CAPITOLO

20

SOFTWARE PER L'ELABORAZIONE DEI DATI: REGRESSIONE MULTIPLA

20.1 INTRODUZIONE

Nel Capitolo 17 si è visto come ottenere la stima del modello di regressione lineare semplice mediante l'uso dei software R ed Excel. In questo capitolo verranno mostrate, molto schematicamente, le procedure che consentono di estendere le tecniche viste in precedenza al caso della regressione lineare multipla.

Il software R contempla un insieme di procedure che permettono la stima dei parametri, la verifica d'ipotesi, la selezione del modello, la verifica delle assunzioni del modello. Anche con Excel, attraverso lo strumento **Analisi dati**, si può applicare con semplicità e qualche limitazione la regressione lineare multipla.

Nell'illustrare le procedure di stima per intervallo e verifica d'ipotesi sui parametri del modello di regressione, utilizzeremo il file di dati *Demo* (.txt o .xls) che contiene i dati di un sondaggio fittizio realizzato su un campione di 6400 persone con informazioni di base di tipo demografico (*Età, Sesso, Stato civile, Istruzione* ecc.) economico e di possesso (*Possesso di TV, fax, stereo, cellulare, PC* ecc.).

20.2 REGRESSIONE LINEARE MULTIPLA CON R

Come abbiamo già visto nel Paragrafo 17.5, per stimare con R un modello di regressione lineare, si può utilizzare la funzione `lm()` che ha sintassi:

```
lm(formula, data, subset, weights, na. action,...).
```

L'argomento `formula` definisce il modello che si vuole stimare e nella regressione lineare multipla ha la forma

$$y \sim x_1 + x_2 + ... + x_k$$

Per il significato degli altri argomenti si rimanda al Paragrafo 17.5.

In questo paragrafo considereremo la variabile *Reddito famigliare* (*REDDITO*) dipendente dall'*Età* (*ETA*), dal *Grado di istruzione* (*ISTRUZ*), dal *Numero di componenti del nucleo famigliare* (*NFAM*), dal *Numero di anni di permanenza nell'impiego attuale* (*IMPIEGO*).

Possiamo quindi definire la formula del nostro modello come:

$$REDDITO \sim ETA + ISTRUZ + NFAM + IMPIEGO$$

Possiamo leggere il data-set e memorizzarlo in un data-frame con il comando:

```
dataf<-read.csv(file ='d://Demo.txt',header = TRUE,
                sep=';', dec =',')
```

Per esaminare le prime 3 righe dei dati con i nomi delle variabili possiamo utilizzare il comando `head(dataf,3)`:

```
> head (dataf,3)
   ETA STATOCIV REDDITO CATREDD ISTRUZ IMPIEGO SODDLAV PENSIONE SESSO NFAM CELLUL
1  55         1      72 50 - 74      1      23       2        0     1     4      0
2  56         0     153     > 75      1      35       1        0     2     1      1
3  28         1      28 25 - 49      3       4       0        0     1     3      1
   LINEEMUL SEGRET CERCAPER TV VIDEOREG STEREO PC FAX
1         0      1        0  1        1      1  0   0
2         0      1        1  1        1      1  0   0
3         0      1        0  1        1      1  1   0
```

Sappiamo che il reddito è generalmente una variabile molto asimmetrica. Possiamo analizzarne la distribuzione tramite le istruzioni:

```
plot(density(dataf$REDDITO), main="Density Plot: REDDITO", ylab="Densità",
    sub = paste ("Skewness:", round (e1071:: skewness (dataf$REDDITO), 2)))
    polygon (density (dataf$REDDITO), col =" red")
```

Nel grafico della Figura 20.2.1 si osserva un'asimmetria elevata della variabile che è opportuno trattare.

Introducendo la trasformazione logaritmica otteniamo una distribuzione molto meno asimmetrica (Figura 20.2.2), come è testimoniato dall'indice di asimmetria che passa da 4,51 a 0,59.

La funzione `lm()` crea un oggetto da cui possono essere estratte molte informazioni. In particolare, indicando con `fm` l'oggetto creato dalla funzione, possiamo utilizzare il comando `summary (fm)` per ottenere tutti i risultati principali del modello,

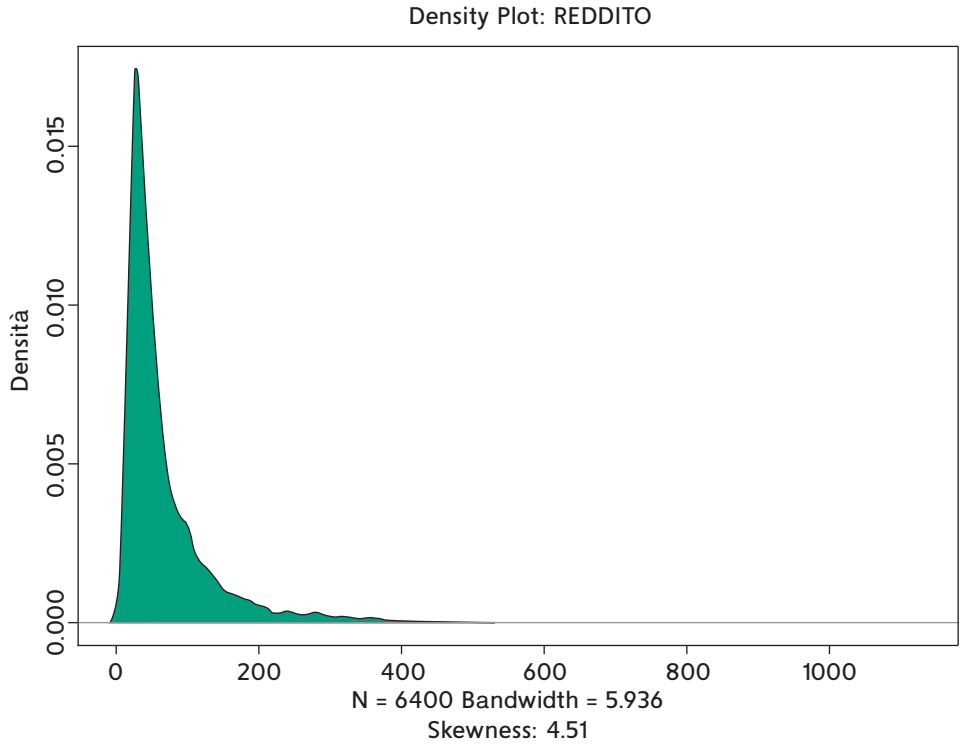

Figura 20.2.1
Grafico di densità per la variabile *REDDITO*.

