INTRODUZIONE A R
Lezione 4

Silvia Bacci*e Silvia Pandolfi†

1 La regressione lineare

1.1 Regressioni lineare semplice

Per applicare il metodo di regressione lineare scegliamo il dataset cars relativo alle velocità (in miglia orarie) di alcune automobili ed al loro spazio di frenata (in piedi - dati degli anni '20):

```
> data(cars)
```

Forniamo una rappresentazione grafica dei dati tramite una nuvola di punti (scatterplot)

```
> plot(cars$speed, cars$dist)
```

Come si può osservare dal grafico c'è una relazione di proporzionalità crescente tra la velocità del veicolo ed il suo spazio di frenata, confermata anche dal coefficiente di correlazione di Bravais-Pearson:

```
> cor(cars$speed, cars$dist)
[1] 0.8068949
```

Per effettuare un regressione dello spazio di frenata rispetto la velocità si utilizza la funzione `lm()`:

```
> reg = lm(cars$dist ~ cars$speed)
> reg

Call:
lm(formula = cars$dist ~ cars$speed)

Coefficients:
(Intercept)   cars$speed
    -17.579        3.932
```

Gli oggetti contenuti in `reg` sono denominati nei seguenti modi:

*silvia.bacci@unipg.it
†pandolfi@stat.unipg.it

```
> names(reg)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

Una più completa ed esauriente sintesi dei risultati della regressione si ottiene col comando `summary()`:

```
Call:
lm(formula = cars$dist ~ cars$speed)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

La retta di regressione stimata è data da:

$$y = -17.579 + 3.932x$$

e può essere visualizzata sul medesimo grafico dei dati osservati in modo da esprimere un giudizio immediato sulla sua bontà di adattamento:

```
> plot(cars$speed, cars$dist)
> abline(reg)
```

La significatività statistica dei due coefficienti di regressione stimati (intercetta e coefficiente angolare) viene valutata tramite il test t: la terza colonna (`t value`) dell'output di `summary(reg)` dà il valore osservato della statistica test e l'ultima colonna (`Pr(>|t|)`) dà il p-value corrispondente. L'ipotesi nulla verificata è la seguente: $H_0 : \beta_h = 0$ con $h = 0$ nel caso dell'intercetta e $h = 1$ nel caso del coefficiente angolare, contro l'ipotesi alternativa $H_1 : \beta_h \neq 0$. Pertanto, valori sufficientemente piccoli del p-value indicano che il coefficiente stimato è significativamente diverso da zero. Nel nostro esempio, si conclude che la velocità dell'automobile ha un effetto significativo sullo spazio di frenata e, più precisamente, per ogni miglio orario in più lo spazio di frenata aumenta mediamente di 4 piedi.

I coefficienti di regressione possono essere estratti da `reg` in uno dei seguenti modi:

```
> coefficients(reg)
> # oppure
> reg$coefficients
(Intercept)  cars$speed
 -17.579095    3.932409
```

I corrispondenti intervalli di confidenza al 95% sono invece ottenuti tramite

```
> confint(reg, level = 0.95)
                 2.5 %    97.5 %
(Intercept) -31.167850 -3.990340
cars$speed    3.096964  4.767853
```

I valori previsti dello spazio di frenata ($\hat{y}_i$) in corrispondenza dei valori osservati delle velocità (x_i, $i = 1, \ldots, n$) si ottengono nel seguente modo:

```
> fitted(reg)
> # oppure:
> reg$fitted.values
```

Invece, le stime dei residui sono date da:

```
> residuals(reg)
> # oppure:
> reg$residuals
```

Per prevedere il valore atteso della Y in funzione di uno specifico valore di x è sufficiente sostituire il valore della variabile esplicativa nella retta di regressione, con l'accortezza di aggiungere un 1 per l'intercetta. Ad es., lo spazio di frenata medio per un veicolo che procede a 30 miglia l'ora è dato da 100.39 piedi:

```
> x_new = c(1,30)
> yhat_new = x_new%*%coefficients(reg)
> yhat_new
         [,1]
[1,] 100.3932
```

Infine, una misura sintetica della bontà di adattamento della retta ai dati è data dal coefficiente di determinazione R^2, che indica la quota di varianza totale della Y spiegata dalla retta stimata. Questo indice è "contenuto" all'interno di `summary(reg)`, per cui basta semplicemente richiamarlo con il simbolo del dollaro:

```
> summary(reg)$r.squared
[1] 0.6510794d
```

