LEZIONE n.7 (a cura di Michael Fop)

E' possibile esprimere l'indice r^2 pure in un' altra forma.
Considerando infatti la **scomposizione della Devianza Totale** Dev(Y) :

$$\text{Dev(Y)} = \sum_i \left(Y_i - \bar{Y}\right)^2 = \sum_i \left[\left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)\right]^2 =$$
$$= \sum_i \left[\left(Y_i - \hat{Y}_i\right)^2 + \left(\hat{Y}_i - \bar{Y}\right)^2\right] + 2\sum_i \left[\left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)\right]$$

dove $2\sum_i \left[\left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)\right] = 0$

per cui
$$\text{Dev(Y)} = \sum_i \left(Y_i - \hat{Y}_i\right)^2 + \sum_i \left(\hat{Y}_i - \bar{Y}\right)^2$$

DEVIANZA RESIDUA (RSS) — **DEVIANZA SPIEGATA** (ESS)

RSS = *Residual Sum of Squares*
ESS = *Explained Sum of Squares*

ESS è una misura della variabilità spiegata dal modello di regressione, mentre RSS èla devianza dei residui e rappresenta una misura della variabilità non spiegata (perciò “residua”) dal modello. Pertanto RSS è una misura dell' errore di previsione legato all' utilizzo del modello, più questa è elevata peggiore è la qulità del modello.
Infatti un buon modello dovrebbe avere:

RSS ⟶ 0 , per avere una possibilità nulla di sbagliare utilizzando il modello

ESS ⟶ Dev(Y), a fronte del fatto che 0 < ESS < Dev(Y); affinché il modello sia in grado di inglobare tutta la variabilità di Y

Si può pertanto scrivere r^2 come :

$$r^2 = 1 - \frac{RSS}{Dev(Y)} = \frac{ESS}{Dev(Y)}$$

per cui :
_ se **ESS = 0** allora **r^2 = 0** , perciò la retta stimata è parallela all' asse delle ascisse. In questo caso la retta di regressione non ha alcuna capacità esplicativa nei confronti di Y
_ se **ESS = Dev(Y)** , allora **r^2 = 1** e risulta **RSS = 0,** pertanto il modello si adatta perfettamente ai dati osservati e tutti i valori stimati per Y si pongono sulla retta di regressione

Un ulteriore modo per esprimere r^2 è

$$r^2 = \frac{\left[Cod(X,Y)\right]^2}{Dev(X)\,Dev(Y)} = \hat{\beta}_1 \frac{Cod(X,Y)}{Dev(Y)} \qquad \text{dato che} \qquad \beta_1 = \frac{Cod(X,Y)}{Dev(X)}$$

pertanto si ha che $r^2 = 0$ soltanto se $\beta_1 = 0$, ovvero se X e Y sono incorrelate.

La bontà dell' accostamento della retta di regressione ai dati può essere pure interpretata tramite **tabella ANOVA** (*Analysis Of Variance*), verificando in tale contesto l'effettiva capacità della variabile eplicativa X nello spiegare la variabilità di Y

Dev	**Gradi di Libertà**		**F**
$ESS = \sum_i \left(\hat{Y}_i - \bar{Y}\right)^2$	1	$S.^2 = ESS$	$\frac{ESS}{S^2}$
$RSS = \sum_i \left(Y_i - \hat{Y}_i\right)^2$	n - 2	$S^2 = \frac{RSS}{n-2}$	

$$Dev(Y) = \sum_i \left(Y_i - \bar{Y}\right)^2$$

In particolare si denota cheil rapporto F è una misura di significatività per l'intera regeressione.
Il test – F in questo caso è infatti utile a controllare la validità del modello, in quanto con questo si vuole verificare idealmente l'ipotesi nulla “il modello **NON** è esplicativo” contro l'alternativa che sia esplicativo per la variabile Y.
Difatti:

sotto H0 : $\beta_1 = 0$ il rapporto F ~ F(1, n – 2) e si ha che se
$F < F_\alpha \Rightarrow si\ accetta\ H_0$ che vale pure se $p - value(F) \geq \alpha = 0{,}05$

Pertanto se il valore F osservato è abbastanza elevato vuol dire che l' ipotesi nulla è da rifiutare e la variabile esplicativa X ha influenza sulla Y, perciò il modello è utile a “spiegare” il fenomeno Y tramite X.
Se invece si ottiene un F osservato nella zona di non rifiuto di H0 (perciò si ha il non rifiuto di $\beta_1 = 0$) si è per ritenere che il modello non sia utile a stimare la relazione esistente tra X e Y.