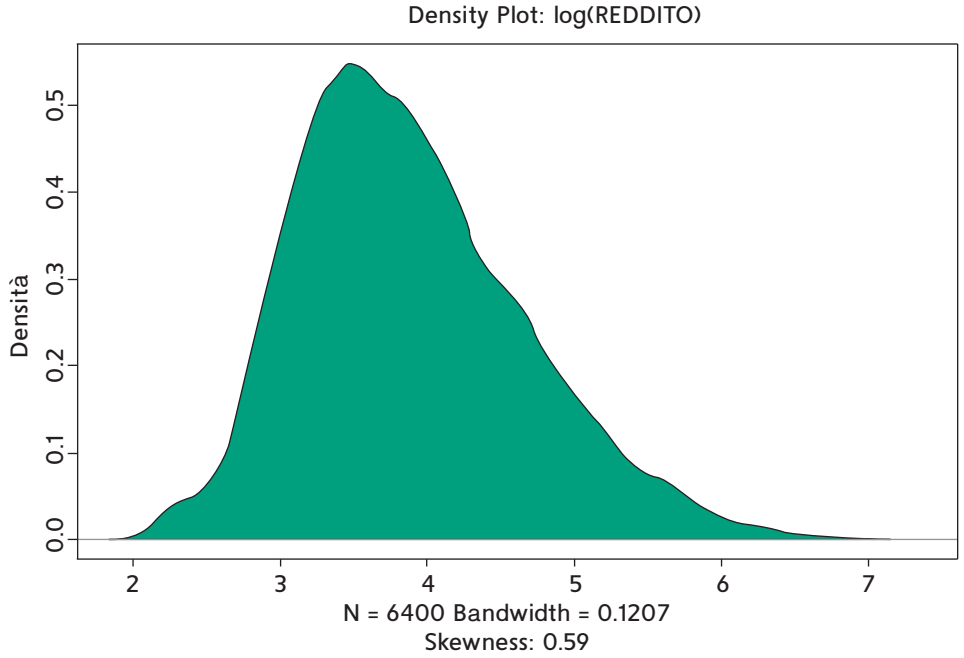

Figura 20.2.2
Grafico di densità per la variabile *log(REDDITO)*.

che includono le stime dei coefficienti, il test di significatività, l'analisi dei residui, l'indice R^2, il test F sul modello. Possiamo a questo punto confrontare il modello in cui si trova la variabile orginaria *REDDITO* con quello che si ottiene considerando il `log(REDDITO)` in cui il logaritmo è a base naturale.

Con la variabile non trasformata otteniamo l'output:

```
> fm <- lm(REDDITO ~ ETA + ISTRUZ + NFAM + IMPIEGO, dataf)
> summary(fm)

Call:
lm(formula = REDDITO ~ ETA + ISTRUZ + NFAM + IMPIEGO, data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-257.66  -24.91   -4.46   15.99  963.07

Coefficients:
             Estimate  Std. Error  t value    Pr(>|t|)
(Intercept) -19.49932    3.98143   -4.898    9.94e-07 ***
ETA          -0.24848    0.0811    -3.061     0.00222 **
ISTRUZ       15.65774    0.65786   23.801     < 2e-16 ***
NFAM          1.07319    0.53958    1.989     0.04675 *
IMPIEGO       5.33769    0.10238   52.135     < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61.47 on 6395 degrees of freedom
  (4175 observations deleted due to missingness)
Multiple R-squared:  0.3906, Adjusted R-squared:  0.3902
F-statistic:  1025 on 4 and 6395 DF,  p-value: < 2.2e-16
```

Considerando invece come variabile dipendente il `log(REDDITO)`, otteniamo:

```
> LOGREDD <- log(dataf$REDDITO)
> fm <- lm(LOGREDD ~ ETA + ISTRUZ + NFAM + IMPIEGO, dataf)
> summary(fm)

Call:
lm(formula = LOGREDD ~ ETA + ISTRUZ + NFAM + IMPIEGO, data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6755 -0.3103  0.0121  0.3374  2.5443

Coefficients:
                Estimate   Std. Error t value Pr(>|t|)
(Intercept)    2.9563312    0.0369769    79.951     < 2e-16 ***
ETA           -0.0033404    0.0007540    -4.430    9.56e-06 ***
ISTRUZ         0.1762391    0.0061098    28.845     < 2e-16 ***
NFAM           0.0082589    0.0050113     1.648       0.0994 .
IMPIEGO        0.0570896    0.0009509    60.040     < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5709 on 6395 degrees of freedom
Multiple R-squared:  0.4574, Adjusted R-squared:  0.4571
F-statistic:  1348 on 4 and 6395 DF,  p-value: < 2.2e-16
```

La valutazione dell'adattamento, condotta con l'indice R^2, ci dice che la variabilità spiegata è sempre piuttosto bassa, quindi ci dobbiamo aspettare una forte variabilità della nuvola dei punti rispetto all'iperpiano di regressione. Comunque, l'adattamento con la variabile trasformata migliora in maniera sensibile e quindi il modello è preferibile. La statistica *F* ci dice inoltre che il modello complessivamente è significativo.

Guardando il *p-value* dei coefficienti di regressione (`Pr(>|t|)`), occorre osservare che il coefficiente di *NFAM* risulta non significativo, e quindi la variabile potrebbe essere eliminata. In effetti, si può verificare che eliminando *NFAM* otteniamo fondamentalmente lo stesso adattamento con un modello più semplice.

Il modello finale ottenuto è quindi il seguente:

$$\widehat{LOGREDD} \sim 2.96 - 0.003 \cdot ETA + 0.176 \cdot ISTRUZ + 0.057 \cdot IMPIEGO$$

Non potendo realizzare e valutare una rappresentazione a 5 dimensioni, è utile analizzare il grafico dei valori predetti e dei valori osservati della variabile dipendente (Figura 20.2.3), oppure i grafici dei residui (Figura 20.2.4) che si possono ottenere facilmente con le seguenti istruzioni:

```
plot (fitted(fm), LOGREDD, col = 3, xlab='Valori stimati')
layout(matrix(c(1,2,3,4), 2, 2))
plot(fm)
```

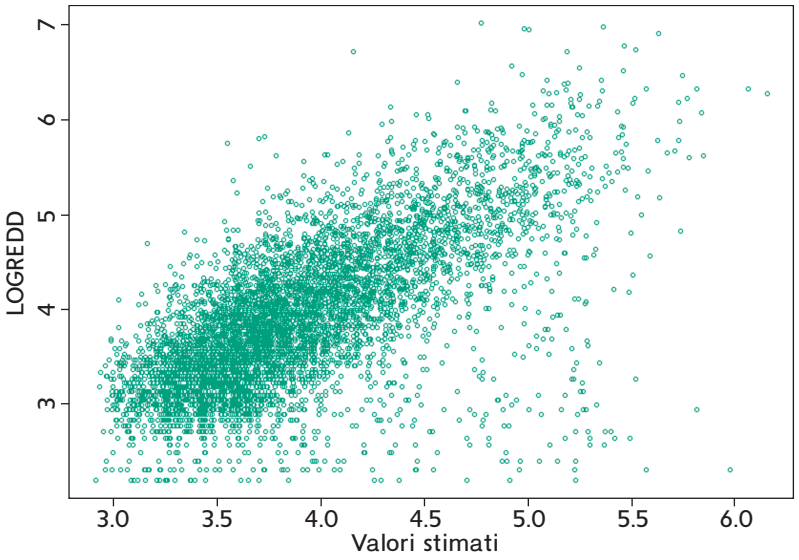

Figura 20.2.3
Grafico di dispersione tra valori osservati e stimati di *log(REDDITO)*.