1.2 Analisi dei residui

Sulla base dei valori previsti delle Y e dei residui stimati è possibile ottenere delle rappresentazioni grafiche utili per valutare il rispetto delle ipotesi di base del modello di regressione lineare classico.

```
# Grafico dei valori stimati  verso i valori osservati delle Y
plot(reg$fitted.values, cars$dist)
# Grafico dei valori stimati delle Y verso i residui stimati
plot(reg$fitted.values, reg$residuals, xlab = "Fitted values", ylab = "Residuals")
# Normal Probability Plot
qqnorm(reg$residuals, ylab = 'Residuals')
qqline(reg$residuals)
```

```
# Oppure:
par(mfrow=c(2,2))
plot(reg)
```

Affinchè il modello lineare sia un buon modello per i dati oggetto di studio, dovrei osservare i seguenti andamenti:

- nel grafico `plot(reg$fitted.values, cars$dist)` i punti dovrebbero disporsi sulla bisettrice del primo e terzo quadrante;
- nel grafico `plot(reg$fitted.values, reg$residuals, xlab = "Fitted values", ylab = "Residuals")` i punti dovrebbero equidistribuirsi sopra e sotto la retta $y = 0$ con variabilità costante;
- nel Normal Probability Plot i punti dovrebbero giacere sulla retta.

1.3 Regressione lineare multipla

Nella pratica dell'analisi dei fenomeni collettivi è usuale studiare più di due variabili simultaneamente e le relazioni che intercorrono tra esse. Per questo scopo uno strumento spesso utile è rappresentato dalla regressione lineare multipla, che non è altro che una logica estensione della regressione lineare semplice. Dal punto di vista dell'implementazione in `R`, non è prevista nessuna novità sostanziale, essendo anch'essa basata sulla funzione `lm()`.

Si consideri il dataset `covariate_useless.txt`, che le osservazioni su 100 famiglie relativamente a reddito, numero di figli, metri quadri dell'abitazione e spesa alimentare.

```
> dati <- read.table("covariate_useless.txt", header = TRUE, sep = "\t")
> attach(dati)
> head(dati)
  reddito figli metri  spesa
1  20.458     3    74 10.751
2  21.518     0    50  6.487
3  20.667     0    50  6.267
4  21.247     2    88 10.035
5  35.547     3    56 16.110
6  40.077     0    50  8.407
> str(dati)
'data.frame': 100 obs. of  4 variables:
 $ reddito: num  20.5 21.5 20.7 21.2 35.5 ...
 $ figli  : int  3 0 0 2 3 0 0 1 1 2 ...
 $ metri  : int  74 50 50 88 56 50 50 50 51 87 ...
 $ spesa  : num  10.75 6.49 6.27 10.04 16.11 ...
```

Adesso stimiamo un primo modello di regressione lineare che spieghi la spesa in funzione del solo reddito e poi aggiungiamo le variabili `figli` e `metri`.

```
> mod1 = lm(spesa ~ reddito)
> summary(mod1)
Call:
```

```
lm(formula = spesa ~ reddito)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9672 -2.0540  0.2695  1.7259  4.4951

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.93504    0.80384   6.139 1.77e-08 ***
reddito      0.20329    0.02514   8.087 1.66e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.416 on 98 degrees of freedom
Multiple R-squared:  0.4002,Adjusted R-squared:  0.3941
F-statistic: 65.39 on 1 and 98 DF,  p-value: 1.657e-12

> mod2 = lm(spesa ~ reddito + figli)
> summary(mod2)
Call:
lm(formula = spesa ~ reddito + figli)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96163 -0.49748  0.03782  0.48178  1.65137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.947288   0.267644   7.276 8.97e-11 ***
reddito     0.202603   0.007786  26.023  < 2e-16 ***
figli       2.019169   0.066401  30.409  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7483 on 97 degrees of freedom
Multiple R-squared:  0.9431,Adjusted R-squared:  0.9419
F-statistic: 803.2 on 2 and 97 DF,  p-value: < 2.2e-16

> mod3 = lm(spesa ~ reddito + figli + metri)
> summary(mod3)
Call:
lm(formula = spesa ~ reddito + figli + metri)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96117 -0.53267  0.04849  0.47407  1.65851

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.212657   0.478503   4.624 1.17e-05 ***
reddito      0.202209   0.007830  25.825  < 2e-16 ***
figli        2.070632   0.101665  20.367  < 2e-16 ***
metri       -0.005110   0.007629  -0.670    0.505
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7504 on 96 degrees of freedom
Multiple R-squared:  0.9433,Adjusted R-squared:  0.9416
F-statistic: 532.6 on 3 and 96 DF,  p-value: < 2.2e-16
```

Nota che, delle tre covariate inserite, `metri` non risulta statisticamente significativa, essendo il corrispondente p-value pari a 0.505. Nonostante ciò l'R^2 multiplo risulta leggermente superiore in `mod3` rispetto a `mod2`, diversamente dall'R^2 corretto:

```
cbind("R-sq" =  c(summary(mod1)$r.squared, summary(mod2)$r.squared,
+ summary(mod3)$r.squared),  "Adj-R-sq" = c(summary(mod1)$adj.r.squared,
+ summary(mod2)$adj.r.squared, summary(mod3)$adj.r.squared))
          R-sq  Adj-R-sq
[1,] 0.4002215 0.3941013
[2,] 0.9430565 0.9418824
[3,] 0.9433214 0.9415502
```

In conclusione, la covariata `metri` non è necessaria per spiegare l'andamento della spesa alimentare.

