Regressione Multipla

Regressione Multipla

Regressione Multipla

- Il modello di regressione multipla è una semplice generalizzazione del modello semplice già visto
- In esso si assume che una variabile sia spiegata linearmente da un certo numero di altre variabili, ovvero

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_2 x_{i2} + \varepsilon_i$$

Regressione Multipla

Per valutare il legame lineare complessivo tra le variabili si usano diversi modi

- Coefficiente di correlazione multiplo
- Matrice di correlazione
- Diagrammi a dispersione a coppie

Regressione Multipla: Coefficiente di correlazione multipla

Vediamo un esempio di calcolo del coefficiente di correlazione regressione multipla nel caso di due variabili indipendenti. Abbiamo y variabile dipendente e x_1, x_2 variabili indipendenti. Indichiamo con $r_{yx_i}, i = 1, 2$ il coefficiente di correlazione tra y ed x_i e con $r_{x_1x_2}$ quello tra le variabili indipendenti. Allora il coefficiente di correlazione multipla è dato da

$$R = \frac{\sqrt{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_1}r_{yx_2}}}{\sqrt{1 - r_{x_1x_2}}}$$

Questa statistica misura l'intensità del legame lineare tra le tre variabili considerate

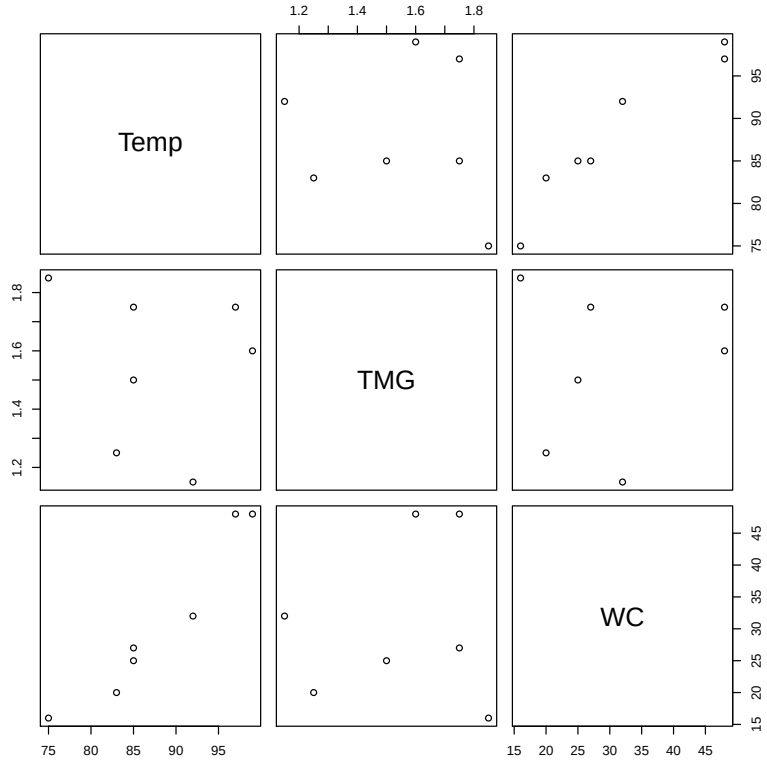
Regressione Multipla: Esempio I

Si rilevano le temperature in Farenheit in 7 giorni dell'estate per una persona che su di un periodo di tre ore registra anche quanto tempo trascorre falciando il prato e quante volte beve acqua.

Temperature (F)	Time mowing the grass (hours)	Water Consumption (ounces)
75	1.85	16
83	1.25	20
85	1.5	25
85	1.75	27
92	1.15	32
97	1.75	48
99	1.6	48

Si vuole verificare il legame tra consumo d'acqua e queste due caratteristiche.

Regressione Multipla: Esempio I



Osservando i grafici a coppie si nota come la temperatura e il consumo d'acqua abbiano un chiaro legame lineare, mentre il ruolo del tempo non è del tutto chiaro quanto sia influente

Regressione Multipla: Esempio I

Calcoliamo i coefficienti di correlazione e poi l'R multiplo. I coefficienti sono organizzati nella *matrice di correlazione*

	Temperature	Time	Water
Temperature	1.000	-0.155	0.963
Time	-0.155	1.000	0.106
Water	0.963	0.106	1.000

$R = 0.996822$ complessivamente il legame lineare è forte.

Regressione Multipla: Esempio I

Stimiamo prima il modello di regressione tra Water Consumption e la sola Temperature:

```
> summary(y1)

Call:
lm(formula = WC ~ Temp, data = dati1)

Residuals:
      1       2       3       4       5       6       7
 4.0080 -3.6013 -1.5037  0.4963 -4.6618  4.0824  1.1801

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -96.8452    16.0924  -6.018 0.001821 **
Temp          1.4512     0.1821   7.967 0.000503 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.777 on 5 degrees of freedom
Multiple R-squared: 0.927,  Adjusted R-squared: 0.9124
F-statistic: 63.47 on 1 and 5 DF,  p-value: 0.0005026
```