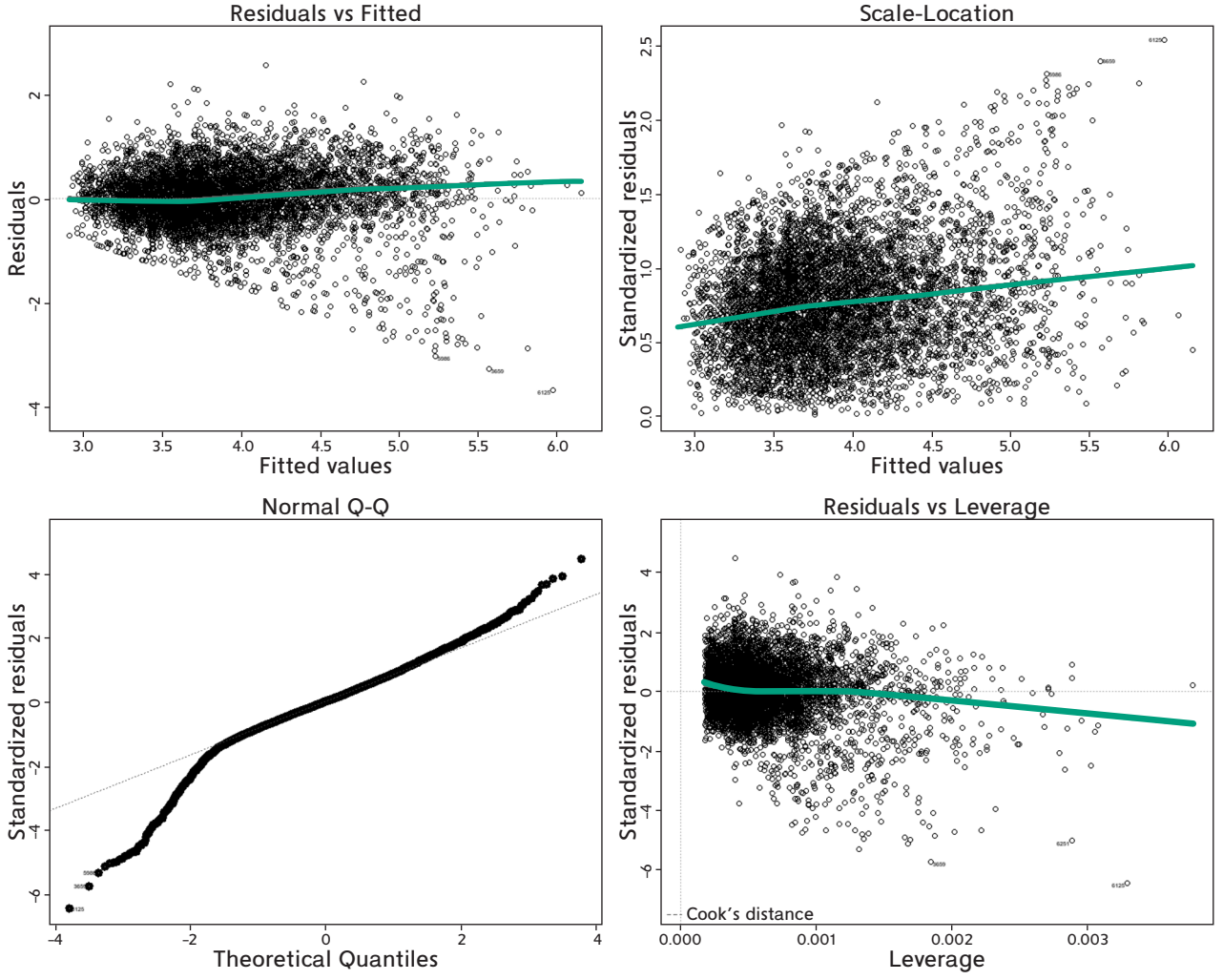

Figura 20.2.4
Grafici relativi ai residui del modello.

20.3 REGRESSIONE LINEARE MULTIPLA CON EXCEL

Lo strumento **Analisi dati** di Excel consente di stimare i parametri del modello di regressione lineare sia semplice sia multiplo. Partendo dal menu **Dati** si seleziona **Analisi dati** e quindi la voce **Regressione**, da cui si passa alla finestra mostrata in Figura 20.3.1. In questa finestra si devono indicare le aree del foglio in cui sono presenti i dati di input corrispondenti alla variabile dipendente e a quelle indipendenti (queste ultime devono essere disposte sul foglio in ordine consecutivo, da sinistra a destra, ed è possibile immettere fino a un massimo di 16 variabili). Si deve specificare, inoltre, se calcolare gli intervalli di confidenza per i coefficienti del modello (indicando anche il livello di confidenza) e i grafici dei residui da analizzare.

Consideriamo il file di dati *Demo* già citato nell'introduzione.

In questo paragrafo considereremo la variabile *Reddito famigliare* (*REDDITO*) dipendente dall'*Età* (*ETA*), dal *Grado di istruzione* (*ISTRUZ*), dal *Numero di componenti del nucleo famigliare* (*NFAM*), dal *Numero di anni di permanenza nell'impiego attuale* (*IMPIEGO*).

Occorre quindi spostare la variabile *Y*, nel nostro caso *REDDITO*, come prima colonna e mettere le altre variabili esplicative di seguito.

Selezionando quindi nella finestra della Figura 20.3.1 le variabili del modello e le statistiche da visualizzare come nella Figura 20.3.2, si ottengono le tabelle di Figu-

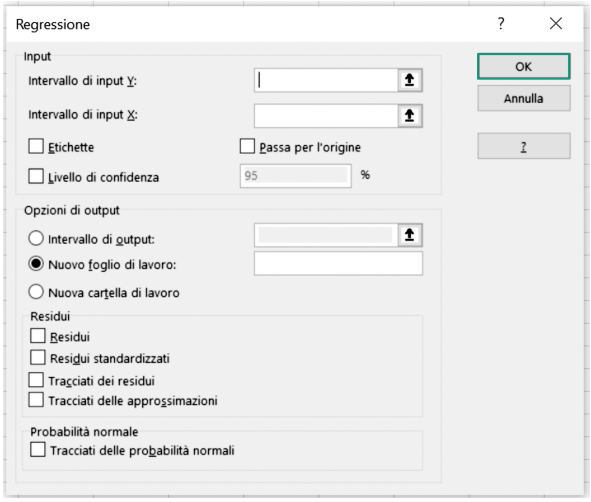

Figura 20.3.1
Finestra Regressione.

