dove, in qeusto caso, ϵ_i non è una variabile aleatoria, ma solamente la differenza tra la retta e il valore osservato, che è il primo approccio che è stato usato nella regressione e si può vedere come un ottimizzazione, i.e. trovare la "migliore" retta che approssima i dati. Faccio questo esempio per far vedere cosa guadagnamo se usiamo un approccio statistico Ci chiediamo adesso come trovare delle stima di α e β. Dobbiamo decidere una funzione da ottimizzare, e la scelta più naturale è

$$SS_\epsilon = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

Stima dei parametri II

dove SS sta per **sum of squares**, e vogliamo quindi minimizzare l'errore al quadrato:

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

- Minimizziamo rispetto a α, scrivendo

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

abbiamo già visto che il valori di α che minimizza la funzione è

$$\hat{\alpha} = \bar{z} = \sum_{i=1}^{n} \frac{y_i - \beta x_i}{n} = \bar{y} - \beta \bar{x}$$

Stima dei parametri III

- Minimizziamo rispetto a β, scrivendo

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

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\beta\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + \beta^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = SS_y - 2\beta SS_{xy} + 2\beta^2 SS_x$$

Visto che vogliamo minimizzare, calcoliamo la derivata che ci da

$$-2SS_{xy} + 2\beta SS_x = 0 \Rightarrow \beta = \frac{SS_{xy}}{SS_x}$$

e visto che la derivata seconda è positiva, è un minimo.

Stima dei parametri IV

Abbiamo quindi che gli stimatore sono

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

e

$$\hat{\beta} = \frac{SS_{xy}}{SS_x}$$

Se applichiamo questi calcoli al caso delle temperature, otteniamo

$$\hat{\alpha} = -20.52, \qquad \hat{\beta} = 0.036$$

quindi cè un aumento di 0.036 gradi ogni anno. La domanda che ci dovremmo porre è se 0.036 è veramente l'effetto vero, o se non cè un effetto temporale, ma il numero che abbiamo ottenuto è dovuto "al caso". Anche perchè, se io al posto dell'anno avessi preso una qualsiasi variabile, calcolando $\hat{\beta} = \frac{SS_{xy}}{SS_x}$ non otterrei mai esattamente 0, neanche se non ci fosse nessuna relazione tra x e y.

Stima dei parametri V

In questo caso ci sarebbe utile poter applicare un qualche test d'ipotesi, del tipo $H_0 : \beta = 0$, contro $H_1 : \beta \neq 0$, ma non lo possiamo fare perchè niente nella regresione è variabile aleatoria, e quindi $\hat{\beta}$ non è variabile aleatoria.

Reintroduciamo le ipotesi sull'errore:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2), \quad \epsilon_i \perp \epsilon_{i'}$$

adesso proviamo a trovare degli stimatori per i parametri del modello. La cosa che possiamo fare è utilizzare il metodo della massima verosimiglianza

$$L(\alpha, \beta, \sigma^2|\mathbf{y}) = f(\mathbf{y}|\alpha, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right)$$

Stima dei parametri VI

come nel caso di variabili iid da una normale, la massimizzazine di (α, β) è indipendente da quella di σ^2 perchè per qualsiasi valore di σ^2 noi dobbiamo sempre trovare i valori che massimizzano

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

che equivale a dire, i valore che minimizzano

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

che è la stessa funzione che abbiamo minimizzato prima e quindi gli stimatori sono ancora una volta

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

e

$$\hat{\beta} = \frac{SS_{xy}}{SS_x}$$

Stima dei parametri VII

Per σ^2 dobbiamo massimizzare la verosimiglianza di una normale, che ci da il classico stimatore di massima verosimiglianza

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n}$$

anche se viene spesso sostituito con

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

che non è distorto.

Rispetto al caso precedente abbiamo guadagnato che

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

e

$$\hat{\beta} = \frac{SS_{xy}}{SS_x}$$

sono funzioni di variabili aleatorie, le y_i, e quindi sono variabili aleatorie esse stesse.

Qualche risultato:

Stima dei parametri VIII

- abbiamo che

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

e, usando la relazione

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

possiamo scrivere

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

che ci mostra come $\hat{\beta}$ sia combinazione lineari di normali, e quindi è distribuita come una normale.

Stima dei parametri IX

- Per il parametro di intercetta abbiamo

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \bar{x}\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} y_i \left(\frac{1}{n} - \bar{x}\frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

  che è quindi distribuita come una normale.

Abbiamo quindi

$$\hat{\alpha} \sim N\left(\mu_\alpha, \sigma_\alpha^2\right)$$

e

$$\hat{\beta} \sim N\left(\mu_\beta, \sigma_\beta^2\right)$$

con parametri da determinare.

$$\mu_\beta = E(\hat{\beta}) = E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Stima dei parametri X

ma visto che

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

abbiamo che

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

quindi lo stimatore di β è **corretto**.

$$\sigma^2_\beta=Var(\hat{\beta})=Var\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{SS_x}\right)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2Var(y_i)}{SS_x^2}=$$

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

Stima dei parametri XI

Facciamo gli stessi calcoli per $\hat{\alpha}$

$$\mu_\alpha = E(\hat{\alpha}) = E(\bar{y} - \hat{\beta}\bar{x}) = \frac{\sum_{i=1}^n E(y_i)}{n} - \beta\bar{x} = \frac{\sum_{i=1}^n \alpha}{n} + \beta\frac{\sum_{i=1}^n x_i}{n} - \beta\bar{x} = \alpha$$

mostrando che è corretto. La sua varianza è

$$\sigma_\alpha^2 = Var(\hat{\alpha}) = Var\left(\sum_{i=1}^n y_i\left(\frac{1}{n} - \bar{x}\frac{(x_i-\bar{x})}{SS_x}\right)\right) = \sum_{i=1}^n Var(y_i)\left(\frac{1}{n} - \bar{x}\frac{(x_i-\bar{x})}{SS_x}\right)^2 =$$