C'è un buon adattamento, vediamo se l'aggiunta del tempo migliora il modello

Regressione Multipla: Esempio I

```
> summary(y2)

Call:
lm(formula = WC ~ Temp + TMG, data = dati1)

Residuals:
      1       2       3       4       5       6       7
 1.0441  0.4642 -0.6935 -1.8264  0.1061  1.0252 -0.1197

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -121.65500    6.54035 -18.601 4.92e-05 ***
Temp           1.51236    0.06077  24.886 1.55e-05 ***
TMG           12.53168    1.93302   6.483  0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.245 on 4 degrees of freedom
Multiple R-squared: 0.9937, Adjusted R-squared: 0.9905
F-statistic: 313.2 on 2 and 4 DF,  p-value: 4.027e-05
```

L'adattamento in termini di R^2 migliora. Cosa è però l'*Adjusted R-squared*?

Regressione Multipla: Adjusted R^2

Un problema che si riscontra con i modelli di regressione multipla è che al crescere del numero delle variabili indipendenti aumenta il valore di R^2 anche se le variabili non contribuiscono più alla spiegazione della variabile dipendente.
E' quindi opportuno far riferimento ad una versione corretta di questo indice di adattamento che tenga conto del numero di variabili presenti nel modello e permetta di costruire un modello davvero significativo. Quindi se p è il numero delle variabili indipendenti, n il numero di osservazioni disponibili:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Nel nostro esempio l'aggiunta del tempo al modello migliora l'adattamento sia in termini di R^2 che di R^2 corretto.

Regressione Multipla: procedura stepwise

Come scegliere allora il numero di variabili più adatto? o in altre parole come costruire il modello migliore quando abbiamo a disposizione più variabili indipendenti.
Una possibilità è partire dal modello completo e poi stimare i modelli con tutte le combinazioni delle variabili esplicative fino al modello con la sola intercetta, confrontando i valori del Multiple Rsquared e dall'adjusted R-squared
In R si può usare una procedura automatica che svolge questo compito anche se va sempre usata con attenzione. Il modello finale deve avere anche un senso logico..

Regressione Multipla: procedura stepwise

procedura stepwise,il confronto tra i vari modelli si basa su più indicatori di qualità tra cui l'AIC (akaike information criterion) basato sulla verosimiglianza e quindi sull'assunzione di gaussianità.
La procedura stepwise può lavorare sia provando a stimare tutti i modelli partendo dal modello completo e levando una variabile alla volta fino al modello contenente la sola intercetta, che usare la strada inversa cioè partendo dal modello con la sola intercetta arrivare al modello saturo.

Regressione Multipla: test sui coefficienti

- Tramite il valore di $\bar{R}^2$ possiamo definire quanto è "sensato" globalmente un modello di regressione multipla
- Volendo capire quale sia il contributo di ciascuna variabile indipendente dobbiamo condurre un test di ipotesi sui coefficienti del modello (in realtà i software statistici fanno questo test automaticamente)
- Il problema di ipotesi è:

$$\begin{aligned} H_0 &: \beta_i = 0 \\ H_1 &: \beta_i \neq 0 \end{aligned}$$

dove β_i è il coefficiente della variabile i-esima

Regressione Multipla: test sui coefficienti

- la statistica test è:

$$t_\beta = \frac{\hat{\beta}_i}{s_{\beta_i}} \tilde{t}_{n-2}$$

  che si distribuisce come una t di Student con $n - p$ gradi di libertà (stimiamo p coefficienti). $\hat{beta}_i$ è il valore stimato del coefficiente i-esimo e s_{β_i} l'errore standard corrispondente.
- quindi se al valore della statistica test corrisponde un p-value minore del livello di significatività scelto, rifiutiamo H_0 e la variabile indipendente corrispondente ha influenza sulla variabile risposta.

Regressione multipla: confronti tra coefficienti

- Per confrontare il ruolo delle diverse variabili indipendenti sulla variabile risposta è bene ricordare che queste sono di solito misurate su scale molto diverse
- Prima di fare confronti è necessario stimare il modello con le variabili (tutte) riportate alla stessa scala (standardizzate)
- Osserviamo come cambia il modello se standardizziamo tutte le variabili, ricordiamo che in questo caso tutte le variabili avranno media 0 e varianza= 1, se $z_i = \frac{y_i - \bar{y}}{\sigma_y}$ e $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}$, $j = 1, \ldots, p$

$$z_i = \beta_1 \tilde{x}_{i1} + \ldots + \beta_p \tilde{x}_{ip} + \varepsilon_i, \quad i = 1, \ldots, n$$

non c'é più l'intercetta che ora vale zero per costruzione.

Regressione Multipla: Esempio II

Abbiamo a disposizione dei valori di concentrazione (medie giornaliere) di Ozono rilevati a New York in 153 giorni tra maggio e settembre, inoltre sono state rilevate diverse variabili che possono contribuire alla presenza di ozono:

1. Solar.R Radiazione solare (lang)
2. Wind Velocità del vento (mph)
3. Temp TemperaturA (gradi F)
4. Month Mese (1-12)
5. Day Giorno del mese (1-31)

Su questi dati procediamo con l'esercitazione. (Script su AirQuality)