1.3.1 Variabili esplicative qualitative e test Anova

Spesso una o più variabili esplicative sono costituite da fattori, cioè da variabili qualitative con almeno due modalità. Quando le modalità sono soltanto due, non si presentano particolari problemi in sede di stima del modello, salvo tenere presente che il modello è interpretabile soltanto quando la covariata binaria assume uno dei due valori possibili. Quando le modalità sono più di due, invece, occorre fare attenzione alla natura della variabile: se questa non è memorizzata come fattore, infatti, verrà considerata come variabile quantitativa, con conseguenze sui coefficienti di regressione stimati e sull'interpretazione dei risultati della regressione. Vediamo un esempio.

```
> dati2 = read.table("var_qualitative.txt", header = TRUE)
> str(dati2)
’data.frame’: 500 obs. of  4 variables:
 $ eta          : int  24 44 43 41 67 46 46 55 47 38 ...
 $ sesso        : Factor w/ 2 levels "F","M": 1 1 2 2 1 1 1 1 1 2 ...
 $ titolo_studio: Factor w/ 4 levels "Diploma","Laurea",..: 3 1 3 1 3 1 4 3 4 1 ...
 $ reddito      : num  24.2 42.8 41.7 41.4 55.9 ...
```

Come si può osservare sia `sesso` sia `titolo_studio` sono riconosciute come fattori con 2 e 4 livelli, rispettivamente. Qualora `titolo_studio` non fosse stata memorizzata come fattore sarebbe necessario trasformarla usando la funzione `as.factor`:

```
> # titolo_studio = as.factor(dati2$titolo_studio)
```

Adesso stimo il modello che spiega il reddito in funzione delle tre covariate:

```
> mod1 = lm(dati2$reddito ~ dati2$eta+dati2$sesso+dati2$titolo_studio)
> summary(mod1)

Call:
lm(formula = dati2$reddito ~ dati2$eta + dati2$sesso + dati2$titolo_studio)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5597 -1.3567 -0.0939  1.3125  5.7759

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      12.33739    0.30251  40.783  < 2e-16 ***
dati2$eta                         0.70030    0.00517 135.458  < 2e-16 ***
dati2$sessoM                      1.88446    0.17848  10.558  < 2e-16 ***
dati2$titolo_studioLaurea         2.80431    0.26645  10.525  < 2e-16 ***
dati2$titolo_studioLicenza_Media -1.94484    0.23085  -8.425 3.97e-16 ***
dati2$titolo_studioMaster         7.72731    0.23698  32.608  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.98 on 494 degrees of freedom
Multiple R-squared:  0.9761,Adjusted R-squared:  0.9758
F-statistic:  4030 on 5 and 494 DF,  p-value: < 2.2e-16
```

Nota che, benchè le covariate inserite nel modello siano tre (più l'intercetta), i coefficienti di regressione stimati sono cinque (più l'intercetta): infatti, per il fattore a quattro livelli `titolo_studio` vengono create automaticamente 4 variabili dummy e la prima (`Diploma`) viene presa per default come categoria di riferimento. Sulla base dei coefficienti stimati è evidente il vantaggio di possedere una laurea o un master rispetto ad un diploma di scuola superiore e lo svantaggio di possedere al più la licenza di scuola dell'obbligo, sempre rispetto ad un diploma di scuola superiore.

Qualora di voglia modificare la categoria di riferimento occorre riordinare i livelli dei fattori; una possibilità è offerta dalla funzione `relevel`:

```
> ?relevel
> mod1b = lm(dati2$reddito ~ dati2$eta+dati2$sesso+relevel(dati2$titolo_studio, ref=2))
> summary(mod1b)

Call:
lm(formula = dati2$reddito ~ dati2$eta + dati2$sesso + relevel(dati2$titolo_studio,
    ref = 2))

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-5.5597 -1.3567 -0.0939  1.3125  5.7759

Coefficients:
                                                    Estimate Std. Error t value
(Intercept)                                         15.14170    0.36541   41.44
dati2$eta                                            0.70030    0.00517  135.46
dati2$sessoM                                         1.88446    0.17848   10.56
relevel(dati2$titolo_studio, ref = 2)Diploma        -2.80431    0.26645  -10.53
relevel(dati2$titolo_studio, ref = 2)Licenza_Media -4.74915    0.28719  -16.54
relevel(dati2$titolo_studio, ref = 2)Master          4.92300    0.29345   16.78
                                                    Pr(>|t|)
(Intercept)                                          <2e-16 ***
dati2$eta                                            <2e-16 ***
dati2$sessoM                                         <2e-16 ***
relevel(dati2$titolo_studio, ref = 2)Diploma         <2e-16 ***
relevel(dati2$titolo_studio, ref = 2)Licenza_Media   <2e-16 ***
relevel(dati2$titolo_studio, ref = 2)Master          <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.98 on 494 degrees of freedom
Multiple R-squared:  0.9761,Adjusted R-squared:  0.9758
F-statistic:  4030 on 5 and 494 DF,  p-value: < 2.2e-16
```

