# *R lezione #2: analisi di regressione multivariata e interazioni fattoriali*

*Nicola Romanò*

---

## *Introduzione*

L'anno scorso abbiamo parlato di modelli lineari e del loro uso per eseguire la regressione e l'analisi della varianza (ANOVA). Abbiamo considerato solo situazioni semplici con una variabile indipendente che influenza la variabile di uscita (ANOVA a una via) o due fattori (ANOVA a due vie) che non interagiscono tra loro. Nelle lezioni abbiamo parlato delle interazioni e di come cambiano la nostra interpretazione dei modelli lineari. In questo seminario daremo un'occhiata a come affrontare le interazioni in R.

## *Obiettivi formativi*

Dopo aver completato questo seminario sarai in grado di:

- Utilizzare modelli lineari per eseguire regressione multipla e analisi della varianza con più fattori
- Interpretare l'output di un modello lineare
- Confrontare due modelli per scegliere quello che si adatta meglio ai dati
- Interpretare i risultati della tua analisi in presenza di interazioni

## *Sezione 1 - Un ripasso dei modelli lineari*

Iniziamo questo seminario con un piccolo ripasso dei modelli lineari. Un modello lineare è un modello statistico che mette in relazione le variazioni di una variabile dipendente ($Y$) con le variazioni di una o più variabili indipendenti ($X_1, X_2, ..., X_n$).

L'equazione generale per tale modello è:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$

Dove:

- $Y$ sono le nostre misure
- $X_1, ..., X_n$ sono i fattori (o predittori) che influenzano $Y$. In genere sono le altre variabili nel set di dati, o le loro trasformazioni/combinazioni [1].
- $\beta_1, ...\beta_n$ sono i coefficienti di regressione, fattori di scala per i predittori.

[1] Ad esempio, potremmo aver raccolto il peso dei soggetti nel nostro studio, ma usare *log(peso)* come predittore per il nostro modello. Oppure potremmo aver raccolto due valori diversi e utilizzare il loro rapporto come parametro del modello

- $\epsilon$ è l'errore, o residuo. Rappresenta la differenza tra ciò che è spiegato dalla previsione del modello e ciò che abbiamo osservato. Include l'effetto di tutti i fattori che non abbiamo misurato nella nostra configurazione sperimentale, così come gli errori di misurazione. Generalmente assumiamo che sia distribuito normalmente [2].

Quando usiamo R (o qualsiasi altro software!) Per generare il modello, ciò che fa è stimare i coefficienti $\beta$ in modo tale da minimizzare l'errore [3].

In questa formula, ciascun predittore agisce indipendentemente dagli altri. In altre parole, se abbiamo due predittori, $X_1$ e $X_2$, l'effetto di $X_1$ su $Y$ sarà sempre lo stesso, indipendentemente dal valore di $X_2$. Ciò non è sempre il caso.

### *Regressione lineare semplice*

Come primo esempio consideriamo il set di dati `pressure-workshop2.csv`. In questo studio è stato studiato l'effetto di un farmaco sulla riduzione della pressione arteriosa (misurata in mmHg) su 150 pazienti di età, peso (in kg) e sesso diversi.

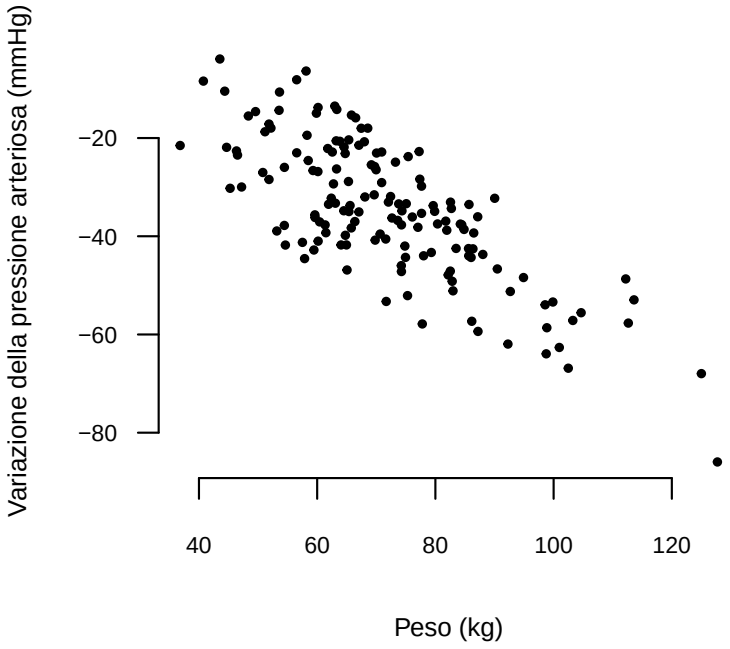
Inizia a familiarizzare con i dati. Quanti uomini e donne ci sono? Quale gamma di età e peso? Traccia le varie variabili l'una contro l'altra e vedi se emergono particolari modelli [4].

