Laboratorio 6

Regressione multipla

6.1 Analisi del dataset HOOK.DAT

I dati contenuti nel file `hook.dat`, raccolti da Joseph Hooker sulle montagne dell'Himalaya, riportano le temperature di ebollizione dell'acqua (in gradi Farenheit) relative a diversi valori della pressione atmosferica. Si vuole studiare la relazione tra le due variabili.

```
> hook <- read.table("I:\\modelli\\hook.dat",
                     col.names=c("temp", "press"))
> hook
    temp  press
 1 210.8 29.211
 2 210.2 28.559
 ...
30 181.0 15.919
31 180.6 15.376

> attach(hook)
```

I dati appaiono ordinati per temperature decrescenti. Al decrescere della temperatura anche la pressione decresce. Questo suggerisce l'esistenza di un legame tra le variabili.
Per esplorare graficamente la relazione analizziamo il diagramma di dispersione.

```
> plot(temp~press)
```

Il grafico mostra una evidente relazione tra le due variabili. Proviamo quindi ad adattare un modello di regressione lineare semplice.

```
> fit <- lm(temp~press)
> summary(fit)

Call:
lm(formula = temp ~ press)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.6735 -0.6805  0.2203  0.5296  1.3976

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 146.67290    0.77641  188.91   <2e-16 ***
press         2.25260    0.03809   59.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.806 on 29 degrees of freedom
Multiple R-Squared: 0.9918,     Adjusted R-squared: 0.9915
F-statistic:  3498 on 1 and 29 DF,  p-value:     0
```

Appare evidente che entrambi i coefficienti sono fortemente significativi. Dato il risultato osservato sul coefficiente angolare, non sorprende il risultato del test F per la bontà complessiva del modello.
Il quadrato del coefficiente di correlazione (`Multiple R-Squared:  0.9918`) ci dice che il 99% della variabilità della temperatura è spiegata dalla sua relazione lineare con la pressione.
Possiamo aggiungere la retta stimata nel diagramma di dispersione precedente con il comando:

```
> abline(coef(fit))
```

Passiamo ora all'analisi dei residui.

```
> plot(fit)
```

Il grafico dei residui rispetto ai valori adattati indica un chiaro andamento parabolico dei residui. Questo ci dice che i valori bassi e alti previsti dal modello sono sistematicamente sovrastimati, mentre i valori centrali sono sottostimati.
In maniera meno evidente, le lacune del modello possono essere viste anche dal grafico dei valori osservati sui valori stimati.

```
> plot(temp, fitted(fit))
> abline(0,1)
```

Il `qqnorm` dei residui mostra qualche scostamento dalla normalità per i residui di segno positivo. Da questo potremmo desumere che i residui sono un po' asimmetrici. Possiamo provare a vederlo anche con i soliti strumenti:

```
> res <- rstandard(fit)
> hist(res)
> boxplot(res)
```

In conclusione, l'analisi dei residui non appare soddisfacente, nonostante i risultati ottenuti nei test di significatività.
Come possiamo rimediare?
Il commento fatto per il grafico dei residui rispetto ai valori adattati ci fa capire che deve esserci una relazione tra residui e pressione. Proviamo a fare un diagramma a dispersione.

```
> plot(res~press)
```

Effettivamente, il grafico mostra una relazione di tipo quadratico. Questo significa che anche tra la temperatura e la pressione può esistere una relazione di tipo quadratico.
Quindi il modello potrebbe essere migliorato introducendo un termine quadratico come ulteriore variabile esplicativa. In questo modo passiamo da una regressione semplice ad una regressione multipla di tipo polinomiale.

```
> fit1 <- lm(temp ~ press + I(press^2))
> summary(fit1)

Call:
lm(formula = temp ~ press + I(press^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.79906 -0.26314 -0.01578  0.25139  0.73891

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.701623   2.112363  59.981  < 2e-16 ***
press         4.157627   0.199069  20.885  < 2e-16 ***
I(press^2)   -0.043754   0.004552  -9.612 2.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3956 on 28 degrees of freedom
Multiple R-Squared: 0.9981,     Adjusted R-squared: 0.998
F-statistic:  7307 on 2 and 28 DF,  p-value:     0
```

La funzione `I( )` permette di specificare nella formula del modello che si vuole considerare `press` al quadrato come una variabile esplicativa. Il modello contenuto in `fit1` si poteva ottenere a partire dal precedente modello (`fit`), con il comando

```
> fit1 <- update(fit, .~. + I(press^2))
```