In quest'ultimo caso i coefficienti stimati per la variabile `titolo_studio` indicano l'effetto differenziale sul reddito di un certo titolo di studio (cioè diploma di scuola superiore, al più scuola dell'obbligo e master) rispetto alla laurea.

Per comprendere meglio l'importanza di trattare le variabili categoriali come fattori e non come variabili quantitative, vediamo che cosa succede se alla covariata `titolo_studio` venissero assegnati dei valori numerici arbitrari.

```
# assegno degli scores arbitrari al titolo di studio
> q1 = as.numeric(dati2$titolo_studio)
> is.factor(q1)
[1] FALSE

# stimo il modello con titolo_studio inserito tramite scores
> mod2 = lm(dati2$reddito ~ dati2$eta+dati2$sesso+q1)
> summary(mod2)
Call:
lm(formula = dati2$reddito ~ dati2$eta + dati2$sesso + q1)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5252  -2.1336   0.5046   2.3277   8.0666

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.149801   0.592861  17.120  < 2e-16 ***
```

```
dati2$eta     0.701056   0.009119  76.879  < 2e-16 ***
dati2$sessoM  1.567365   0.314983   4.976 8.97e-07 ***
q1            1.726046   0.132014  13.075  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.507 on 496 degrees of freedom
Multiple R-squared:  0.9247,Adjusted R-squared:  0.9242
F-statistic:  2029 on 3 and 496 DF,  p-value: < 2.2e-16
```

Confrontando i modelli `mod1`, cioè il modello con titolo di studio inserito come fattore, e `mod2`, cioè il modello con titolo di studio inserito come variabile numerica, si osserva che:

- `mod1` si adatta meglio di `mod2`

  ```
  > summary(mod1)$adj.r.squared
  [1] 0.9758291
  > summary(mod2)$adj.r.squared
  [1] 0.9241955
  ```

- `mod1` consente una più chiara interpretazione dell'effetto delle singole modalità di `titolo_studio` sulla variabile risposta `reddito`;
- in `mod2` l'effetto della licenza media (`q1=3`) rispetto al diploma risulta positivo, mentre in `mod1` è negativo, come è più ragionevole attendersi;
- in mod2 l'intercetta non è intepretabile, perchè `q1=0` non ha senso.

