MODELLI DI REGRESSIONE SEMPLICE e MULTIPLA (1)

DESCRIZIONE OUTPUT EXCEL (pag. 9 e 10)

DATASET : IMPIEGATI

Contiene dati, su un campione di 474 impiegati, relativi ad alcune variabili; vengono prese in considerazione le seguenti:

- SALARY/1000 = retribuzione annua (in migliaia di dollari)
- EDUC : anni di istruzione
- PREVEXP : esperienza lavorativa precedente (in mesi)

Per ottenere l'output relativo al modello di regressione semplice (con variabile dipendente SALARY/1000 e variabile esplicativa EDUC) e l'output relativo al modello di regressione multipla (con variabile dipendente SALARY/1000 e variabili esplicative EDUC e PREVEXP) si procede come segue (dopo aver installato il componente aggiuntivo ANALISI DEI DATI):

- dal menù DATI, si seleziona ANALISI DATI;
- dal menù che appare si seleziona REGRESSIONE;
- nella finestra che si apre si selezionano i dati relativi alla variabile dipendente (Y) e alla variabile indipendente (X) (o alle variabili indipendenti, tutte in blocco, nel caso della regressione multipla);
- se incluse le etichette, si seleziona la cella relativa;
- eventualmente, si sceglie un livello di confidenza aggiuntivo a quello fornito (0,95) in automatico.

1) REGRESSIONE SEMPLICE [i valori sono arrotondati al quarto decimale]

Y = SALARY/1000

X = EDUC

- Nella I tabella vengono riportati l'indice R^2, pari a 0.4363, che indica la proporzione di variabilità spiegata sulla variabilità totale della variabile dipendente. Viene inoltre indicata l'ampiezza del campione n = 674. Si tralasciano le rimanenti quantità.

- Nella II tabella vengono riportati i valori della DEVIANZA TOTALE (SST, indicata come TOTALE, uguale a 137916.4954), della DEVIANZA RESIDUA SSE (indicata come RESIDUO, pari a 77738.2777), della DEVIANZA SPIEGATA SSR (indicata come REGRESSIONE, pari a 60178.2178).
Le rimanenti quantità presenti nella II tabella possono essere tralasciate (nella regressione semplice).

- Nella III tabella vengono riportate le quantità necessarie per l'inferenza su β_1 (e β_0); Fissiamo l'attenzione su β_1, corrispondente alla II riga, indicata con EDUC (che è la variabile alla quale corrisponde il coefficiente β_1).

Colonna coefficienti = stima di β_1 (ovvero $b_1 = 3.9098$) (e stima di β_0 (cioè $b_0 = -18.3312$)). Da qui si ricava l'equazione stimata $\hat{y} = -18.3312 + 3.9098x$.

Colonna errore standard: standard error di b_1, ovvero

$$s_{b_1} = 0,2045$$

Colonna Stat t = valore della statistica test osservata, ovvero

$$T_{oss} = \frac{b_1}{s_{b_1}} = 19.1150$$

Colonna valore di significatività = p-value relativo al test bilaterale corrispondente alle ipotesi $H_0: \beta_1 = 0$ contro $H_1: \beta_1 \neq 0$, pari a $9.6397 \cdot 10^{-61}$ (quindi praticamente uguale a 0, per cui si rifiuta H_0 a qualunque livello)

Colonne Inferiore 95% e Superiore 95% = estremi inferiore e superiore dell'intervallo di confidenza al 95% per β_1 (l'intervallo è quindi $(3.5080, 4.3118)$)

Colonne Inferiore 99% e Superiore 99% = come il precedente, ma al 99%

2) REGRESSIONE MULTIPLA [valori sempre arrotondati al quarto decimale]

Y = SALARY/1000

X_1 = EDUC

X_2 = PREVEXP

- Nella prima tabella sono riportati gli indici $R^2 = 0.4415$, $R = 0.6644$ (radice quadrata di R^2, uguale al valore assoluto del coefficiente di correlazione lineare tra Y e $\hat{Y}$, indicato come R multiplo).

- Nella seconda tabella sono riportati:

1° riga:
- gdl = 2, numero di variabili esplicative (k);
- SQ = 60884.1162, devianza spiegata = SSR;
- MQ = SQ/gdl = SSR/k = 30442.06;
- $F = \frac{SSR/k}{SSE/n-k-1} = \frac{30442.06}{163.5507} = 186.1322$, che rappresenta il valore osservato della statistica F relativa al test di "significatività globale" del modello, per le ipotesi $H_0: \beta_1 = \beta_2 = 0$ contro H_1 = almeno uno dei due coefficienti è $\neq 0$.
- Significatività F, ovvero il p-value relativo al test descritto al punto precedente.

2° riga:
- gdl = 471 = n-k-1 (gradi di libertà residui del modello);

- SQ = 77032,3812 # devianza residua = SSE ;
- MQ = SQ/gdl = $\frac{SSE}{n-k-1} = \frac{77032,3812}{471} = 163.5508$;

3° npx :
- gdl = n-1 = 473
- SQ = 137916.4854, devianza totale = SST.

- Nella terza tabella sono riportate le quantità relative all'inferenza (test e intervalli, oltre alle stime puntuali) per β_0, β_1 (coefficiente della variabile EDUC) e β_2 (coefficiente della variabile PREVEXP). La descrizione è identica a quella riportata nella parte relativa al modello di regressione lineare semplice.