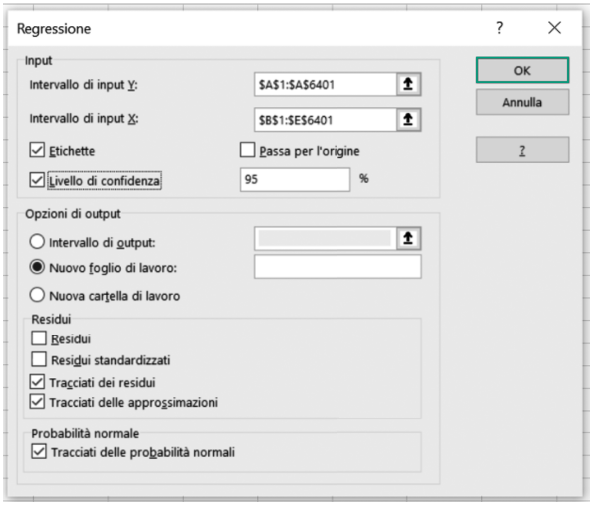

Figura 20.3.2
Compilazione della finestra Regressione.

	A	B	C	D	E	F	G
1	OUTPUT RIEPILOGO						
2							
3	*Statistica della regressione*						
4	R multiplo	0,625					
5	R al quadrato	0,391					
6	R al quadrato corretto	0,390					
7	Errore standard	61,470					
8	Osservazioni	6400,000					
9							
10	ANALISI VARIANZA						
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
12	Regressione	4,0	15487985,1	3871996,3	1024,7	0,000	
13	Residuo	6395,0	24164136,9	3778,6			
14	Totale	6399,0	39652121,9				
15							
16		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
17	Intercetta	-19,499	3,981	-4,898	0,0000	-27,304	-11,694
18	ETA	-0,248	0,081	-3,061	0,0022	-0,408	-0,089
19	ISTRUZ	15,658	0,658	23,801	0,0000	14,368	16,947
20	NFAM	1,073	0,540	1,989	0,0468	0,015	2,131
21	IMPIEGO	5,338	0,102	52,135	0,0000	5,137	5,538

Figura 20.3.3
Output della procedura Regressione.

ra 20.3.3. Affinché il programma legga i nomi delle variabili, presenti nella prima riga, occorre biffare la casella Etichette.

L'output di riepilogo, mostrato nella Figura 20.3.3, presenta il valore degli indici di bontà di adattamento, l'analisi della varianza, la stima dei parametri del modello. In particolare, nella tabella in cui sono riportate le stime dei parametri del modello sono presenti anche i corrispondenti test (valore della statistica test e del *p-value* associato) e gli intervalli di confidenza.

L'output del programma consiste anche nei grafici *Tracciati dei residui*, *Tracciati delle approssimazioni* e *Tracciati delle probabilità Normali*. I *Tracciati dei residui* consistono in un grafico di dispersione per ciascuna variabile che contrappone i valori della variabile con quelli dei residui del modello. Nel *Tracciato delle approssimazioni* si confrontano i valori di una variabile esplicativa con i valori osservati e stimati della variabile dipendente. Come esempio, abbiamo riportato questi due grafici solo per la variabile esplicativa *IMPIEGO*.

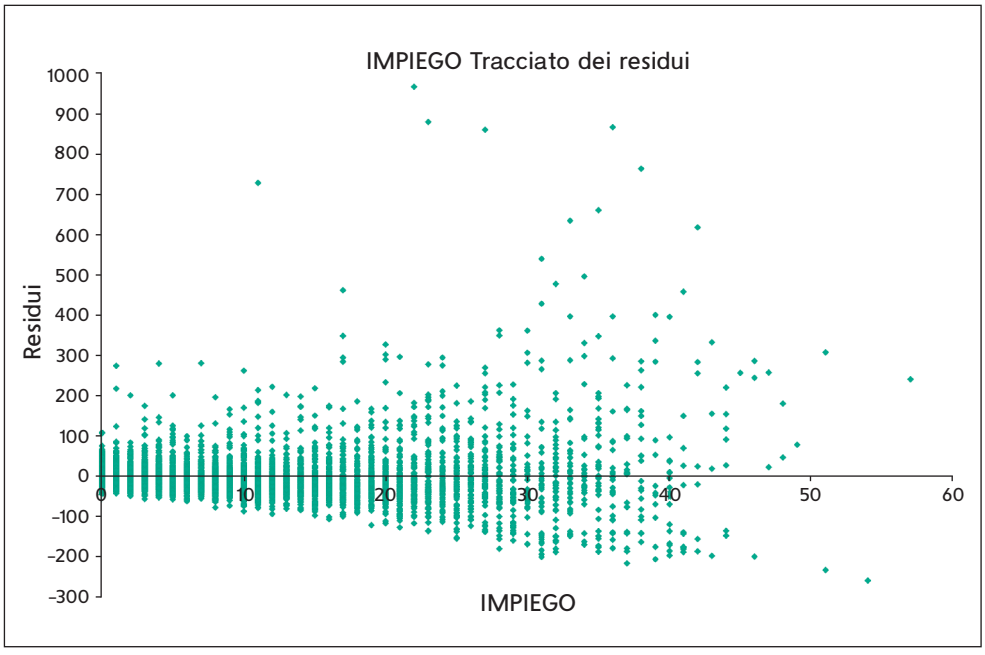

Figura 20.3.4
REDDITO: residui rispetto al carattere *IMPIEGO*.