Un problema comune con i fattori si incontra quando una o alcune (ma non tutte) le modalità del fattore non sono significativamente diverse da zero: in una tale situazione come è possibile affermare che la variabile, considerata nel suo insieme, è statisticamente significativa? In tali casi è necessario ricorrere al test d'ipotesi di analisi della varianza che consente, appunto, di verificare se i valori medi di una certa variabile Y si differenziano in modo significativo al variare dei livelli di un certo fattore. La funzione da utilizzare è la funzione `anova`.

```
#> mod1 = lm(dati2$reddito ~ dati2$eta+dati2$sesso+dati2$titolo_studio)
> # test d'ipotesi congiunta del fattore (quando i livelli sono > 2)
> # con il test F verifico che la variabile titolo_studio sia
> # complessivamente significativa
> anova(mod1)
Analysis of Variance Table

Response: dati2$reddito
                    Df Sum Sq Mean Sq   F value    Pr(>F)
dati2$eta            1  72461   72461 18478.566 < 2.2e-16 ***
dati2$sesso          1    292     292    74.399 < 2.2e-16 ***
dati2$titolo_studio  3   6265    2088   532.557 < 2.2e-16 ***
```

```
Residuals           494   1937       4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> # adesso provo ad inserire un’interazione tra sesso e titolo
> mod2 = lm(dati2$reddito ~ dati2$eta + dati2$sesso + dati2$titolo_studio
+ dati2$sesso:dati2$titolo_studio)
> anova(mod2)
Analysis of Variance Table

Response: dati2$reddito
                                Df Sum Sq Mean Sq    F value Pr(>F)
dati2$eta                        1  72461   72461 18382.1527 <2e-16 ***
dati2$sesso                      1    292     292    74.0106 <2e-16 ***
dati2$titolo_studio              3   6265    2088   529.7785 <2e-16 ***
dati2$sesso:dati2$titolo_studio  3      2       1     0.1408 0.9355
Residuals                      491   1935       4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> mod3 = lm( dati2$reddito ~ dati2$eta + dati2$sesso +
dati2$titolo_studio + dati2$eta:dati2$titolo_studio)
> summary(mod3)


Call:
lm(formula = dati2$reddito ~ dati2$eta + dati2$sesso + dati2$titolo_studio +
    dati2$eta:dati2$titolo_studio)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6250 -1.3476 -0.0492  1.3480  5.9261

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              11.967320   0.449990  26.595  < 2e-16
dati2$eta                                 0.707522   0.008314  85.096  < 2e-16
dati2$sessoM                              1.909494   0.180033  10.606  < 2e-16
dati2$titolo_studioLaurea                 3.057156   0.862578   3.544 0.000431
dati2$titolo_studioLicenza_Media         -1.189339   0.727252  -1.635 0.102609
dati2$titolo_studioMaster                 8.368134   0.703725  11.891  < 2e-16
dati2$eta:dati2$titolo_studioLaurea      -0.005279   0.015601  -0.338 0.735209
dati2$eta:dati2$titolo_studioLicenza_Media -0.015222   0.013882  -1.096 0.273411
dati2$eta:dati2$titolo_studioMaster      -0.013026   0.013536  -0.962 0.336368

(Intercept)                              ***
```

```
dati2$eta                                    ***
dati2$sessoM                                 ***
dati2$titolo_studioLaurea                    ***
dati2$titolo_studioLicenza_Media
dati2$titolo_studioMaster                    ***
dati2$eta:dati2$titolo_studioLaurea
dati2$eta:dati2$titolo_studioLicenza_Media
dati2$eta:dati2$titolo_studioMaster
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.983 on 491 degrees of freedom
Multiple R-squared:  0.9761,Adjusted R-squared:  0.9758
F-statistic:  2512 on 8 and 491 DF,  p-value: < 2.2e-16

> anova(mod3)
Analysis of Variance Table

Response: dati2$reddito
                              Df Sum Sq Mean Sq    F value Pr(>F)
dati2$eta                      1  72461   72461 18425.5350 <2e-16 ***
dati2$sesso                    1    292     292    74.1853 <2e-16 ***
dati2$titolo_studio            3   6265    2088   531.0288 <2e-16 ***
dati2$eta:dati2$titolo_studio  3      6       2     0.5274 0.6636
Residuals                    491   1931       4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> mod0 = lm( dati2$reddito ~  dati2$eta +  dati2$sesso)

> anova(mod0, mod1, mod3)
Analysis of Variance Table

Model 1: dati2$reddito ~ dati2$eta + dati2$sesso
Model 2: dati2$reddito ~ dati2$eta + dati2$sesso + dati2$titolo_studio
Model 3: dati2$reddito ~ dati2$eta + dati2$sesso + dati2$titolo_studio +
    dati2$eta:dati2$titolo_studio
  Res.Df    RSS Df Sum of Sq        F Pr(>F)
1    497 8202.2
2    494 1937.1  3    6265.0 531.0288 <2e-16 ***
3    491 1930.9  3       6.2   0.5274 0.6636
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1.3.2 Test F sulle combinazioni lineari dei coefficienti di regressione

Infine, un problema più generale consiste nel verificare un'ipotesi su una combinazione lineare qualsiasi di coefficienti di regressione. A questo proposito si utilizza la funzione

`linearHypothesis` del pacchetto `car`.

```
> library(car)
> help(linearHypothesis)
```

Supponiamo di voler verificare l'ipotesi che, nel modello `mod1`, l'effetto della laurea e l'effetto della licenza media sul reddito siano tra loro uguali. In altre parole voglio verificare l'ipotesi nulla $H_0: \beta_{\text{Laurea}} - \beta_{\text{LicMedia}} = 0$:

```
> # verifico l'ipotesi che beta_Laurea - beta_LicMedia = 0
> # Specifico i vincoli sui 6 coefficienti di regressione del modello mod1
> Hyp = rbind(c(0,0,0,1,-1,0))
> # Specifico il valore della combinazione lineare sotto l'ipotesi nulla
> Hyp.val = 0
> linearHypothesis(mod1, hypothesis.matrix = Hyp, rhs = Hyp.val)
Linear hypothesis test

Hypothesis:
dati2$titolo_studioLaurea - dati2$titolo_studioLicenza_Media = 0

Model 1: restricted model
Model 2: dati2$reddito ~ dati2$eta + dati2$sesso + dati2$titolo_studio

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    495 3009.4
2    494 1937.2  1    1072.3 273.45 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Come si può osservare, la funzione `linearHypothesis` pone a confronto il modello stimato con un modello ristretto, specificato dall'ipotesi nulla. Un valore elevato di F e, quindi, un valore piccolo del p-value portano a rifiutare l'ipotesi nulla H_0 di adeguatezza del modello ristretto in favore del modello più generale: pertanto, si rifiuta l'ipotesi che l'effetto della laurea e della licenza media sul reddito siano uguali tra loro.

