Laboratorio 10

Analisi della covarianza

10.1 Analisi del dataset CATS.DAT

I dati contenuti nel file `cats.dat` presentano il peso del corpo ed il peso del cuore di alcuni gatti di sesso femminile (1) e maschile (2). Si vuole verificare se esistono differenze, in media, nel peso del cuore tra i due sessi, tenendo presente però la relazione che esiste tra peso del cuore e peso del corpo.
Leggiamo i dati.

```
> cats <- read.table("I:\\modelli\\cats.dat", col.names=c("B","H","S"))
> cats
     B    H S
 1 2.3  9.6 1
 2 3.0 10.6 1
 .....
46 2.7 10.4 2
47 3.2 11.6 2
48 3.0 10.6 2
```

Dobbiamo dichiarare il sesso come fattore.

```
> cats$S <- factor(cats$S)
> attach(cats)
```

In via preliminare, confrontiamo la distribuzione del peso del cuore nei due sessi.

```
> plot(H~S)
```

Il grafico mostra una chiara differenza tra i due sessi. Il peso del cuore nelle femmine è, infatti, mediamente più basso.
Se volessimo verificare l'esistenza di una differenza in media tra i due gruppi, potremmo utilizzare il test t di Student, previa verifica della normalità ed omoschedasticità delle due distribuzioni.

```
> par(mfrow=c(1,2), pty="s")
> qqnorm(H[S==1])
> qqline(H[S==1])
> qqnorm(H[S==2])
> qqline(H[S==2])
```

La distribuzione del peso del cuore nelle femmine devia, sulle code, dalla normalità (questo era da attendersi visto il boxplot), mentre l'ipotesi di normalità pare più che accettabile per i maschi.
Per quanto riguarda l'omoschedasticità:

```
> var.test(H[S==1], H[S==2])

        F test to compare two variances

data:  H[S == 1] and H[S == 2]
F = 0.4799, num df = 23, denom df = 23, p-value = 0.08496
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2076113 1.1094088
sample estimates:
ratio of variances
         0.4799227
```

L'ipotesi è accettabile, per cui procediamo con il test nonostante le incertezze sulla normalità della distribuzione delle femmine. Vista la natura del confronto, possiamo utilizzare il test t ad alternativa unilaterale:

$$\begin{aligned} \mathrm{H}_0 &: \mu_F = \mu_M \\ \mathrm{H}_1 &: \mu_F < \mu_M \,, \end{aligned}$$

dove μ_F indica la media della popolazione dei gatti di sesso femminile e μ_M la media della popolazione dei gatti di sesso maschile. Il test t è ottenuto da

```
> t.test(H[S==1], H[S==2], alternative="less", var.equal=T)

        Two Sample t-test

data:  H[S == 1] and H[S == 2]
t = -4.8419, df = 46, p-value = 7.455e-06
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.412780
sample estimates:
mean of x mean of y
   8.8875   11.0500
```

Come era da attendersi, il test rifiuta l'ipotesi H_0.
Nella formulazione del problema si chiedeva però di tenere conto nel confronto della relazione esistente tra peso del corpo e peso del cuore.
Proviamo quindi a vedere che relazione esiste tra le due variabili.

```
> par(mfrow=c(1,1))
> plot(H~B)
```

Il grafico mostra una chiara relazione tra peso del cuore e del corpo: come è logico attendersi, all'aumentare del peso del corpo aumenta anche il peso del cuore.
Proviamo a vedere la distribuzione del peso del corpo nei due sessi.

```
> plot(B~S)
```

Come immaginabile, il peso del corpo dei gatti maschi è mediamente più alto di quello delle femmine. Considerata allora la relazione tra peso del corpo e quello del cuore sorge il sospetto che le differenze osservate nel peso medio del cuore tra i due gruppi siano in realtà dovute alla dipendenza tra peso del corpo e quello del cuore, più che all'appartenenza ai due sessi.

```
> par(mfrow=c(1,3), pty="s")
> plot(B[S==1], H[S==1], pch=1)
> plot(B[S==2], H[S==2] ,pch=2, col=2)
> plot(H~B, type="n")
> points(B[S==1], H[S==1], pch=1)
> points(B[S==2], H[S==2], pch=2, col=2)
```