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

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

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

Stima dei parametri XII

e ricordando che

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

allora

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

Possiamo anche trovare la covarianza tra i due stimatori

$$Cov(\hat{\alpha}, \hat{\beta}) = Cov\left(\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}\bar{x}, \hat{\beta} \right) = Cov\left(\frac{\sum_{i=1}^{n} y_i}{n}, \hat{\beta} \right) - Cov\left(\hat{\beta}\bar{x}, \hat{\beta} \right) =$$

$$Cov\left(\frac{\sum_{i=1}^{n} y_i}{n}, \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{SS_x} \right) - Cov\left(\hat{\beta}\bar{x}, \hat{\beta} \right) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})Var(y_i)}{nSS_x} - \bar{x}\frac{\sigma^2}{SS_x} =$$

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

Stima dei parametri XIII

Che dice delle cose interessanti

- La covarianza/correlazione è negativa;
- si annulla solo se la media campionaria di x è zero (per questo motivo spesso si standardizzano le x)
- quindi

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{nSS_x} & -\bar{x}\frac{\sigma^2}{SS_x} \\ -\bar{x}\frac{\sigma^2}{SS_x} & \frac{\sigma^2}{SS_x} \end{pmatrix}\right)$$

Prima di passare al test d'ipotesi, dobbiamo ancora provare un paio di risultati, perchè se volessimo dare delle ipotesi sui parametri regressivi, abbiamo bisogno di una statistica per construire il test, ma non possiamo usare, per esempio

$$Z = \frac{\hat{\beta} - \beta_0}{\sqrt{Var(\hat{\beta})}}$$

Stima dei parametri XIV

se $H_o : \beta = \beta_0$, perchè $Var(\hat{\beta})$ dipende da σ^2. Potremmo cercare una variabile $J \sim \chi(m)$, e utilizzare

$$T = \left(\frac{\hat{\beta} - \beta_0}{\sqrt{Var(\hat{\beta})}} \right) / \sqrt{J/m}$$

scegliendo J in modo tale che T non dipenda più da σ^2 (come facciamo con S^2 per i dati iid). Ma per usare S^2 dobbiamo definire cosè S^2 in questo contesto e dimostrare che è indipendente da $\hat{\alpha}$ e $\hat{\beta}$.
Introduciamo i **residui**, che sono la stima dell'errore

$$\hat{\epsilon}_j = y_j - \hat{\alpha} - \hat{\beta} x_j = y_j - \hat{y}_j$$

dove $\hat{y}_j$ è la **stima o previsione** di y_j, e calcoliamo

$$Cov(\hat{\epsilon}_j, \hat{\beta}) = Cov(y_j - \hat{\alpha} - \hat{\beta} x_j, \hat{\beta}) = Cov(y_j, \hat{\beta}) - Cov(\hat{\alpha}, \hat{\beta}) - Cov(\hat{\beta} x_j, \hat{\beta}) =$$

Stima dei parametri XV

$$Cov(y_j, \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{SS_x}) + \bar{x}\frac{\sigma^2}{SS_x} - x_j\frac{\sigma^2}{SS_x} = \sigma^2\frac{(x_j - \bar{x})}{SS_x} + \bar{x}\frac{\sigma^2}{SS_x} - x_j\frac{\sigma^2}{SS_x} = 0$$

e abbiamo anche che

$$Cov(\hat{\epsilon}_j, \hat{\alpha}) = Cov(y_j - \hat{\alpha} - \hat{\beta}x_j, \hat{\alpha}) = Cov(y_j, \hat{\alpha}) - Cov(\hat{\alpha}, \hat{\alpha}) - x_j Cov(\hat{\beta}, \hat{\alpha}) =$$

$$Cov(y_j, \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}\bar{x}) - \sigma^2\frac{\sum_{i=1}^{n} x_i^2}{nSS_x} + \frac{\sigma^2 x_j\bar{x}}{SS_x} =$$

$$Cov(y_j, \frac{\sum_{i=1}^{n} y_i}{n}) - \bar{x}Cov(y_j, \hat{\beta}) - \sigma^2\frac{\sum_{i=1}^{n} x_i^2}{nSS_x} + \frac{\sigma^2 x_j\bar{x}}{SS_x} =$$

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

$$\sigma^2\frac{SS_x - n\bar{x}(x_j - \bar{x}) - \sum_{i=1}^{n} x_i^2 + nx_j\bar{x}}{nSS_x} = \sigma^2\frac{SS_x - \sum_{i=1}^{n} x_i^2 - n\bar{x}(x_j - \bar{x} - x_j)}{nSS_x} =$$

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

Stima dei parametri XVI

ma visto che

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

allora

$$Cov(\hat{\epsilon}_j, \hat{\alpha}) = \sigma^2 \frac{-n\bar{x}^2 + n\bar{x}^2}{nSS_x} = 0$$

Da qeusti due risultati $Cov(\hat{\epsilon}_j, \hat{\alpha}) = Cov(\hat{\epsilon}_j, \hat{\beta}) = 0$, abbiamo anche che

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

è indipendente da $\hat{\alpha}$ e $\hat{\beta}$. Si può dimostrare (**Teorema di Cochran**) che se

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{(n-2)}$$

Stima dei parametri XVII

allora

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}$$

Possiamo adesso tornare ai test sui parametri e ricordando che

$$\frac{Z}{\sqrt{V/(m)}} \sim T_m$$

se $Z \sim N(0,1)$ e $V \sim \chi_m$, prendento

$$Z_\alpha = \frac{\hat{\alpha} - \alpha}{\sqrt{\frac{\sigma^2 \sum_{i=1}^n x_i^2}{nSS_x}}} \sim N(0,1)$$

$$Z_\beta = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{SS_x}}} \sim N(0,1)$$

Stima dei parametri XVIII

e

$$V = \frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}$$

allora

$$T_\alpha = \frac{Z_\alpha}{\sqrt{V/(n-2)}} = \frac{\hat{\alpha} - \alpha}{\sqrt{\frac{\sigma^2 \sum_{i=1}^n x_i^2}{nSS_x}} \sqrt{\frac{(n-2)S^2}{\sigma^2} \frac{1}{n-2}}} = \frac{\hat{\alpha} - \alpha}{\sqrt{\frac{S^2 \sum_{i=1}^n x_i^2}{nSS_x}}} \sim T_{n-2}$$

e

$$T_\beta = \frac{Z_\beta}{\sqrt{V/(n-2)}} = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{SS_x}} \sqrt{\frac{(n-2)S^2}{\sigma^2} \frac{1}{n-2}}} = \frac{\hat{\beta} - \beta}{\sqrt{\frac{S^2}{SS_x}}} \sim T_{n-2}$$