Altri esempi:

```
# verifico l'ipotesi che beta_Laurea + beta_LicMedia = 0
Hyp = rbind(c(0,0,0,1,1,0))
Hyp.val = 0
linearHypothesis(mod1, hypothesis.matrix = Hyp, rhs = Hyp.val)
# verifico l'ipotesi che beta_Intercetta = 10
Hyp = rbind(c(1,0,0,0,0,0))
Hyp.val = 10
linearHypothesis(mod1, hypothesis.matrix = Hyp, rhs = Hyp.val)
```

1.3.3 Metodi di selezione delle variabili esplicative

Un problema comune della stima di un modello di regressione lineare multipla consiste nello scegliere una strategia ragionevole per la selezione delle covariate statisticamente

significative. Tra le strategie più note si ricordano la backward selection e la forward selection.

Con la backward selection si inseriscono tutte le covariate disponibili nel modello e poi si procede eliminandone una alla volta, partendo da quella con il p-value più elevato, purchè superiore ad una soglia prefissata (ad es., 5%). Dopo ogni eliminazione si ristima il modello con le covariate residue e si procede eliminando una nuova covariata. Ci si ferma quando tutte le covariate rimaste risultano statisticamente significative.

Con la forward selection si procede in modo speculare, inserendo una variabile alla volta nel modello, a partire da quella con p-value più piccolo, purchè inferiore ad una soglia prefissata (ad es., 5%) e R^2 corretto più elevato. Al passo successivo, si stima il modello con la covariata aggiunta e si procede inserendo una nuova covariata. Ci si ferma quando tutte le covariate rimaste escluse risultano statisticamente non significative.

Qualora tra le variabili esplicative ci siano uno o più fattori con oltre due livelli, l'eliminazione (procedura backward) o l'inserimento (procedura forward) di ciascun fattore nel modello deve essere valutato per tutti i livelli contemporaneamente: è, pertanto, necessario ricorrere ad un test F di tipo Anova (vedi sopra) che verifichi l'ipotesi nulla che tutti i coefficienti di regressione relativi allo stesso fattore siano uguali a 0.

Vediamo adesso come si eseguono con le due procedure.

```
> data = read.table("covariates_useless2.txt", header = TRUE)
> head(data)
> str(data)

> # separo le variabili
> spesa = data$spesa      # spesa alimentare
> reddito = data$reddito    # reddito
> figli = data$figli    # numero figli
> metri = data$metri    # metri quadri abitazione
> sesso = data$sesso   # sesso del capofamiglia (1=M)
> genitori = data$genitori   # dummy per entrambi i genitori che lavorano (1=si)
> zona = data$zona    #  zona di abitazione (1=Urbana)
```

Iniziamo con la procedura backward selection:

```
# Stimo il modello completo (con tutte le covariate)
> mod = lm(spesa ~ reddito + figli + metri + sesso + genitori + zona)
> summary(mod)

> # Siccome la variabile con il p-value più elevato (e > 0.05) è sesso,
> # elimino sesso e stimo il modello senza sesso.
> mod1 = lm(spesa ~ reddito + figli + metri + genitori + zona)
> summary(mod1)

> # Adesso, il p-value più grande (e maggiore di 0.05) è quello della
> covariata metri.  Quindi, elimino metri.
> mod2 = lm(spesa ~ reddito + figli + genitori + zona)
> summary(mod2)
```

A questo punto tutte le covariate inserite in `mod2` risultano statisticamente significative al 5%. Pertanto, non ci sono altri candidati all'eliminazione. Il modello definitivo è `mod2`:

```
lm(spesa ~ reddito + zona+ figli +genitori)
```

Qualora il livello di significatività sia fissato all'1%, anche la variabile genitori andrebbe eliminata.