Per verificare l'ipotesi prima formulata, dovremmo confrontare il peso del cuore di maschi e femmine che hanno un uguale peso del corpo. Possiamo allora modellare la relazione tra `H` e `B` nei due gruppi:

$$\begin{aligned} \mu_M &= \alpha_M + \beta_M B \\ \mu_F &= \alpha_F + \beta_F B \end{aligned}$$

e verificare se i coefficienti delle relazioni lineari sono uguali, ovvero

$$H_0\text{: } \alpha_M = \alpha_F,\ \beta_M = \beta_F,$$

Per condurre l'analisi della covarianza è possibile utilizzare la funzione `lm()`. Costruiamo quindi un modello lineare che spieghi il peso del cuore (`H`) in funzione del peso del corpo (`B`) e del sesso (`S`).
Per stimare questo modello, è necessario codificare la variabile sesso come una variabile numerica. `R` effettua questa conversione automaticamente in vari modi. Noi utilizzeremo l'opzione di default. Per vedere come `R` codifica il fattore sesso quando la deve trattare numericamente, è possibile fare:

```
> contrasts(S)
  2
1 0
2 1
```

Questo ci dice che `R` associa al livello 1 (femmine) il valore 0 ed al livello 2 (maschi) il valore 1. Quindi la colonna sesso viene trattata come un vettore del tipo `c(0,0,...,1,1)`.
Adattiamo allora il seguente modello:

```
> fit <- lm(H ~ B + S + B:S)
```

Il modello stimato con la precedente chiamata ad `lm()` è del tipo:

$$y_i = \beta_0 + \beta_1 B_i + \beta_2 S_i + \beta_3 B_i S_i + \varepsilon_i \,.$$

Un'istruzione equivalente alla precedente è:

```
> fit <- lm(H ~ B*S)
```

che include automaticamente gli effetti marginali (B e S) e l'effetto dell'interazione (B:S).
Vediamo il risultato dell'analisi.

```
> summary(fit)

Call:
lm(formula = H ~ B + S + B:S)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9813 -0.9589 -0.1629  0.8573  2.6277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9318     2.1105   1.389  0.17178
B             2.5525     0.8975   2.844  0.00674 **
S2           -0.2849     3.0313  -0.094  0.92554
B:S2          0.4177     1.1784   0.354  0.72466
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.28 on 44 degrees of freedom
Multiple R-Squared: 0.5664,     Adjusted R-squared: 0.5368
F-statistic: 19.16 on 3 and 44 DF,  p-value: 4.269e-08
```

I risultati dei test sulla significatività dei coefficienti ci suggeriscono che i parametri β_2 e β_3 possono essere eliminati. Questo significa che le relazioni lineari nei due sessi hanno uguale intercetta e uguale coefficiente angolare. In definitiva, non esiste differenza nel peso medio del cuore nei due sessi se si tiene conto del peso del corpo. Notiamo che anche l'intercetta β_1 non risulta significativa.
Costruiamo il modello finale che meglio spiega il peso del cuore in funzione del peso del corpo.

```
> fit1 <-  lm(H ~ B - 1)
> summary(fit1)

Call:
lm(formula = H ~ B - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0372 -0.9521 -0.0969  0.8620  3.0732

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
B   3.8507     0.0714   53.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.292 on 47 degrees of freedom
Multiple R-Squared: 0.9841,     Adjusted R-squared: 0.9838
F-statistic:  2909 on 1 and 47 DF,  p-value:     0

> plot(fit1)
```

Il modello pare abbastanza buono.

```
detach()
```

10.2 Analisi del dataset INSULATE.DAT

I dati riportati nel dataset `insulate.dat` sono stati raccolti dal proprietario di una casa per valutare l'effetto dell'isolamento termico della costruzione sul consumo di gas per riscaldamento. Il termometro dell'impianto è stato regolato a 20 gradi Celsius e sono stati registrati:

`Temp`: la temperatura media esterna, misurata in gradi Celsius;
`Cons`: il consumo settimanale di gas (in 1000 ft^3).

Queste misurazioni sono state condotte per 26 settimane prima e per 30 settimane dopo l'installazione del sistema di isolamento termico.