Quindi, se abbiamo un ipotesi del tipo

$$H_0 : \beta = \beta_0, \qquad H_1 : \beta \neq \beta_0$$

Stima dei parametri XIX

ci da la seguente regione di accettazione

$$A(\mathbf{Y}) = \{\mathbf{Y} : \beta_0 - t_{n-2,\alpha/2}\sqrt{\frac{S^2}{SS_x}} \leq \hat{\beta} \leq \beta_0 + t_{n-2,\alpha/2}\sqrt{\frac{S^2}{SS_x}}\}$$

e intervallo di confidenza

$$[\hat{\beta} - t_{n-2,\alpha/2}\sqrt{\frac{S^2}{SS_x}}, \hat{\beta} + t_{n-2,\alpha/2}\sqrt{\frac{S^2}{SS_x}}]$$

oppure Quindi, se abbiamo un ipotesi del tipo

$$H_0 : \alpha = \alpha_0, \qquad H_1 : \alpha \neq \alpha_0$$

ci da la seguente regione di accettazione

$$A(\mathbf{Y}) = \{\mathbf{Y} : \alpha_0 - t_{n-2,\alpha/2}\sqrt{\frac{S^2\sum_{i=1}^n x_i^2}{nSS_x}} \leq \hat{\alpha} \leq \alpha_0 + t_{n-2,\alpha/2}\sqrt{\frac{S^2\sum_{i=1}^n x_i^2}{nSS_x}}\}$$

Stima dei parametri XX

e intervallo di confidenza

$$[\hat{\alpha} - t_{n-2,\alpha/2}\sqrt{\frac{S^2\sum_{i=1}^n x_i^2}{nSS_x}}, \hat{\alpha} + t_{n-2,\alpha/2}\sqrt{\frac{S^2\sum_{i=1}^n x_i^2}{nSS_x}}]$$

Si può utilizzare anche la stastistica

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}$$

per fare test come

$$H_0 : \sigma^2 = c \qquad H_1 : \sigma^2 \neq c$$

Stima dei parametri XXI

I comandi di R, con l'output (vedere anche il file) con cui si fa una regressione sono i seguenti

```
reg = lm(y ~ x)# comando
summary(reg) # comando
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.35543 -0.65074 -0.08616  0.64231  3.01540

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -20.522834  15.897593  -1.291     0.202
x             0.036921   0.008188   4.509 0.0000322 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.098 on 58 degrees of freedom
Multiple R-squared:  0.2596,    Adjusted R-squared:  0.2468
F-statistic: 20.33 on 1 and 58 DF,  p-value: 0.00003218
```

Stima dei parametri XXII

In cui la colonna "Estimates" da le stime, "t-value" il valore della statistica t osservata, "$Pr(>|t|)$" è il p-value, mentre "Std. Error" sono gli errori standard, cioè il denominatore di $\hat{\alpha}$ ($\sqrt{\frac{S^2 \sum_{i=1}^{n} x_i^2}{n SS_x}}$) e $\hat{\beta}$ ($\sqrt{\frac{S^2}{SS_x}}$). Il test che fa è un test in cui valuta se i aprametri sono pari a 0.

R^2 e previsione I

La domanda che dovrebbe venire spontanea è "quanto la regressione approssima bene i dati". La risposta a questa domanda non è immediata e bisogna un po' ragionare su cosa si intende. La cosa più naturale da chiedersi è quanto siano piccoli i residui o al loro somma

$$SS_r = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(SS_r è residuals SS) che però dipende dall scala in cui le y sono misurate e assume valori su tutto $\mathbb{R}$. Cerchiamo invece un valore che è indipendente dalla scala e che ci dia un valore compreso tra 0 (modello inutile) a 1 (modello "perfetto"). Prendiamo

$$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R^2 e previsione II

che indica quanta varianza cè nei dati prima di fare una regressione, e vediamo come questa è collegata a SS_r:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) =$$

$$SS_r + SS_{reg} + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

dove indichiamo $SS_{reg} = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ come i quadrati dovuti alla regressione. Concentriamoci sull'ultimo termine

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

e visto che

$$\hat{y}_i - \bar{y} = \hat{\alpha} + \hat{\beta}x_i - \bar{y} = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i - \bar{y} = \hat{\beta}(x_i - \bar{x})$$

R^2 e previsione III

e

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) = (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})$$

abbiamo che

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}))\hat{\beta}(x_i - \bar{x}) =$$

$$\hat{\beta}\sum_{i=1}^{n}((y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}(x_i - \bar{x})^2) =$$

$$\hat{\beta}(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2) = \hat{\beta}(SS_{xy} - \hat{\beta}SS_x)$$

$$\hat{\beta}(SS_{xy} - \frac{SS_{xy}}{SS_x}SS_x) = 0$$

quindi

$$SS_y = SS_r + SS_{reg}$$

R^2 e previsione IV

e appare natural allora definire un indice (chiamato r-quadro o r-quadrato)

$$R^2 = 1 - \frac{SS_r}{SS_y} = \frac{SS_{reg}}{SS_y}$$

che mi dice la percentuale di variazione totale SS_y viene spiegata dalla regressione SS_{reg}.

R^2 e previsione V

per concludere la parte sulla regressione semplice, vediamo come si fa **previsione**. L'idea della previsione e di dire qualcosa su una y_o non osservata, di cui però abbiamo le covariate (nel caso della temperature, potremmo cheiderci cosa succederà l'anno dopo l'ultima osservazione). Assumiamo σ^2 noto.
Sappiamo che

$$y_o = \alpha + \beta x_o + \epsilon_o$$

Noi non conosciamo i parametri ma le stime e possiamo dire che abbiamo

$$y_o = \hat{\alpha} + \hat{\beta} x_o + \epsilon_o.$$

Possiamo fare diverse previsioni

R^2 e previsione VI

- **Prevedo il valore medio di** y_o, che è $\mu_0 = \alpha + \beta x_o$ con $\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_o$. Visto che è combinazioni di normali, è ancora normale con

$$E(\hat{y}_0) = E(\hat{\alpha} + \hat{\beta} x_o) = \alpha + \beta x_o$$

quindi è corretto, e varianza

$$Var(\hat{y}_0) = Var(\hat{\alpha} + \hat{\beta} x_o) = Var(\hat{\alpha}) + x_o^2 Var(\hat{\beta}) + 2 x_o Cov(\hat{\alpha}, \hat{\beta}) =$$

$$\frac{\sigma^2 \sum_{i=1}^n x_i^2}{n SS_x} + \frac{x_o^2 \sigma^2}{SS_x} - 2\frac{x_o \sigma^2 \bar{x}}{SS_x} = \frac{\sigma^2}{SS_x}\left(\frac{1}{n}\sum_{i=1}^n x_i^2 + x_o^2 - 2x_0\bar{x} + \bar{x}^2 - \bar{x}^2\right) =$$

$$\frac{\sigma^2}{SS_x}\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n} n\bar{x}^2 + (x_o^2 - 2x_0\bar{x} + \bar{x}^2)\right) =$$