Adesso vediamo come si sviluppa la procedura forward.

```
> # Stimo un modello per ciascuna covariata candidata alla selezione
> mod1 = lm(spesa ~ reddito)
> mod2 = lm(spesa ~ figli)
> mod3 = lm(spesa ~ metri)
> mod4 = lm(spesa ~ sesso)
> mod5 = lm(spesa ~ genitori)
> mod6 = lm(spesa ~ zona)
> summary(mod1)
> summary(mod2)
> summary(mod3)
> summary(mod4)
> summary(mod5)
> summary(mod6)
```

Sia reddito che zona presentano i p-value più piccoli, inoltre reddito presenta un R^2 aggiustato più elevato. Quindi aggiungo reddito.

```
> # Stimo un modello (con la variabile reddito)
> #per ciascuna altra variabile candidata alla selezione
> mod11 = lm(spesa ~ reddito + figli)
> mod21 = lm(spesa ~ reddito + metri)
> mod31 = lm(spesa ~ reddito + sesso)
> mod41 = lm(spesa ~ reddito + genitori)
> mod51 = lm(spesa ~ reddito + zona)
> summary(mod11)
> summary(mod21)
> summary(mod31)
> summary(mod41)
> summary(mod51)
```

La variabile zona presenta il p-value più piccolo e l'R^2 aggiustato più elevato. Quindi aggiungo zona.

```
> mod12 = lm(spesa ~ reddito + zona+figli)
> mod22 = lm(spesa ~ reddito + zona+metri)
> mod32 = lm(spesa ~ reddito + zona+sesso)
> mod42 = lm(spesa ~ reddito + zona+genitori)
> summary(mod12)
> summary(mod22)
> summary(mod32)
> summary(mod42)
```

La variabile figli presenta il p-value più piccolo e l'R^2 aggiustato più elevato. Quindi aggiungo figli.

```
> mod13 = lm(spesa ~ reddito + zona+ figli +metri)
> mod23 = lm(spesa ~ reddito + zona+ figli +sesso)
> mod33 = lm(spesa ~ reddito + zona+ figli +genitori)
> summary(mod13)
> summary(mod23)
> summary(mod33)
```

La variabile genitori è significativa al 5%. Inserisco genitori. Nessuna altra variabile è significativa. Quindi il modello selezionato è:

```
lm(spesa ~ reddito + zona+ figli +genitori)
```

Il modello selezionato è lo stesso della procedura backward, ma tale risultato è tutt'altro che scontato.

In R è possibile eseguire le procedure backward e forward in maniera automatica, tramite la funzione `regsubsets` del pacchetto `leaps`. Ad esempio, nel caso della procedura backward si procede nel seguente modo (per la procedura forward basta sostituire `method=backward` con `method=forward`):

```
> library(leaps)
> mod_back = regsubsets(spesa ~ reddito + figli + metri + sesso + genitori + zona,
+ data=data, method = "backward")
> summary(mod_back)
Subset selection object
Call: regsubsets.formula(spesa ~ reddito + figli + metri + sesso +
    genitori + zona, data = data, method = "backward")
6 Variables  (and intercept)
         Forced in Forced out
reddito      FALSE      FALSE
figli        FALSE      FALSE
metri        FALSE      FALSE
sesso        FALSE      FALSE
genitori     FALSE      FALSE
zona         FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: backward
         reddito figli metri sesso genitori zona
1  ( 1 ) "*"     " "   " "   " "   " "      " "
2  ( 1 ) "*"     " "   " "   " "   " "      "*"
3  ( 1 ) "*"     "*"   " "   " "   " "      "*"
4  ( 1 ) "*"     "*"   " "   " "   "*"      "*"
5  ( 1 ) "*"     "*"   "*"   " "   "*"      "*"
6  ( 1 ) "*"     "*"   "*"   "*"   "*"      "*"
```

Nell'output della funzione `regsubsets` viene riportata una tabella con una riga per ogni possibile modello che può essere costruito con le covariate indicate (nel nostro caso, abbiamo 6 possibili modelli) e, per ciascuno, vengono indicate con "*" le covariate che vengono selezionate. Ad es., se vogliamo un modello con 4 covariate, allora le covariate selezionate sono `reddito`, `figli`, `genitori`, `zona`. Come si può osservare sia l'ordine di uscita (o

ingresso) delle covariate sia il risultato finale a parità di numero di covariate è uguale a quello ottenuto sopra con la procedura "manuale". Tuttavia, la funzione `regsubsets` non consente di visualizzare il livello di probabilità in corrispondenza del quale una certa variabile esce dal modello. Per compensare questa mancanza di informazione, si possono calcolare i valori dell'R^2 aggiustato per ciascuno dei 6 modelli possibili e selezionare il modello in corrispondenza del quale si presenta l'ultimo "salto" dell'R^2:

```
>  summary(mod_back)$adjr2
[1] 0.7702616 0.9401186 0.9995132 0.9995152 0.9995148 0.9995144
```

Nel nostro esempio, il modello selezionato è il numero 3, costituito dalle covariate `reddito`, `figli`, `zona`.