1. Qual è la relazione tra temperature esterna e consumo di gas?
2. La eventuale relazione varia tra prima e dopo la coibentazione?

```
> insulate <- read.table("I:\\modelli\\insulate.dat",
+                        col.names=c("quando", "temp", "cons"))
> insulate
   quando  temp cons
 1  prima  -0.8  7.2
```

```
 ...
56   dopo   9.7  1.5

> attach(insulate)
```

Il quesito 1 ci chiede di esplorare la relazione esistente tra temperatura e consumo. Per farci una idea della relazione, iniziamo con qualche analisi grafica.

```
> plot(cons~temp)
```

Si nota una evidente relazione decrescente tra temperatura e consumo: all'aumentare della temperatura calano i consumi.
Proviamo a vedere se questa relazione è diversa prima e dopo l'isolamento termico (quesito 2):

```
> plot(cons~temp, type="n")
> points(temp[quando=="prima"], cons[quando=="prima"], pch=1)
> points(temp[quando=="dopo"], cons[quando=="dopo"], pch=2, col=2)
```

Si nota chiaramente come la relazione rimanga decrescente, ma i livelli di consumo si abbassino, a parità di temperatura, dopo l'isolamento termico. Questo parrebbe suggerire, per spiegare il legame tra temperatura e consumo prima e dopo l'isolamento, un modello che preveda due rette di regressione con diversa intercetta. Non è chiaro se le due rette debbano avere anche diverso coefficiente angolare.
Procediamo allora con l'analisi della covarianza.

```
> fit <- lm(cons ~ temp + quando + temp:quando)
> summary(fit)

Call:
lm(formula = cons ~ temp + quando + temp:quando)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.18011  0.03757  0.20930  0.63803

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.72385    0.11810  40.000  < 2e-16 ***
temp             -0.27793    0.02292 -12.124  < 2e-16 ***
quandoprima       2.12998    0.18009  11.827 2.22e-16 ***
temp:quandoprima -0.11530    0.03211  -3.591  0.00073 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-Squared: 0.9277,     Adjusted R-squared: 0.9235
F-statistic: 222.3 on 3 and 52 DF,  p-value:     0
```

Tutti i parametri risultano significativi. Pertanto non solo il livello medio di consumo cambia prima e dopo la coibentazione, ma cambia anche la forza del legame tra consumo e temperatura. Il test F conferma la validità del modello. Questo risponde al quesito 2.

Esercizio: eseguire e commentare l'analisi dei residui del modello.

Possiamo aggiungere le rette stimate nel grafico precedente con i comandi:

```
> abline(4.72385+2.12998, -0.27793-0.11530)
> abline(4.72385, -0.27793, col=2)
```

Infatti, tenendo presente che:

```
> contrasts(quando)
      prima
dopo      0
prima     1
```

le due regressioni risultanti sono le seguenti:

prima della coibentazione: $(4.72385+2.12998) - (0.27793+0.11530) * \text{temp}$
dopo la coibentazione: $4.72385 - 0.27793 * \text{temp}$

In questo caso, la stima delle due rette di regressione ottenuta mediante l'analisi della covarianza coincide con la stima di due rette di regressione semplice per i due gruppi di dati separatamente:

```
> fitprima <- lm(cons~temp, subset=(quando=="prima"))
> summary(fitprima)

Call:
lm(formula = cons ~ temp, subset = (quando == "prima"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.62020 -0.19947  0.06068  0.16770  0.59778

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.85383    0.11842   57.88  < 2e-16 ***
temp        -0.39324    0.01959  -20.08 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2813 on 24 degrees of freedom
Multiple R-Squared: 0.9438,     Adjusted R-squared: 0.9415
```

```
F-statistic: 403.1 on 1 and 24 DF,  p-value: 1.11e-16

> fitdopo <- lm(cons~temp, subset=(quando=="dopo"))
> summary(fitdopo)

Call:
lm(formula = cons ~ temp, subset = (quando == "dopo"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.11082  0.02672  0.25294  0.63803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974   36.41  < 2e-16 ***
temp        -0.27793    0.02518  -11.04 1.05e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,     Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
```

Il primo approccio offre, tuttavia, il vantaggio di poter condurre test sull'uguaglianza dei coefficienti nei 2 modelli di regressione.
Per riottenere il grafico con le due rette stimate possiamo procedere in questo modo:

```
> plot(cons~temp, type="n")
> points(temp[quando=="prima"], cons[quando=="prima"], pch=1)
> points(temp[quando=="dopo"], cons[quando=="dopo"], pch=2, col=2)
> abline(fitprima$coeff)
> abline(fitdopo$coeff, col=2)

> detach()
```