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

R^2 e previsione VII

$$\frac{\sigma^2}{SS_x}\left(\frac{1}{n}SS_x + (x_o - \bar{x})^2\right) = \sigma^2\left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x}\right)$$

che ci dice che questo aumenta se x_o è distante dalla media campionaria.

- **Prevedo il valore di** y_o con

$$y_o = \hat{\alpha} + \hat{\beta}x_o + \epsilon_o.$$

In questo caso la "prevision" non è un previsione nel vero senso della parola, perchè ϵ_o non lo conosco, ma posso dire che y_o è una normale con valore medio uguale alla migliore stima che ho, cioè $\hat{\alpha} + \hat{\beta}x_o$.
Essendo combinazione lienare di normali è ancora normale con media

$$E(y_0) = E(\hat{\alpha} + \hat{\beta}x_o + \epsilon_o) = \alpha + \beta x_o$$

e varianza

$$V(y_0) = V(\hat{\alpha} + \hat{\beta}x_o + \epsilon_o) = V(\hat{\alpha} + \hat{\beta}x_o) + V(\epsilon_o)$$

R^2 e previsione VIII

non devo considerare la covarianza perchè abbiamo assunto che le ϵ sono indipendenti, e quindi $\hat{\alpha} + \hat{\beta} x_o$ è indipendente da ϵ_0. QUindi

$$V(y_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x} \right) + \sigma^2 = \sigma^2 \left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x} + 1 \right)$$

Visto che le previsioni sono variabili aleatorie, posso fare anche dei test del tipo

$$H_0 : \mu_0 = c \qquad H_1 : \mu_0 \neq c$$

che mi porta all'usuale intervallo di confidenza

$$\left[\hat{y}_0 - t_{n-2,\alpha/2} \sqrt{S^2 \left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x} \right)}, \hat{y}_0 + t_{n-2,\alpha/2} \sqrt{S^2 \left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x} \right)} \right]$$

con $\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_o$

R^2 e previsione IX

Se non assumiamo σ^2 noto, dobbiamo tener conto che anche questa è una variabile aleatoria. Si può fare, ma i calcoli diventano molto più complessi.

Oppure

$$H_0 : Y_o = c \qquad H_1 : Y_o \neq c$$

che mi porta all'usuale intervallo di confidenza

$$\left[\hat{y}_0 - t_{n-2,\alpha/2}\sqrt{S^2\left(1+\frac{1}{n}+\frac{(x_o-\bar{x})^2}{SS_x}\right)}, \hat{y}_0 + t_{n-2,\alpha/2}\sqrt{S^2\left(1+\frac{1}{n}+\frac{(x_o-\bar{x})^2}{SS_x}\right)}\right]$$

con $\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_o$

Regressione Multivariata I

nei casi reali, abbiamo sempre più di una x con cui vogliamo comprendere la y. Nel caso più generale, assumiamo

$$y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \cdots + \beta_p x_{i,p} + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$, iid. Lo stesso modello sopra si può scrivere in maniera compatta come

$$y_i = \sum_{i=1}^{p} \beta_i x_{i,p} + \epsilon_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$$

con $x_{i,1} = 1$, $\mathbf{x}_i = (1, x_{i,2}, x_{i,3}, \ldots, x_{i,p})$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.
Per avere senso e poter avere risultati interpretabili, dobbiamo avere $n >> p$, altrimenti diventa un problema "matematico" e non statistico. Naturalmente non avremo più una retta di regressione, ma un iperpiano.
anche in questo caso valgono le ipotesi di Gauss

- $E(\epsilon_i) = 0$

Regressione Multivariata II

- $Var(\epsilon_i) = \sigma^2$, tutto uguali, e si di che sono **omoschedastiche**
- $Cov(\epsilon_i, \epsilon_j) = 0$, se $j \neq i$

a cui va aggiunta anche un'altra, come nella univariata

$$E(\epsilon_i | x_{i,1}, \ldots, x_{i,p}) = 0$$

e

$$Var(\epsilon_i | x_{i,1}, \ldots, x_{i,p}) = \sigma^2$$

Nella regressione multivariata, è più facile lavorare con le matrici. Assumiamo che $\mathbf{X}$ sia una matrice $n \times p$, con riga i-esima pari a $\mathbf{x}_i$, e $\mathbf{y} = (y_1, \ldots, y_n)'$ e $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$, allora

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Regressione Multivariata III

oppure

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

L'interpretazione dei parametri è molto simile, visto che la media di y_i aumenta di β_j se aumento di 1 il valore di x_j

$$E(y_i) = \beta_1 + \beta_2 x_{i,2} + \beta_2 x_{i,3} + \cdots + \beta_p x_{i,p}$$

e ipotizziamo che $x_{i',2} = x_{i,2} + 1$, e $x_{i',p} = x_{i,p}$, per $p \neq 2$

$$E(y_{i'}) = \beta_1 + \beta_2(x_{i,2} + 1) + \beta_3 x_{i,3} + \cdots + \beta_p x_{i,p} = E(y_i) + \beta_2$$

Passiamo adesso al problema della stima dei parametri, che facciamo massimizzando la verosimiglianza:

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right)$$

Regressione Multivariata IV

Come sempre, lo possiamo prima trovare lo stimatore di $\boldsymbol{\beta}$ visto che è valido per qualsiasi valore di σ^2. Usiamo la log verosimiglianza

$$logL(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = -\frac{n}{2}log(2\pi\sigma^2) - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

che equivale a minimizzare

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$$

Per trovare il minimo calcoliamo la derivata e poniamola uguale a zero

$$\frac{dlogL(\boldsymbol{\beta}, \sigma^2|\mathbf{y})}{d\boldsymbol{\beta}} = 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y} = 0$$

che risolto per $\boldsymbol{\beta}$ mi da

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Regressione Multivariata V

che possiamo facilmente vedere che è un minimo perchè la derivata seconda è

$$2\mathbf{X}'\mathbf{X}$$

che è una matrice simmetrica e semidefinita positiva.
Una volta che abbiamo $\hat{\boldsymbol{\beta}}$, lo stimtore di massima verosimiglianza di σ^2 è

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2}{n}$$

Attenzione Per poter invertire $\mathbf{X}'\mathbf{X}$, il rango di $\mathbf{X}$ deve essere p, i.e., le colonne di $\mathbf{X}$ devono essere linearmente indipendenti. Da un punto di vista matematico basta questo, ma se le colonne di $\mathbf{X}$ sono linearmente indipendenti, ma la correlazione è comunque alta $Cor([\mathbf{X}]_{.j}, [\mathbf{X}]_{.j'})$, dove la correlazione qui va intesa come correlazione campionaria

Regressione Multivariata VI

(visto che le x non sono variabili aleatorie), allora le stime sono **instabili**, e non interpretabili (questo problema si chiama **multicollinearità**).

Cerchiamo di capire cosa succede se cè multicollinearità. Nello script R cè un esempio in sono state prese delle persone che faceva jogging al parco, e misurate alcune "qualità". Noi siamo interessati a valutare il valore dell'ossigenazioni rispetto alle pulsazioni massime durante la corsa, e le pulsazioni medie. facciamo un grafico delle variabili di interesse

Regressione Multivariata VII

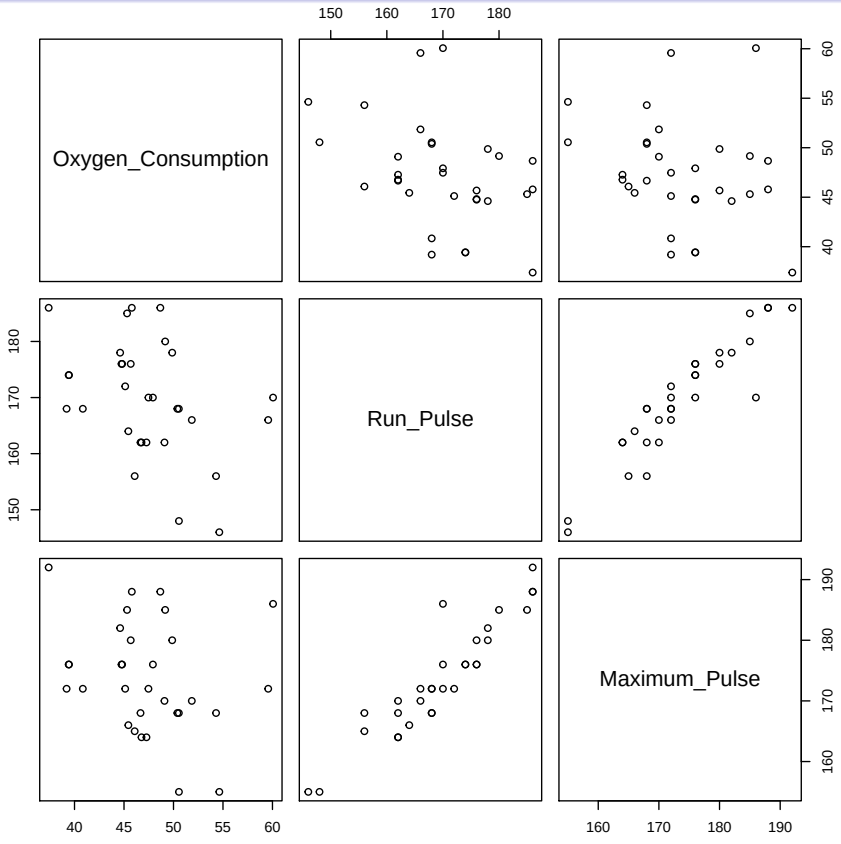

Figure: regressione multipla

Regressione Multivariata VIII

Dalla figura è chiaro come le due covariate (le pulsazioni) non siano linearmente dipendenti, ma neanche completamente indipendenti (in termini di correlazione). Vediamo cosa succede se proviamo a ottenere le stime:

```
reg = lm(Oxygen_Consumption~Run_Pulse+Maximum_Pulse,data = data )# comando
summary(reg)# comando

Call:
lm(formula = Oxygen_Consumption ~ Run_Pulse + Maximum_Pulse,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2523 -1.5567  0.3653  2.2084 10.7222

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.7223    16.3055   3.908 0.000537 ***
Run_Pulse     -0.6822     0.2248  -3.034 0.005159 **
Maximum_Pulse  0.5719     0.2515   2.274 0.030837 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regressione Multivariata IX

```
Residual standard error: 4.648 on 28 degrees of freedom
Multiple R-squared:  0.2896,    Adjusted R-squared:  0.2389
F-statistic: 5.708 on 2 and 28 DF,  p-value: 0.008331
```

Sembrerebbe che le pulsazioni medie facciano diminuire l'ossigenazione, ma le massime la fanno aumentare, anche se la due rappresentano la stessa cosa (sono molto correlate), questo è perchè la matrice $\mathbf{XX}'$ è "quasi" singolare. Quindi, per avere risultati robusti e che hanno senso, dovremmo usare solo variabili indipendenti tra loro. Se faccio dure regresisoni semplici, vedo che i risultati delle due variabiloi sono in accordo

```
Call:
lm(formula = Oxygen_Consumption ~ Run_Pulse, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5161 -2.2374 -0.3812  2.7489 12.7576

Coefficients:
```

Regressione Multivariata X

```
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 82.47111   15.04463   5.482 0.00000665 ***
Run_Pulse   -0.20687    0.08853  -2.337     0.0266 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.971 on 29 degrees of freedom
Multiple R-squared:  0.1585,    Adjusted R-squared:  0.1294
F-statistic: 5.461 on 1 and 29 DF,  p-value: 0.02656