1.3.4 Diagnosi della multicollinearità

Oltre all'analisi dei residui già vista a proposito della regressione lineare semplice, nel caso di regressione linear multipla è importante procedere ad un'analisi della multicollinearità, cioè della presenza nel modello di covariate altamente correlate le une alle altre. Si ricorda che la multicollinearità è un problema più sottile rispetto alla perfetta collinearità, che si riscontra quando una covariata è ottenibile da un'altra tramite una trasformazione lineare. Il caso della perfetta collinearità porta ad un modello non identificabile e il software dà sempre un messaggio di errore, consentendo una rapida diagnosi del problema e una soluzione immediata (basta eliminare dal modello una delle due covariate problematiche). La multicollinearità, invece, non è facilmente diagnosticabile se non specificamente indagata e, se ignorata, porta a stime inaffidabili e instabili.

In genere, è ragionevole sospettare la presenza di multicollinearità tra una coppia di covariate quando:

- la correlazione tra le due variabili è molto elevata;
- l'R^2 corretto del modello di regressione è molto elevato, ma numerosi coefficienti di regressione risultano statisticamente non significativi;
- eliminando una covariata dal modello, un'altra covariata, prima non significativa, diventa significativa.

Vediamo il seguente esempio.

```
> multi = read.table("multicollinear.txt",header=T)
> str(multi)
'data.frame': 100 obs. of  4 variables:
 $ x1: num  -1.505 -0.0677 0.1954 0.4551 -0.4586 ...
 $ x2: num  0.425 -0.878 0.499 -1.939 0.501 ...
 $ x3: num  0.46 -0.905 0.432 -1.996 0.545 ...
 $ y : num  0.327 -0.311 1.898 -2.553 1.31 ...
> attach(multi)

> # Calcolo i coefficienti di correlazione tra le coppie di variabili esplicative
> cor(x2, x3)
[1] 0.9982573
> cor(x1, x3)
```

```
[1] -0.01804989
> cor(x1, x2)
[1] -0.0161666
> # Indizio num. 1: correlazione tra x2 e x3 molto elevata

> # Stimo il modello di regressione multipla
> mod = lm(y ~ x1+x2+x3)
> summary(mod)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32854 -0.30499 -0.00849  0.28972  1.13196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.90045    0.05106  17.635   <2e-16 ***
x1           1.03253    0.04622  22.338   <2e-16 ***
x2           1.38902    0.89827   1.546    0.125
x3           0.57242    0.90226   0.634    0.527
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.499 on 96 degrees of freedom
Multiple R-squared:  0.9504,Adjusted R-squared:  0.9488
F-statistic: 612.7 on 3 and 96 DF,  p-value: < 2.2e-16
> # Indizio num. 2: R-quadro aggiustato del modello molto elevato
> # e due coeff. di regress. su tre non sono significativamente diversi da zero

> # Stimo il modello con due sole covariate
>  mod = lm(y ~ x1+x2)
>  summary(mod)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32669 -0.31419 -0.01282  0.25676  1.14822

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.89582    0.05038   17.78   <2e-16 ***
x1           1.03158    0.04606   22.40   <2e-16 ***
x2           1.95792    0.05283   37.06   <2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4974 on 97 degrees of freedom
Multiple R-squared:  0.9502,Adjusted R-squared:  0.9491
F-statistic: 924.5 on 2 and 97 DF,  p-value: < 2.2e-16
> # Indizio num. 3: togliendo x3, x2 diventa significativa
```

Una misura sintetica usata per fornire una diagnosi più oggettiva della multicollinearità è rappresentata dal *Variance Inflaction Factor* (VIF), che viene definito come:

$$VIF_j = \frac{1}{1 - R_j^2},$$

dove j indica la regressione della covariata x_j rispetto a tutte le altre covariate inserite nel modello. Valori elevati di R_j^2 e, quindi, di VIF_j denotano un'elevata correlazione tra x_j e le altre covariate; in generale, un $VIF > 5$ denota un problema di multicollinearità rilevante.

```
> # Regredisco x1 su x2 e x3
> mod = lm(x1 ~ x2+x3)
> r2_1=summary(mod)$adj.r.squared
> VIF1 = 1/(1-r2_1)
> VIF1
[1] 0.9810837

> # Regredisco x2 su x1 e x3
> mod = lm(x2 ~ x1+x3)
> r2_2=summary(mod)$adj.r.squared
> VIF2 = 1/(1-r2_2)
> VIF2
[1] 281.6376

> # Regredisco x3 su x1 e x2
> mod = lm(x3 ~ x2+x1)
> r2_3=summary(mod)$adj.r.squared
> VIF3 = 1/(1-r2_3)
> VIF3
[1] 281.6557
```

Concludo che sussiste multicollinearità, causata dalla compresenza nel modello delle variabili `x2` e `x3`.