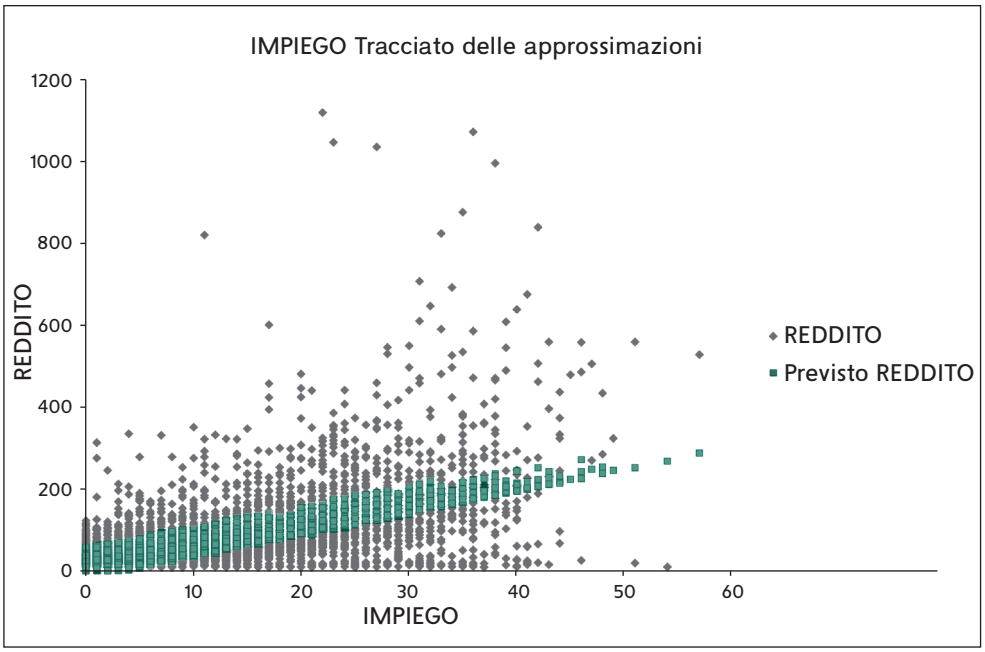

Figura 20.3.5
REDDITO e *REDDITO* stimato rispetto alla variabile *IMPIEGO*.

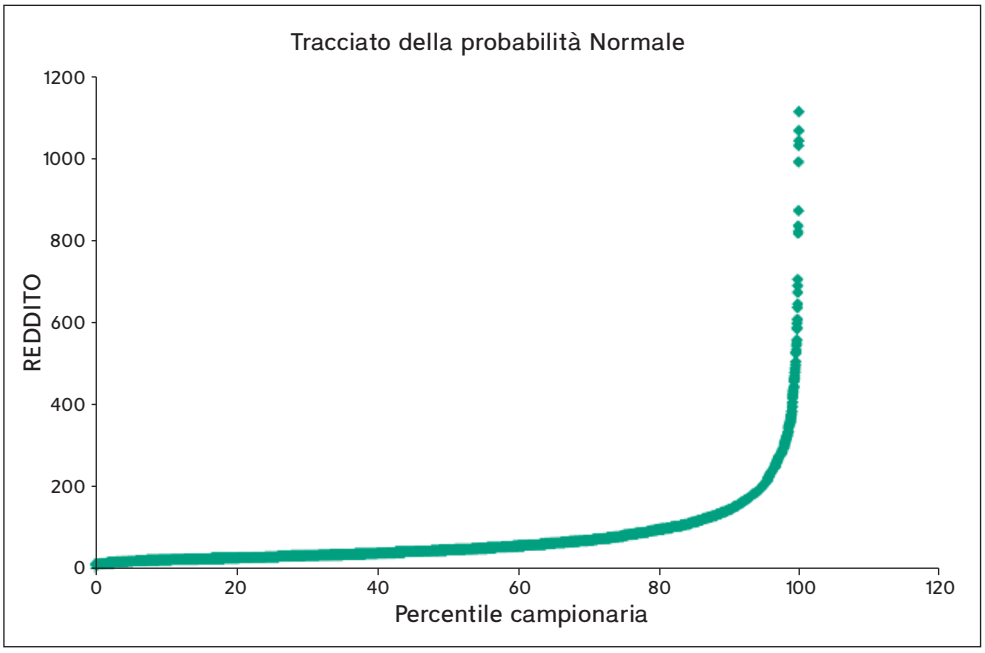

Figura 20.3.6
Tracciato della probabilità Normale.

Avendo biffato l'opzione *Tracciati delle probabilità Normali*, otteniamo anche il grafico di Figura 20.3.6 che ci permette di affermare che la distribuzione del carattere *REDDITO* è ben lontana da una distribuzione Normale.

Considerando la forte asimmetria e disnormalità del *REDDITO*, come appare anche nel grafico di Figura 20.3.6, possiamo ridefinire l'asse delle ordinate corrispondente al carattere *REDDITO* su una scala logaritmica (base 10) utilizzando le opzioni del grafico. Il grafico successivo mostra come la trasformazione sia stata utile per ottenere una distribuzione più vicina alla Normale.

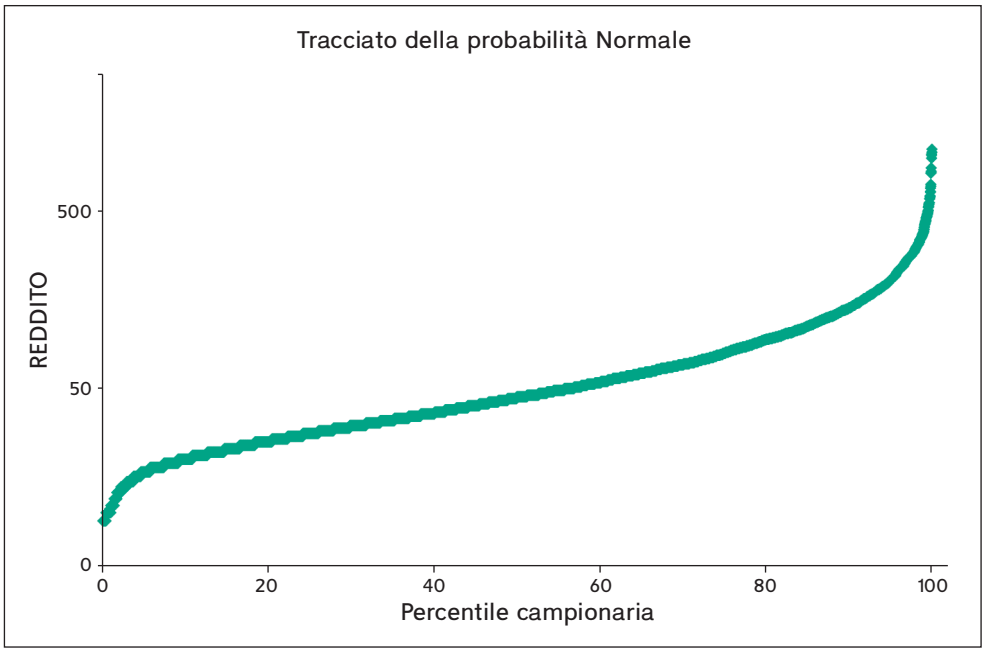

Figura 20.3.7
Tracciato della probabilità Normale con scala logaritmica.