Call:
lm(formula = Oxygen_Consumption ~ Maximum_Pulse, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4200 -2.4097 -0.5206  2.8109 14.3671

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    71.2959    18.2515   3.906 0.000516 ***
Maximum_Pulse  -0.1377     0.1049  -1.312 0.199696
---
```

Regressione Multivariata XI

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.265 on 29 degrees of freedom
Multiple R-squared:  0.05606,   Adjusted R-squared:  0.02351
F-statistic: 1.722 on 1 and 29 DF,  p-value: 0.1997
```

Regressione Multivariata XII

Come con il caso univariato, noi vogliamo fare inferenza e per questo dobbiamo trovare la distribuzione di

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

che è combinazione di normali con media

$$E(\hat{\boldsymbol{\beta}}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

quindi è non distorto, e

$$Var(\hat{\boldsymbol{\beta}}) = Var((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} =$$

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Quindi è congiuntamente normale $N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, e posso fare qualsiasi test del tipo

$$H_0 : \beta_j = c \qquad H_1 : \beta_j \neq c$$

Regressione Multivariata XIII

visto che

$$\hat{\beta}_j \sim N([\boldsymbol{\beta}]_j, [\sigma^2(\mathbf{X}'\mathbf{X})^{-1})]_{jj})$$

con l'usuale test T, tenendo da conto però che la variabile aleatora

$$\frac{(n-p)S^2}{\sigma^2} \sim \chi_{n-p}$$

cioè un chi quadrato con un numero di gradi di libertà n meno il numero dei regressori. La statistica test è allora

$$t = \frac{\hat{\beta}_j - c}{\sqrt{[S^2(\mathbf{X}'\mathbf{X})^{-1})]_{jj}}} \sim t_{n-p}$$

Regressione Multivariata XIV

Possiamo vedere anche se lo stimatore $\hat{\boldsymbol{\beta}}$ raggiunge il limite di Cramer Rao (considerando σ^2 fisso per semplicità, anche perchè sappiamo che è distorto) nel caso multivariato, e per stimatori non distorti, abbiamo che

$$Var(W(\mathbf{Y})) \geq \mathcal{I}_n^{-1}$$

dove $\mathcal{I}_n$ è l'informazione di Fisher, con

$$[\mathcal{I}_n]_{ij} = E\left(\frac{d}{d\beta_i} log f(\mathbf{Y}|\boldsymbol{\beta}) \frac{d}{d\beta_j} log f(\mathbf{Y}|\boldsymbol{\beta})\right)$$

Il cui calcolo è complesso e laborioso, ma, sotto le stesse condizioni di Cramer Rao univariato, possiamo calcolare $\mathcal{I}_n$ usando le derivate secondo

$$[\mathcal{I}_n]_{ij} = -E\left(\frac{d^2}{d\beta_i \beta_j} log f(\mathbf{Y}|\boldsymbol{\beta})\right)$$

che ci porta a scrivere

$$\mathcal{I}_n = -E\left(\frac{d}{d\boldsymbol{\beta}^2} log f(\mathbf{Y}|\boldsymbol{\beta})\right)$$

Regressione Multivariata XV

Quando abbiamo calcolato lo stimatore di massima verosimiglianza, avevamo visto che era un massimo calcolando proprio la derivata seconda di

$$\frac{d^2}{d\boldsymbol{\beta}^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = 2\mathbf{X}'\mathbf{X}$$

da cui possiamo ricavare che

$$\frac{d}{d\boldsymbol{\beta}^2} log f(\mathbf{Y}|\boldsymbol{\beta}) = \frac{d}{d\boldsymbol{\beta}^2}\left((2\pi\sigma)^{-n/2} - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right) = \frac{\mathbf{X}'\mathbf{X}}{\sigma^2}$$

da cui abbiamo il limite di Cramer Rao che è

$$Var(W(\mathbf{Y})) \geq \mathcal{I}_n^{-1} = \left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

esattamente la varianza dello stimatore di massima verosimiglianza.

Regressione Multivariata XVI

Se volessi fare previsione, assumendo σ^2 nota, sarei interessato alla media

$$\mu_o = \mathbf{x}_o\boldsymbol{\beta}$$

di un'osservazione futura, con vettore di covariate $\mathbf{x}_o$, o al suo valore

$$y_0 = \mathbf{x}_0\boldsymbol{\beta} + \epsilon_o$$

Nel caso della media posso stimarla con

$$\hat{\mu}_o = \mathbf{x}_o\hat{\boldsymbol{\beta}}$$

che è normal con media

$$E(\hat{\mu}_o) = \mathbf{x}_o E(\hat{\boldsymbol{\beta}}) = \mathbf{x}_o\boldsymbol{\beta}$$

e varianza

$$Var(\hat{\mu}_o) = \mathbf{x}_o Var(\hat{\boldsymbol{\beta}})\mathbf{x}_o' = \sigma^2\mathbf{x}_o(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o'$$

Regressione Multivariata XVII

Se voglio vedere la miglior stima che ho di y_0, posso usare

$$\mathbf{x}_0\hat{\boldsymbol{\beta}} + \epsilon_o$$

che ha medie e varianze pari a

$$E(\mathbf{x}_0\hat{\boldsymbol{\beta}} + \epsilon_o) = \mathbf{x}_0\boldsymbol{\beta}$$

e

$$Var(\mathbf{x}_0\hat{\boldsymbol{\beta}} + \epsilon_o) = \sigma^2\mathbf{x}_o(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o' + \sigma^2$$

Nel caso multivariato, l'R^2 si calcola allo stesso modo

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Regressione Multivariata XVIII

dove

$$\hat{y}_i = \boldsymbol{x}_i \hat{\boldsymbol{\beta}}$$

visto che si può dimostrare che

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

ma è molto complesso.