Questo comando permette di aggiornare il modello corrente (`fit`), cambiandone la formula. In questo caso, si è lasciato `temp` come variabile risposta (si noti la presenza del punto a sinistra di `~`) e si è aggiunta l'esplicativa `press` al quadrato.
I test rilevano forte significatività dei coefficienti e della bontà complessiva del modello.
Passiamo a controllare i residui del modello.

```
> par(mfrow=c(2,2))
> plot(fit1)
```

Le analisi dei residui appaiono notevolmente migliorate e completamente soddisfacenti.
Pertanto la relazione tra temperatura di ebollizione e pressione può essere ben descritta da una funzione quadratica.

```
> detach()
```

6.2 Analisi del dataset CHERRY.DAT

Riconsideriamo i dati contenuti nel file `cherry.dat`, riferiti a misurazioni rilevate su alberi di ciliegio. Le variabili sono rispettivamente: diametro (a 4.5 piedi dal suolo, misurato in pollici), altezza (in piedi), volume (in piedi cubici di legname utile). Si vuole studiare la dipendenza del volume di legno utile dall'altezza e dal diametro dell'albero.

```
> cherry <- read.table("I:\\modelli\\cherry.dat",
                       col.names=c("diametro","altezza","volume"))
> cherry
   diametro altezza volume
1       8.3      70   10.3
2       8.6      65   10.3
....
31     20.6      87   77.0

> attach(cherry)
```

Cominciamo con l'analisi grafica preliminare.

```
> par(mfrow=c(2,1))
> plot(volume~diametro)
> plot(volume~altezza)
> cor(diametro, volume)
[1] 0.9671194
> cor(altezza, volume)
[1] 0.5982497
```

Si osserva una dipendenza del volume dal diametro. La relazione con l'altezza non è chiara.

```
> par(mfrow=c(1,1))
> plot(diametro, altezza)
> cor(diametro, altezza)
[1] 0.5192801
```

Anche fra diametro ed altezza c'è una leggera correlazione, come ci si aspetta visto il significato delle variabili. Il significato delle variabili che si stanno trattando in realtà suggerisce di utilizzare un altro tipo di modello. Sembra infatti logico aspettarsi che il volume di legname dipenda in qualche modo dal prodotto delle altre due variabili. Conviene allora trasformare i dati.

```
> plot(log(diametro), log(volume))
> plot(log(altezza), log(volume))
> cor(log(diametro), log(volume))
[1] 0.976665
> cor(log(altezza), log(volume))
[1] 0.648638
```

Le cose sono migliorate. Vediamo come il modello si adatta ai dati.

```
> fit <- lm(log(volume) ~ log(diametro) + log(altezza))
> summary(fit)

Call:
lm(formula = log(volume) ~ log(diametro) + log(altezza))

Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.63162    0.79979  -8.292 5.06e-09 ***
log(diametro)  1.98265    0.07501  26.432  < 2e-16 ***
log(altezza)   1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-Squared: 0.9777,     Adjusted R-squared: 0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value:     0
```

Il modello sembra buono. Si può passare all'analisi dei residui.

```
> res <- rstandard(fit)
> par(mfrow=c(2,1))
> plot(log(diametro), res)
> plot(log(altezza), res)
> par(mfrow=c(2,2))
> plot(fit)
```

I grafici sembrano evidenziare la presenza di alcuni dati anomali. Ad esempio, la 18-esima osservazione. Si può ripetere l'analisi eliminando tale osservazione.

```
> fit1 <- lm(log(volume) ~ log(diametro) + log(altezza),
             subset=-c(18))
> summary(fit1)

Call:
lm(formula = log(volume) ~ log(diametro) + log(altezza),
   subset = -c(18))

Residuals:
     Min       1Q   Median       3Q      Max
-0.17500 -0.05706  0.00624  0.05940  0.11383

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.18549    0.78061  -9.205 8.14e-10 ***
log(diametro)  1.95816    0.07051  27.770  < 2e-16 ***
log(altezza)   1.26100    0.19984   6.310 9.39e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07565 on 27 degrees of freedom
Multiple R-Squared: 0.9814,     Adjusted R-squared:  0.98
F-statistic: 712.3 on 2 and 27 DF,  p-value:     0
```

L'opzione `subset` permette di eseguire l'analisi su uno specificato sottoinsieme di osservazioni. In questo caso si è indicato di prendere tutte le osservazioni, tranne la 18-esima.
Passiamo all'analisi dei residui.

```
> par(mfrow=c(2,2))
> plot(fit1)
```

Esercizio: Ripetere l'analisi escludendo le altre osservazioni anomale.