	A	B	C	D	E	F	G
1	OUTPUT RIEPILOGO						
2							
3	*Statistica della regressione*						
4	R multiplo	0,6763					
5	R al quadrato	0,4574					
6	R al quadrato corretto	0,4571					
7	Errore standard	0,2479					
8	Osservazioni	6400,0000					
9							
10	ANALISI VARIANZA						
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
12	Regressione	4	331,42	82,86	1347,84	0,00	
13	Residuo	6395	393,12	0,06			
14	Totale	6399	724,54				
15							
16		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
17	Intercetta	1,284	0,016	79,951	0,000	1,252	1,315
18	ETA	-0,001	0,000	-4,430	0,000	-0,002	-0,001
19	ISTRUZ	0,077	0,003	28,845	0,000	0,071	0,082
20	NFAM	0,004	0,002	1,648	0,099	-0,001	0,008
21	IMPIEGO	0,025	0,000	60,040	0,000	0,024	0,026
22							

Figura 20.3.8
Output ottenuto con *LOGREDD*.

Possiamo quindi ripetere l'analisi considerando il logaritmo (base 10) del carattere *REDDITO*. Si osservi che Excel utilizza come default il logaritmo base 10 mentre R utilizza il logaritmo base naturale, da cui le differenze nei risultati con quelli riportati nel Paragrafo 20.2.

La valutazione dell'adattamento, considerando l'indice R^2, ci dice che la variabilità spiegata è sempre piuttosto bassa, quindi ci dobbiamo aspettare una forte variabilità della nuvola dei punti rispetto all'iperpiano di regressione. Comunque, l'adattamento con la variabile trasformata migliora in maniera sensibile e quindi il modello è preferibile. La statistica F ci dice che il modello complessivamente è significativo.

Guardando il *p-value* dei coefficienti di regressione (Valore di significatività), occorre osservare che il coefficiente di *NFAM* risulta non significativo, quindi la variabile potrebbe essere eliminata. In effetti, si può verificare che eliminando *NFAM* otterremmo fondamentalmente lo stesso adattamento con un modello più semplice.

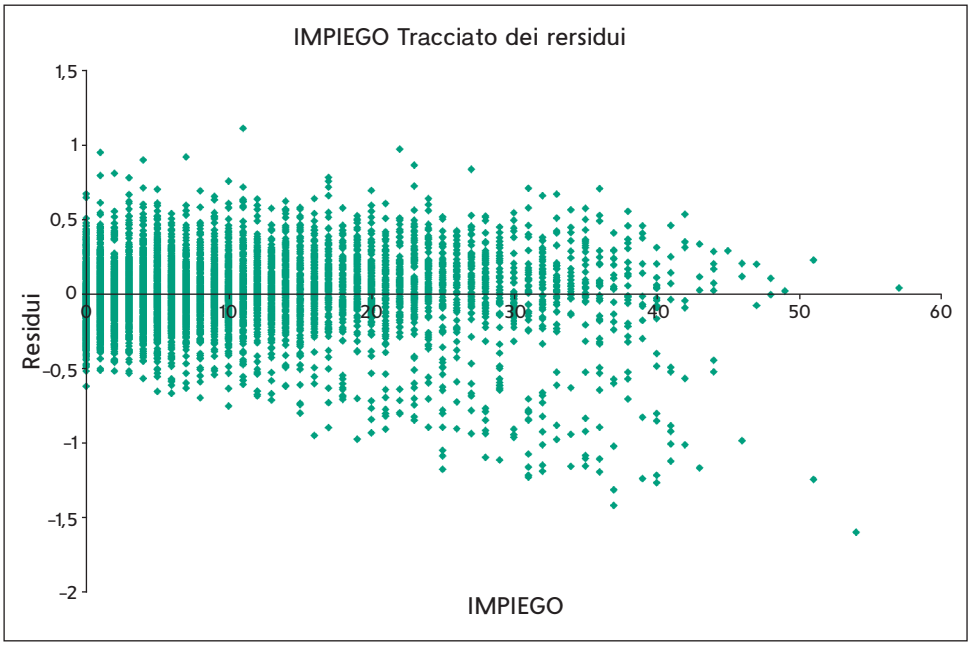

Figura 20.3.9
LOGREDD: residui rispetto al carattere *IMPIEGO*.

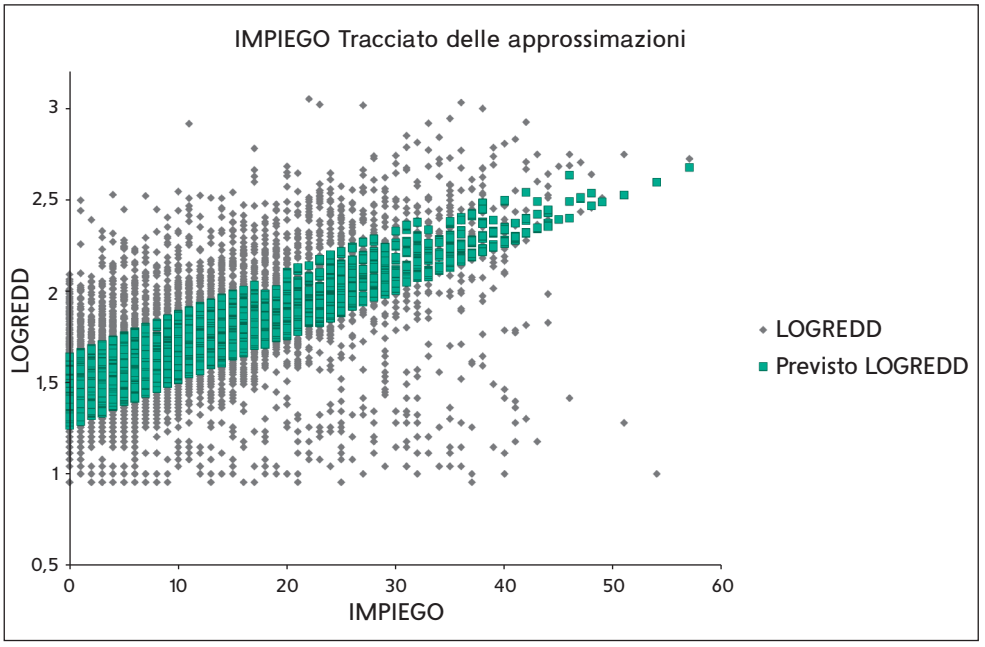

Figura 20.3.10
LOGREDD e *LOGREDD* stimato rispetto alla variabile *IMPIEGO*.

Il modello finale che si ottiene eliminando il carattere *NFAM* presenta tutti i coefficienti significativi ed è il seguente:

$$\widehat{LOGREDD} = 1{,}297 - 0.002 \cdot ETA + 0.076 \cdot ISTRUZ + 0.025 \cdot IMPIEGO$$

Possiamo quindi riprodurre i grafici già visti nelle Figure 20.3.4 e 20.3.5, utilizzando il modello finale ottenuto. I grafici mostrano un leggero miglioramento nell'adattamento del modello in riferimento alla variabile *IMPIEGO*.