Verifica delle Ipotesi I

Un punto che non abbiamo mai toccato è la verifica delle ipotesi. Il modello si basa su alcuni assunti fondamentali

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = \sigma^2$, tutto uguali, e si di che sono **omoschedastiche**
- $Cov(\epsilon_i, \epsilon_j) = 0$, se $j \neq i$

più la normalita di ϵ Tutta la costruzione che abbiamo fatto, ha senso se queste ipotesi sono verificate, altrimenti no. Vedere che la prima è verificata sempre è facile, nel senso che la stima dei residui ha media campionaria uguale a zero. Riprendiamo la funzione che abbiamo minimizzato per trovare le stime di $\boldsymbol{\beta}$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_p x_{i,p})^2 = \sum_{i=1}^{n}\Big((y_i - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_p x_{i,p}) - \hat{\beta}_1\Big)^2 =$$

$$\sum_{i=1}^{n}\Big(z_i - \hat{\beta}_1\Big)^2$$

Verifica delle Ipotesi II

con $z_i = y_i - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_p x_{i,p}$. Per qualsiasi valore di z_i, con $i = 1, \ldots, n$, il minimo di quella funzione si raggiunge con $\hat{\beta}_1 = \bar{z}$, che è quindi lo stimatore di massima verosimiglianza di β_1.
Abbiamo che la stima dei residui è

$$\hat{\epsilon}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i,2} - \cdots - \hat{\beta}_p x_{i,p} = z_i - \hat{\beta}_1 = z_i - \bar{z}$$

quindi la media campionaria dei residui è

$$\sum_{i=1}^{n} \frac{\hat{\epsilon}_i}{n} = \sum_{i=1}^{n} \frac{z_i - \bar{z}}{n} = 0$$

che è sempre zero, ma siolo in un modello **che ha l'intercetta**.

Per le altre ipotesi, i.e., normalità, omoschedasticità, indipendenza tra le ϵ, e tra ϵ e $[\mathbf{X}]_{.j}$, si potrebbero fare dei test, ma si può fare anche visivamente.

Verifica delle Ipotesi III

Facciamo un esempio. Abbiamo un dataset (vedere il codice R), in cui analizziamo le vendite in base a quando abbiamo speso in pubblicità su youtube, facebook, e newspaper. Vediamo dei plot