Per verificare quindi la qualità del modello occorre considerare :
_ $r^2 \geq 0{,}75$ per una elevata Devianza Spiegata dal modello
_ un F osservato elevato (e un p-value(F) minore del livello di significatività prefissato), perché sia verosimile l'ipotesi di dipendenza di Y da X per poter utilizzare il modello

Infine nel modello di regressione semplice con una sola variabile esplicativa il test -F è equivalente al test-t sul coefficiente angolare, pertanto forniscono entrambi lo stesso p-value.

$t = \frac{\hat{\beta}_1}{SE[\hat{\beta}_1]}$ sotto H0 : $\beta_1 = 0$ p-value(F) = p-value(t)

Il modello cosi costruito consente di effettuare previsioni sulla variabile Y, laddove per “previsione” s'intende assegnare valori plausibili ad Y sulla base di nuovi valori di X :

$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ dove $\hat{Y}_0$ e x_0 sono il valore previsto di Y e il nuovo valore di X.

Si ha pertanto che $\hat{Y}_0$ è una variabile casuale in quanto dipende dai dati osservati, difatti cambiando questi cambia pure il valore previsto di Y.

In particolare :

$\hat{Y}_0 \sim N(\mu_0, \sigma_0)$ dove

$$\mu_0 = \beta_0 + \beta_1 x_0$$

$$\sigma_0^2 = \sigma^2 [\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Dev(X)}]$$

Gli **intervalli di confidenza** per $\hat{Y}_0$ sono :

_ se σ^2 è nota

$$\hat{Y}_0 - z_{\alpha/2}\sqrt{\sigma_0^2} \; ; \; \hat{Y}_0 + z_{\alpha/2}\sqrt{\sigma_0^2}$$

_ se σ^2 è incognita

dove

$\hat{\sigma}_0^2 = S^2 [\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Dev(X)}]$ e $t_{\alpha/2}$ è il quantile di una t ~ (n – 2)

In realtà con questo metodo la stima di $\hat{Y}_0$ è una stima sul valore atteso della popolazione Y e non sul valore relativo allo specifico elemento y della stessa.

La previsione si effettua infatti con la relazione

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

la quale è derivante dal modello, che per costruzione è definito da

$$E[Y \, dato \, x_0] = \beta_0 + \beta_1 x_0$$

per questo motivo $\hat{Y}_0$ è una stima del valore medio di Y.

Da ciò deriva pertanto che la stima effettuata sul singolo elemento Y_0 della distribuzione di Y è data dal valore medio di Y più una componente di errore ε

$$Y_0 = E[Y \, dato \, x_0] + \varepsilon_0$$

In questo caso per il singolo elemento Y_0 si parla di **"intervallo predittivo"** e occorre considerare pure il termine di errore ε :