20.4 SELEZIONE DEL MODELLO DI REGRESSIONE CON R ED EXCEL

Nei precedenti paragrafi abbiamo già accennato all'utilità di ridurre il numero di variabili esplicative eliminando dal modello la variabile *NFAM* che risultava non necessaria. In questo modo si cerca di ridurre il rumore presente nel modello ottenendo un modello più semplice e robusto. Se il modello include molte variabili esplicative è necessaria una strategia di selezione che, senza considerare tutte le possibilità di scelta delle variabili esplicative, ci permetta di ottenere il modello ottimale con un tempo di elaborazione ragionevole. In effetti se abbiamo 10 variabili esplicative, il numero di modelli alternativi da considerare sarebbe 2^{10}, ossia 1024.

La tecnica più comune di selezione delle variabili è di tipo **stepwise** (per passi successivi), in pratica si aggiunge (**forward stepwise**) o si elimina (**backward stepwise**) una variabile alla volta fino a ottenere il modello desiderato. L'aggiunta o l'eliminazione avvengono in base a un livello prestabilito di significatività (probabilità) del valore della statistica test F oppure di un indice come AIC o BIC. Questa procedura si presta a ovvie obiezioni, in quanto la sequenza di test che ne derivano rendono i livelli di significatività poco affidabili e il valore di R^2 troppo ottimistico. D'altra parte la procedura risulta generalmente efficace.

Selezione del modello di regressione in R

ESEMPIO 20.4.1

Richiamando il dataset *Demo* inseriamo, oltre alle variabili esplicative già utlizzate nel Paragrafo 20.2.1, anche le variabili *Possesso di telefono cellulare*, *di linee multiple*, *di TV*, *di videoregistratore*, *di fax*, *di computer*. Utilizzando tutte le variabili otteniamo:

```
> lm(formula = LOGREDD ~ ETA + ISTRUZ + NFAM + IMPIEGO + CELLUL +
                    LINEEMUL + TV + VIDEOREG + FAX + PC, data = dataf)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5882 -0.3030  0.0054  0.3327  2.5187

Coefficients:
                Estimate   Std. Error t value     Pr(>|t|)
(Intercept)    1.9569334    0.0796024    24.584     < 2e-16 ***
ETA           -0.0027939    0.0007377    -3.787    0.000154 ***
ISTRUZ         0.1577599    0.0066200    23.831     < 2e-16 ***
NFAM           0.0017907    0.0048879     0.366    0.714114
IMPIEGO        0.0545204    0.0009374    58.162     < 2e-16 ***
CELLUL         0.0897809    0.0161850     5.547    3.02e-08 ***
LINEEMUL      -0.0149803    0.0152774    -0.981    0.326853
TV             0.5340262    0.0744915     7.169    8.41e-13 ***
VIDEOREG       0.5216797    0.0380038    13.727     < 2e-16 ***
FAX            0.0596622    0.0197968     3.014    0.002591 **
PC            -0.0085156    0.0156129    -0.545       0.585484
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5542 on 6389 degrees of freedom
```

```
Multiple R-squared:  0.4892, Adjusted R-squared:  0.4884
F-statistic: 611.8 on 10 and 6389 DF,  p-value: < 2.2e-16
```

Si possono notare alcune variabili con coefficienti non significativi, per cui è consigliabile una procedura di selezione delle variabili. In R ci sono molte librerie specializzate che si potrebbero utilizzare, noi utilizzeremo la solita libreria stats. In questa libreria c'è la funzione `steps()` che consente di applicare una strategia stepwise di selezione basata sull'indice AIC.

```
> step(fm)
Start:  AIC=-7543.98
LOGREDD ~ ETA + ISTRUZ + NFAM + IMPIEGO + CELLUL + LINEEMUL +
    TV + VIDEOREG + FAX + PC

             Df  Sum of Sq      RSS     AIC
- NFAM        1       0.04   1962.3 -7545.8
- PC          1       0.09   1962.4 -7545.7
- LINEEMUL    1       0.30   1962.6 -7545.0
<none>                      1962.3 -7544.0
- FAX         1       2.79   1965.1 -7536.9
- ETA         1       4.41   1966.7 -7531.6
- CELLUL      1       9.45   1971.7 -7515.2
- TV          1      15.78   1978.1 -7494.7
- VIDEOREG    1      57.87   2020.2 -7359.9
- ISTRUZ      1     174.42   2136.7 -7001.0
- IMPIEGO     1    1038.99   3001.3 -4826.5

Step:  AIC=-7545.84
LOGREDD ~ ETA + ISTRUZ + IMPIEGO + CELLUL + LINEEMUL + TV + VIDEOREG +
    FAX + PC

             Df  Sum of Sq      RSS     AIC
- PC          1       0.09   1962.4 -7547.5
- LINEEMUL    1       0.28   1962.6 -7546.9
<none>                      1962.3 -7545.8
- FAX         1       2.80   1965.1 -7538.7
- ETA         1       4.75   1967.1 -7532.4
- CELLUL      1       9.48   1971.8 -7517.0
- TV          1      15.79   1978.1 -7496.6
- VIDEOREG    1      58.18   2020.5 -7360.8
- ISTRUZ      1     174.54   2136.9 -7002.5
- IMPIEGO     1    1039.09   3001.4 -4828.2

Step:  AIC=-7547.54
LOGREDD ~ ETA + ISTRUZ + IMPIEGO + CELLUL + LINEEMUL + TV + VIDEOREG +
    FAX

             Df  Sum of Sq      RSS     AIC
- LINEEMUL    1       0.34   1962.8 -7548.4
<none>                      1962.4 -7547.5
- FAX         1       2.73   1965.2 -7540.6
- ETA         1       4.69   1967.1 -7534.3
```

```
- CELLUL      1         9.40    1971.8   -7519.0
- TV          1        15.82    1978.2   -7498.2
- VIDEOREG    1        58.11    2020.5   -7362.8
- ISTRUZ      1       183.80    2146.2   -6976.5
- IMPIEGO     1      1041.04    3003.5   -4825.8

Step:  AIC=-7548.42
LOGREDD ~ ETA + ISTRUZ + IMPIEGO + CELLUL + TV + VIDEOREG + FAX

             Df Sum of Sq     RSS      AIC
<none>                        1962.8   -7548.4
- FAX         1         2.62    1965.4   -7541.9
- ETA         1         4.89    1967.7   -7534.5
- CELLUL      1         9.08    1971.8   -7520.9
- TV          1        15.73    1978.5   -7499.3
- VIDEOREG    1        57.77    2020.5   -7364.8
- ISTRUZ      1       188.50    2151.3   -6963.5
- IMPIEGO     1      1046.53    3009.3   -4815.4

Call:
lm(formula = LOGREDD ~ ETA + ISTRUZ + IMPIEGO + CELLUL + TV +
     VIDEOREG + FAX, data = dataf)

Coefficients:
(Intercept)        ETA     ISTRUZ   IMPIEGO     CELLUL         TV   VIDEOREG
   1.967334  -0.002875   0.155358 0.054440   0.086395   0.532867   0.518718
        FAX
    0.057399
```

La funzione `steps()` senza argomenti effettua una ricerca backward utilizzando l'indice **AIC**. Se si vuole effettuare una selezione più stringente si può utilizzare l'indice **BIC**, aggiungendo alla funzione l'argomento $k = \log(n)$, ossia, nel nostro caso, `steps(fm, k=8.76)`. Molte altre opzioni sono disponibili, si consulti al riguardo la documentazione della libreria stats.