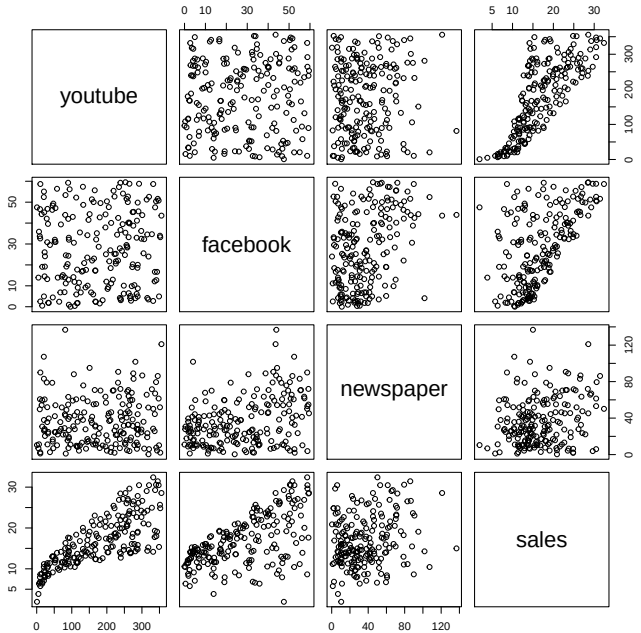

Figure: Regressione

Verifica delle Ipotesi IV

Proviamo a fare un modello in cui spieghiamo sales con youtube.

```
model = lm(sales ~ youtube, data = marketing)# comandi
summary(model) # comandi
Call:
lm(formula = sales ~ youtube, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0632  -2.3454  -0.2295   2.4805   8.6548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.439112   0.549412   15.36   <2e-16 ***
youtube     0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.91 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Verifica delle Ipotesi V

Che ha un R^2 molto alto. Vediamo se i residui sono normali. Vediamo se i residui sono normali

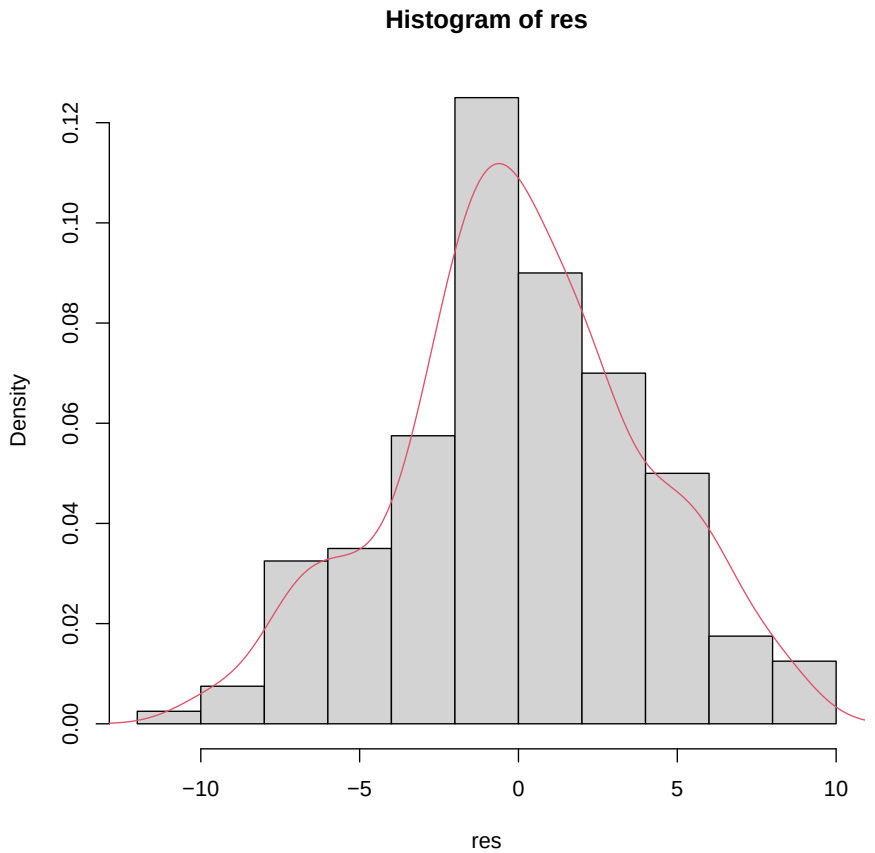

Figure: Regressione

Verifica delle Ipotesi VI

sembrerebbero essere normali (o almeno è palusibile). E vediamo se sono omoschedastiche

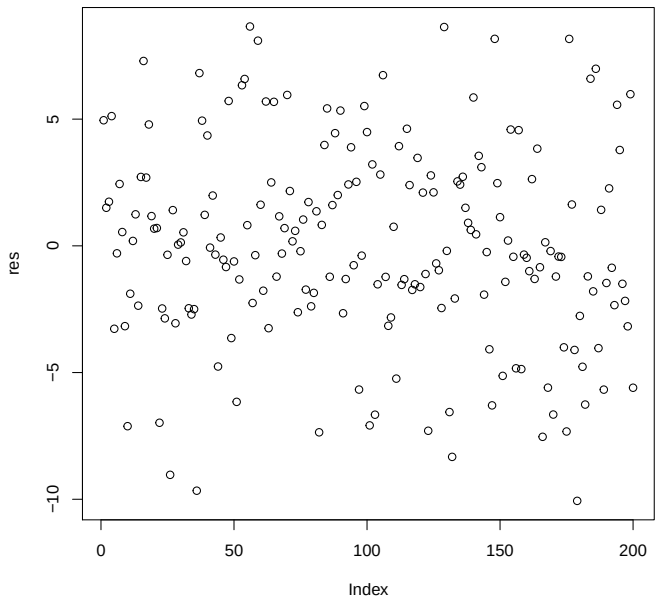

Figure: Regressione

sembra che la varianza sia costante. Adesso plottiamo i residui (asse y) rispetto alla variabile usata nel modello (asse x)

Verifica delle Ipotesi VII

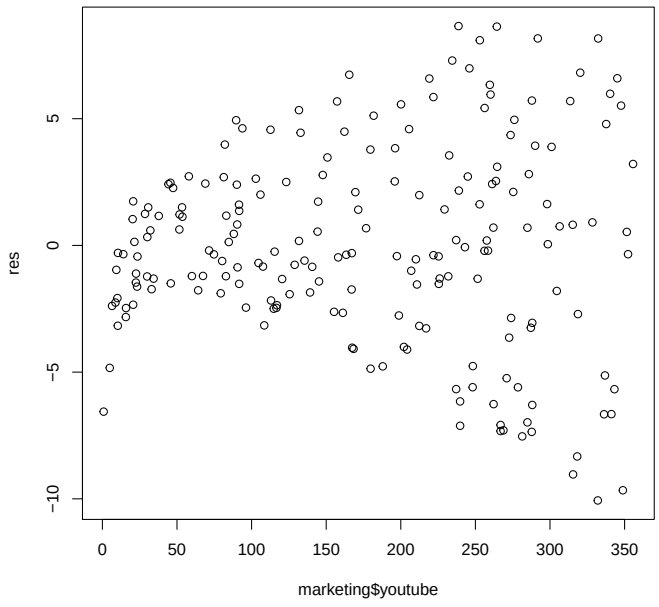

Figure: Regressione

qui vediamo dei problemi. Ci sono dei pattern, il che significa che l'errore è dipendente dalla x. In più, ci sono punti di x in cui la varianza è più piccola di altri, i.e., non sono omoschedastiche. vediamo iresidui rispetto alle variabili non messe nel modello

Verifica delle Ipotesi VIII

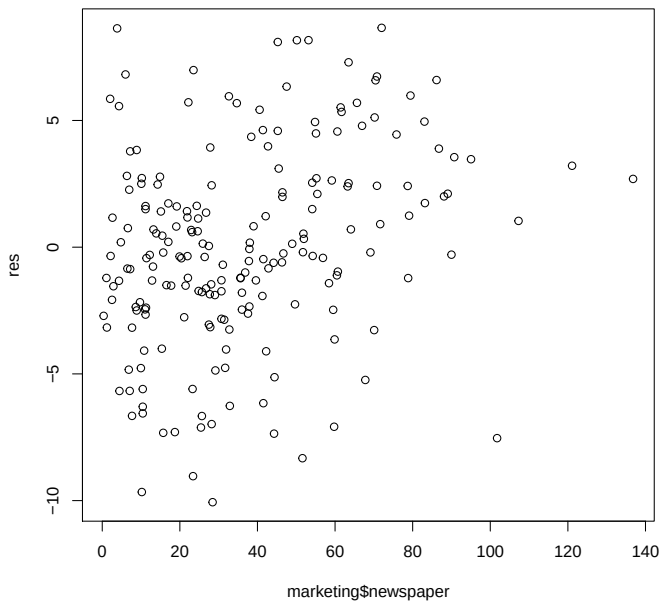

Figure: Regressione

Verifica delle Ipotesi IX

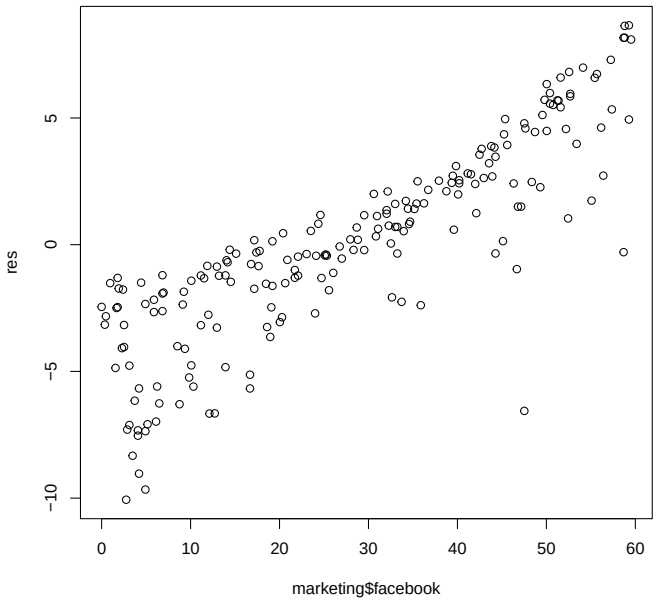

Figure: Regressione

Verifica delle Ipotesi X

Sembra che ci sia dipendenza tra i residui e queste variabili, e questo ci dice che se mettessimo queste variabili nel modello, avrebbero un impatto positivo.

Per concludere, siccono i residui sono eteroschedastiche, e forse dipendenti, la significatività dei test su β non è affidabile. LA regressione va rivista, magari mettendo altre x, oppure mettere delle trasformazioni di x.