L'equazione stimata del modello è:

$$\hat{y} = -20.9783 + 4.0203 \cdot X_1 + 0.0121 \cdot X_2$$

Sulla base dei due output, rispondiamo alle seguenti domande:

1) Nel modello di regressione che include la sola variabile EDUC, questa è significativa per la spiegazione della retribuzione? Se sì, qual è il suo effetto?
2) Il modello che include EDUC e PREVEXP è globalmente significativo?
3) Nel modello che include EDUC e PREVEXP, quest'ultima apporta un contributo significativo alla spiegazione della retribuzione (a livello 0.05)? Se sì, qual è il suo effetto?
4) Qual è la percentuale di variabilità complessivamente spiegata dagli anni di istruzione e dall'esperienza precedente?
5) Si preveda, usando il modello con EDUC e PREVEXP, la retribuzione media degli impiegati con 12 anni di istruzione e 200 mesi di esperienza.
6) Si effettui la stessa previsione richiesta al punto precedente per il modello con la sola EDUC (trascurando quindi il dato su PREVEXP).

1) Si tratta di un modello di regressione semplice. Occorre effettuare il test per $H_0: \beta_1 = 0$ contro $H_1: \beta_1 \neq 0$; dall'output si rileva che il p-value per tale test è $9.6397 \cdot 10^{-61} \simeq 0$, per cui si rifiuta H_0 (a qualunque livello); c'è quindi evidenza che $\beta_1 \neq 0$, per cui EDUC è significativa. L'effetto si rileva dalla stima di β_1, pari a $b_1 = 3.9099$; ad un incremento di un anno di istruzione è associato un incremento (medio) pari a 3909.9 dollari di retribuzione.

2) Si tratta di un modello di regressione multipla, con due variabili esplicative. Per stabilire se è globalmente significativo, occorre effettuare un test F globale, relativo alle ipotesi $H_0: \beta_1 = \beta_2 = 0$ contro H_1 = almeno uno dei due coefficienti è diverso da 0 (β_1 e β_2 sono i coefficienti relativi a EDUC e PREVEXP rispettiv.). Il p-value relativo al suddetto test è $2.7036 \cdot 10^{-60} \simeq 0$, per cui si rifiuta H_0 a qualunque livello $\Rightarrow$ il modello è globalmente significativo.

3) Occorre effettuare, nell'ambito del modello di regressione multipla, il test per la significatività della singola variabile PREVEXP, relativo cioè alle ipotesi $H_0: \beta_2 = 0$ contro $H_1: \beta_2 \neq 0$. Il p-value per questo test è $0.0383 < 0.05$, per cui si rifiuta H_0 (a livello 0.05); dunque PREVEXP è significativa in questo modello. La stima di β_2 è 0.0121 $: ad un incremento di un mese di esperienza, fissati gli anni di istruzione, è associato un incremento medio di retribuzione pari a 12.1 $.

4) La percentuale di variabilità spiegata in questo modello di regressione multipla è pari al 44.15 %. (dato dal valore di R^2).

5) L'equazione stimata del modello è:

$$\hat{y} = -20.9783 + 4.0203 \cdot X_1 + 0.0121 \cdot X_2.$$

La previsione è quindi:

$$\hat{y}_{n+1} = -20.9783 + 4.0203 \cdot 12 + 0.0121 \cdot 200 = 28.6853.$$

6) L'equazione stimata del modello (di regressione semplice) è:

$$\hat{y} = -18.3312 + 3.9098 \cdot X.$$

La previsione è quindi:

$$\hat{y}_{n+1} = -18.3312 + 3.9098 \cdot 12 = 28.5876.$$

OUTPUT RIEPILOGO

OUTPUT REGRESSIONE SEMPLICE

Statistica della regressione	
R multiplo	0.66055891
R al quadrato	0.43633807
R al quadrato corretto	0.43514387
Errore standard	12.8335397
Osservazioni	474

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	1	60178.21776	60178.2178	365.3813749	9.63974E-61
Residuo	472	77738.27768	164.699741		
Totale	473	137916.4954			

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*	*Inferiore 99.0%*	*Superiore 99.0%*
Intercetta	-18.331178	2.821911555	-6.49601438	2.09629E-10	-23.87624179	-12.78611427	-25.62944694	-11.03290912
educ	3.90990671	0.204547037	19.1149516	9.63974E-61	3.507971232	4.311842181	3.380889674	4.438923739

OUTPUT REGRESSIONE MULTIPLA

OUTPUT RIEPILOGO

Statistica della regressione	
R multiplo	0.6644218
R al quadrato	0.4414564
R al quadrato corretto	0.4390846
Errore standard	12.788694
Osservazioni	474

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	60884.11424	30442.06	186.1322301	2.70364E-60
Residuo	471	77032.3812	163.5507		
Totale	473	137916.4954			

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*	*Inferiore 99.0%*	*Superiore 99.0%*
Intercetta	-20.978304	3.087257655	-6.79513	3.28056E-11	-27.04480627	-14.9118009	-28.96290208	-12.99370513
educ	4.0203433	0.210649876	19.08543	1.42045E-60	3.606413509	4.434273168	3.475537926	4.565148751
prevexp	0.0120713	0.005810445	2.077516	0.038295234	0.000653688	0.023488891	-0.002956309	0.027098888