Selezione del modello di regressione in Excel

ESEMPIO 20.4.2

In Excel non è prevista una procedura automatica di selezione delle variabili, pertanto si dovrà procedere, come descritto nel Paragrafo 18.7, eseguendo il test F per modelli annidati. Riprendendo le variabili esplicative dell'esempio precedente, stimiamo con Excel il modello completo (che include tutte le dieci variabili) e uno ridotto con le sette variabili più importanti. Nelle Figure 20.4.1 e 20.4.2 sono mostrati i corrispondenti output.

Considerando le due tabelle ANOVA possiamo applicare la formula (18.7.1) per il calcolo della statistica test F. Si ha:

$$F = \frac{(370{,}20 - 370{,}11)/(10-6)}{370{,}11/(6400-10-1)} = \frac{0{,}0225}{0{,}0579} = 0{,}3884$$

che si deve confrontare con il valore, per un $\alpha = 0{,}05$, ottenuto da una $F - Fisher$ con 4 e 6389 gradi di libertà e pari a $F_{0{,}05} = 2{,}373$. Poiché $0{,}3884 < 2{,}373$, il modello ridotto non può essere rifiutato. Pertanto, si giunge allo stesso risultato visto nell'esempio precedente.

Figura 20.4.1
Modello di regressione con 10 variabili esplicative.

	A	B	C	D	E	F	G
1	OUTPUT RIEPILOGO						
2							
3	*Statistica della regressione*						
4	R multiplo	0,6994					
5	R al quadrato	0,4892					
6	R al quadrato corretto	0,4884					
7	Errore standard	0,2407					
8	Osservazioni	6400					
9							
10	ANALISI VARIANZA						
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
12	Regressione	10	354,43	35,44	611,82	0,00	
13	Residuo	6389	370,11	0,06			
14	Totale	6399	724,54				
15							
16		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
17	Intercetta	0,850	0,035	24,584	0,000	0,782	0,918
18	ETA	-0,001	0,000	-3,787	0,000	-0,002	-0,001
19	ISTRUZ	0,069	0,003	23,831	0,000	0,063	0,074
20	IMPIEGO	0,024	0,000	58,162	0,000	0,023	0,024
21	NFAM	0,001	0,002	0,366	0,714	-0,003	0,005
22	CELLUL	0,039	0,007	5,547	0,000	0,025	0,053
23	LINEEMUL	-0,007	0,007	-0,981	0,327	-0,020	0,007
24	TV	0,232	0,032	7,169	0,000	0,169	0,295
25	VIDEOREG	0,227	0,017	13,727	0,000	0,194	0,259
26	PC	-0,004	0,007	-0,545	0,585	-0,017	0,010
27	FAX	0,026	0,009	3,014	0,003	0,009	0,043
28							

Figura 20.4.2
Modello di regressione con 7 variabili esplicative.

	A	B	C	D	E	F	G
1	OUTPUT RIEPILOGO						
2							
3	*Statistica della regressione*						
4	R multiplo	0,6993					
5	R al quadrato	0,4891					
6	R al quadrato corretto	0,4885					
7	Errore standard	0,2407					
8	Osservazioni	6400					
9							
10	ANALISI VARIANZA						
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*	
12	Regressione	7	354,34	50,62	874,01	0,00	
13	Residuo	6392	370,20	0,06			
14	Totale	6399	724,54				
15							
16		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
17	Intercetta	0,854	0,034	25,317	0,000	0,788	0,921
18	ETA	-0,001	0,000	-3,991	0,000	-0,002	-0,001
19	ISTRUZ	0,067	0,003	24,776	0,000	0,062	0,073
20	IMPIEGO	0,024	0,000	58,379	0,000	0,023	0,024
21	CELLUL	0,038	0,007	5,436	0,000	0,024	0,051
22	TV	0,231	0,032	7,156	0,000	0,168	0,295
23	VIDEOREG	0,225	0,016	13,716	0,000	0,193	0,257
24	FAX	0,025	0,009	2,919	0,004	0,008	0,042
25							

Verifica

Indicare se le seguenti affermazioni sono vere o false. Le risposte sono su WEB.

		V	F
20.1	In R la funzione per svolgere la regressione multipla è presente nella libreria stats.	☐	☐
20.2	In Excel il menu Analisi dati contiene la procedura per svolgere la regressione lineare multipla.	☐	☐
20.3	In R, per eseguire una regressione lineare multipla bisogna selezionare una funzione diversa dalla regressione lineare semplice.	☐	☐
20.4	In R, la funzione `lm()` non permette il test sui coefficienti del modello.	☐	☐
20.5	In R, per ottenere tutti i principali output della regressione si può utilizzare la funzione `summary()`.	☐	☐
20.6	In Excel, il livello di confidenza per l'intervallo di confidenza dei coefficienti di regressione è prestabilito e non può essere cambiato.	☐	☐
20.7	In ExceL, tutte le variabili esplicative che faranno parte del modello di regressione devono essere collocate sul foglio elettronico in colonne successive.	☐	☐
20.8	In R, con la funzione `lm()` della regressione lineare, non è possibile salvare i residui.	☐	☐
20.9	Nell'output di R relativo alle stime dei coefficienti di regressione vengono riportate anche le stime dei coefficienti standardizzati.	☐	☐
20.10	In Excel non è possibile costruire il grafico di dispersione tra i valori predetti e i residui standardizzati.	☐	☐
20.11	In R, tramite la funzione `plot()`, è possibile ottenere il Q-Q plot.	☐	☐
20.12	In Excel con il modulo Regressione non è possibile ottenere il residuo per ogni unità statistica.	☐	☐
20.13	In Excel, nel modulo Regressione l'opzione *Tracciati dei residui* permette di ottenere grafici dei residui per ogni variabile.	☐	☐
20.14	In R, nella libreria stats è presente un modo automatico per selezionare le variabili del modello.	☐	☐
20.15	In Excel, nel modulo Regressione, è possibile selezionare le variabili del modello attraverso una procedura automatica.	☐	☐
20.16	In R, la funzione `steps()` permette la selezione backward stepwise.	☐	☐
20.17	In R, per la funzione di selezione delle variabili *Steps*, è necessario stabilire un valore critico della statistica test *F*.	☐	☐