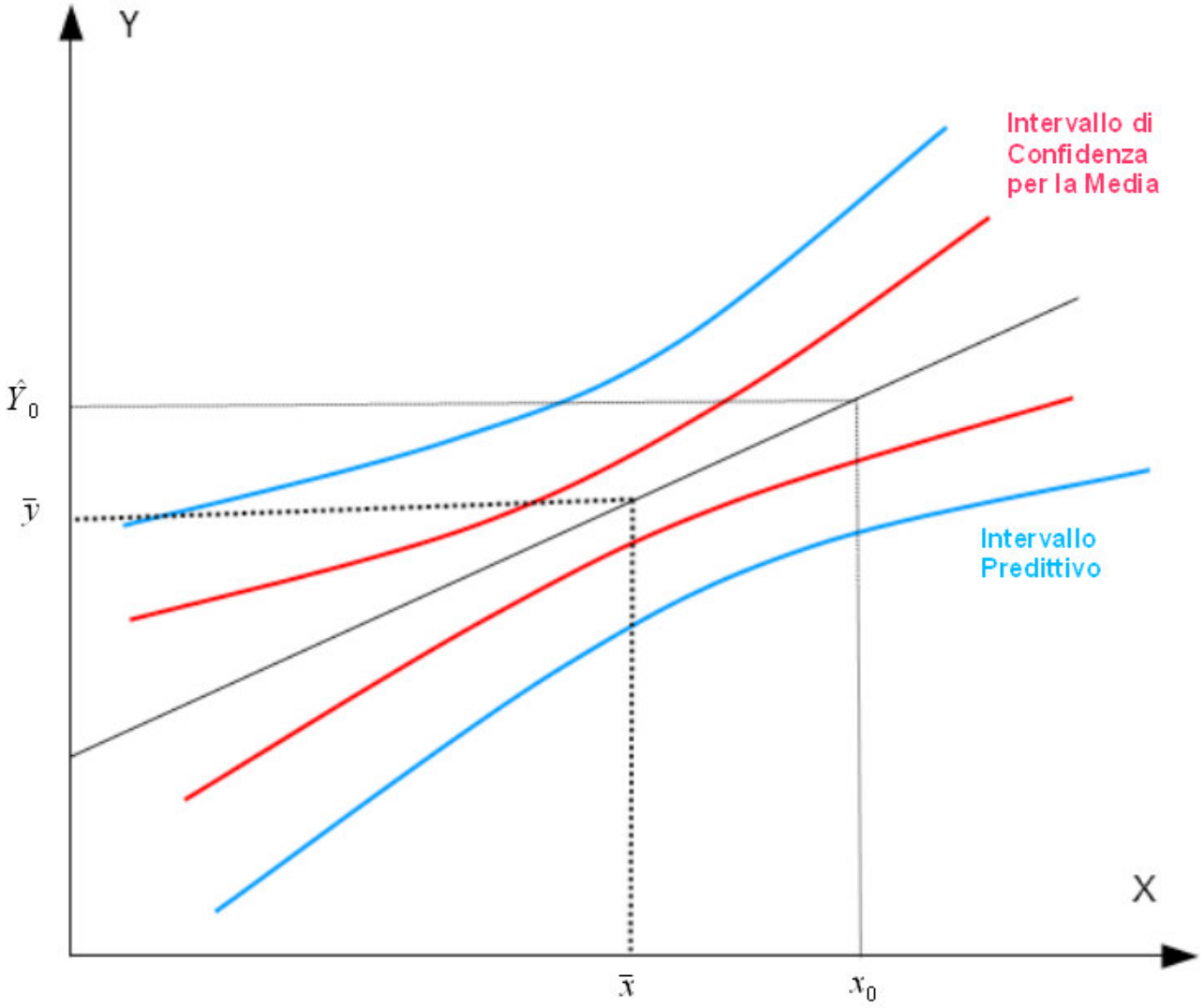
$$\hat{Y} - t_{\alpha/2}\sqrt{\tilde{\sigma}_0^2} \; ; \; \hat{Y} + t_{\alpha/2}\sqrt{\tilde{\sigma}_0^2}$$

dove

$$\tilde{\sigma}_0^2 = S^2 [1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Dev(X)}]$$

Si tiene conto del termine di errore ε che ha varianza uguale a S^2: l'intervallo predittivo per il singolo elemento è più ampio dell'intervallo di confidenza per la media

Graficamente :



Gli intervalli sono più stretti attorno ai valori medi perchè la media è il valore che minimizza l'ampiezza dell' intervallo.
Intuitivamente perchè intorno alle medie c'è “meno incertezza” sul valore previsto.