APPENDICE D
ALGORITMI PER IL CALCOLO DELLA REGRESSIONE LINEARE

Per adattare la retta ai dati sperimentali viene impiegato il metodo dei minimi quadrati, una tecnica di approssimazione ben nota, che consente di minimizzare la somma dei quadrati delle differenze che residuano fra i punti sperimentali e la retta.

Regressione lineare x variabile indipendente

Il modello matematico impiegato presuppone che la *x*, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la *y* sia distribuito normalmente e con una varianza che non cambia al variare del valore della *x*.

Sia allora *n* il numero dei punti aventi coordinate note (x_i,y_i), e siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i: il coefficiente angolare b_{yx} e l'intercetta a_{yx} dell'equazione della retta di regressione *x* variabile indipendente

$$y = a_{yx} + b_{yx} \cdot x$$

che meglio approssima (in termini di minimi quadrati) i dati vengono calcolati rispettivamente come

$$b_{yx} = \Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \Sigma(x_i - \bar{x})2$$

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

Il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2)}$$

E' possibile semplificare i calcoli ricordando che

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

La varianza residua attorno alla regressione viene calcolata come

$$s_0^2 = (\Sigma(y_i - \bar{y})^2 - s_1^2) / (n - 2)$$

essendo

$$s_1^2 = (\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}))^2 / \Sigma(x_i - \bar{x})^2$$

Infine l'errore standard della stima s_{yx} e le deviazioni standard del coefficiente angolare (s_b) e dell'intercetta (s_a), che forniscono una misura rispettivamente della dispersione dei dati attorno alla

retta calcolata, e del grado di incertezza connesso con i valori ottenuti di a_{yx} e di b_{yx}, sono calcolati come

$$s_{yx} = \sqrt{(s_0^2)}$$
$$s_b = s_{yx} \cdot \sqrt{(1 / \Sigma (x_i - \bar{x})^2)}$$
$$s_a = s_b \cdot \sqrt{(\Sigma x_i^2 / n)}$$

Si consideri che la retta di regressione campionaria

$$y = a_{yx} + b_{yx} \cdot x$$

rappresenta la migliore stima possibile della retta di regressione della popolazione

$$y = \alpha + \beta \cdot x$$

Si consideri che il test t di Student per una media teorica nella forma già vista

$$t = (\bar{x} - \mu) / \sqrt{(s^2 / n)}$$

può essere riscritto tenendo conto delle seguenti identità

$$\bar{x} = a_{yx}$$
$$\mu = \alpha$$
$$\sqrt{(s^2 / n)} = s_a$$

assumendo quindi la forma

$$t = (a_{yx} - \alpha) / s_a$$

Questo consente di sottoporre a test la differenza dell'intercetta *a* rispetto a un valore atteso (per esempio rispetto a 0, cioè all'intercetta di una retta passante per l'origine). Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza dell'intercetta rispetto al valore atteso.

Si consideri che il test t di Student per una media teorica può anche essere riscritto tenendo conto delle seguenti identità

$$\bar{x} = b_{yx}$$
$$\mu = \beta$$
$$\sqrt{(s^2 / n)} = s_b$$

assumendo quindi la forma

$$t = (b_{yx} - \beta) / s_b$$

Questo consente di sottoporre a test la differenza del coefficiente angolare *b* rispetto a un valore atteso (per esempio rispetto a 0, cioè al coefficiente angolare di una retta orizzontale, oppure rispetto a 1, cioè al coefficiente angolare di una retta a 45 gradi). Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella

effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza del coefficiente angolare rispetto al valore atteso.

Regressione lineare y variabile indipendente

Il modello matematico impiegato presuppone che la *y*, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la *x* sia distribuito normalmente e con una varianza che non cambia al variare del valore della *y*. Si noti che in questo caso inizialmente la *y* (variabile indipendente) viene posta in ascisse e la *x* (variabile dipendente) viene posta in ordinate.

Sia allora *n* il numero dei punti aventi coordinate note (x_i,y_i), e siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i: il coefficiente angolare b_{xy} e l'intercetta a_{xy} dell'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

che meglio approssima (in termini di minimi quadrati) i dati vengono calcolati rispettivamente come

$$b_{xy} = \Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \Sigma(y_i - \bar{y})^2$$

$$a_{xy} = \bar{x} - b_{xy} \cdot \bar{y}$$

Il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2)}$$

(si noti che, come atteso, esso risulta identico a quello calcolato mediante la regressione *x* variabile indipendente).

E' possibile semplificare i calcoli ricordando che

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma(x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

Per riportare i dati sullo stesso sistema di coordinate cartesiane utilizzato per la regressione *x* variabile indipendente, si esplicita l'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

rispetto alla *y*, ottenendo

$$x - a_{xy} = b_{xy} \cdot y$$

e quindi, dividendo entrambi i membri per b_{xy}

$$y = - a_{xy} / b_{xy} + 1 / b_{xy} \cdot x$$

Quindi l'intercetta a e il coefficiente angolare b che consentono di rappresentare la regressione y variabile indipendente sullo stesso sistema di coordinate cartesiane della regressione x variabile indipendente saranno rispettivamente uguali a

$$a = - a_{xy} / b_{xy}$$
$$b = 1 / b_{xy}$$

Componente principale standardizzata

Il modello matematico impiegato presuppone tanto la x quanto la y siano affette da un errore di misura equivalente.

Sia allora n il numero dei punti aventi coordinate note (x_i,y_i), siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i, sia b_{yx} il coefficiente angolare dell'equazione della retta di regressione x variabile indipendente, e sia b_{xy} il coefficiente angolare dell'equazione della retta di regressione y variabile indipendente.

Il coefficiente angolare b_{cps} dell'equazione della retta di regressione calcolata come componente principale standardizzata è allora uguale a

$$b_{cps} = \sqrt{(b_{yx} \cdot b_{xy})}$$

cioè alla media geometrica tra il coefficiente angolare b_{yx} della regressione x variabile indipendente e il coefficiente angolare b_{xy} della regressione y variabile indipendente, cioè

$$b_{xy} = \sqrt{(\Sigma (y_i - \bar{y})^2 / \Sigma (x_i - \bar{x})^2)}$$

mentre l'intercetta a_{cps} dell'equazione della retta di regressione calcolata come componente principale standardizzata è uguale a

$$a_{cps} = \bar{y} - b_{cps} \cdot \bar{x}$$

Infine il coefficiente di correlazione r viene calcolato come

$$r = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2)}$$

(si noti che, come atteso, esso risulta identico sia a quello calcolato mediante la regressione x variabile indipendente sia a quello calcolato mediante la regressione y variabile indipendente).

E' possibile semplificare i calcoli ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$