Tralasciamo per un momento le altre variabili e concentriamoci sulla relazione tra peso e risposta; sembra che l'effetto maggiore si riscontri nei pazienti più pesanti.

[2] Si noti che sebbene avere residui normalmente distribuiti rende le cose più semplici ... si può ancora avere un modello valido e utilizzabile quando i residui non sono normalmente distribuiti, specialmente se la deviazione dalla normalità è piccola. Per lo più, si riduce all'osservazione critica dei dati e dell'esperienza. Inoltre, parlare con uno statistico è sempre una buona idea!

[3] Nel caso di `lm`, questo è chiamato il metodo dei minimi quadrati. Nelle statistiche di libri e pubblicazioni è possibile visualizzare i parametri stimati indicati come $\hat{\beta}$ (altrimenti detti "beta hat"). Questo per indicare che è il risultato di una stima, ovvero un'approssimazione del vero valore di $\beta$ per la popolazione, che rimane sconosciuta. Possiamo ottenere intervalli di confidenza per queste stime usando `confint(model)`

[4] Se non ricordi come farlo, vedi la Lezione 1.




Possiamo usare un modello lineare per verificare se esiste una tale relazione.

Come sempre, iniziamo affermando la nostra ipotesi nulla _____

Ti ricordi come eseguire una regressione lineare in R? Prova, se non ricordi, leggi la pagina seguente!

```
model <- lm(Response ~ Weight, data = pressure)
```

Questo genera il modello

$Risposta = \beta_0 + \beta_1 * Peso + \epsilon$

Quali sono le ipotesi di questo modello? Ti ricordi come verificare che siano soddisfatte? [5]

[5] Direi che in questo caso le ipotesi sono generalmente soddisfatte, cosa ne pensi?

Questo è uno dei più semplici modelli lineari che possiamo generare, in cui il valore del risultato dipende da un singolo parametro. Questa è chiamata *regressione lineare semplice*.

Diamo un'occhiata all'output del modello

```
summary(model)
```

```
##
## Call:
## lm(formula = Response ~ Weight, data = pressure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.190  -6.354   0.874   6.568  19.554
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.78143    3.32703   3.541 0.000533 ***
## Weight      -0.64852    0.04521 -14.346  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.133 on 148 degrees of freedom
## Multiple R-squared:  0.5817, Adjusted R-squared:  0.5789
## F-statistic: 205.8 on 1 and 148 DF,  p-value: < 2.2e-16
```

Il riassunto ci dà molte informazioni.

Prima di tutto, ci dice i parametri $\beta$ (coefficienti) che sono stati stimati dal modello.

$\hat{\beta}_0 = 11.78$ e $\hat{\beta}_1 = -0.65$

Quindi

Risposta $= 11.78 - 0.65 * \text{Peso} + \epsilon$

Ciò significa che per ogni aumento di 1 Kg di peso c'è una diminuzione di 0,65 mmHg nella pressione arteriosa dopo l'assunzione del farmaco. L'effetto del peso sulla risposta al farmaco è statisticamente significativo ($F_{1,148} = 205.8, p = 2 * 10^{-16}$) [6].

[6] R riporta anche un valore p per l'intercetta; questo è il risultato di un test t di un campione che confronta l'intercetta a 0. In altre parole, in questo caso l'intercetta è statisticamente diversa da 0. L'intercetta è il valore corrispondente a un cambiamento nella pressione sanguigna in cui tutti i fattori (in questo caso il peso) sono uguali a zero. Poiché un peso di 0 non è biologicamente significativo, in questo caso possiamo ignorare questo valore

Un altro valore importante è il coefficiente di determinazione ($R^2$). Questo è una misura di quanto è buono il modello, o di quanto il modello spiega della variabilità dei dati. $R^2$ non è un buon modo per confrontare due diversi modelli, poiché dipende dal numero di parametri; cioè, se aggiungiamo un descrittore in più al nostro modello, $R^2$ aumenterà sempre. Per questo motivo, il software riporta anche una versione "corretta" (`adj.R`). In questo caso $adj.R^2 = 0.5789$; questo significa che il nostro modello descrive $\sim 57{,}9\%$ della variabilità nei nostri dati, che è OK ma non eccezionale. Significa che ci sono altri fattori che non abbiamo considerato come responsabili di$> 40\%$ della variabilità! [7] Quindi, quali sono questi altri fattori?

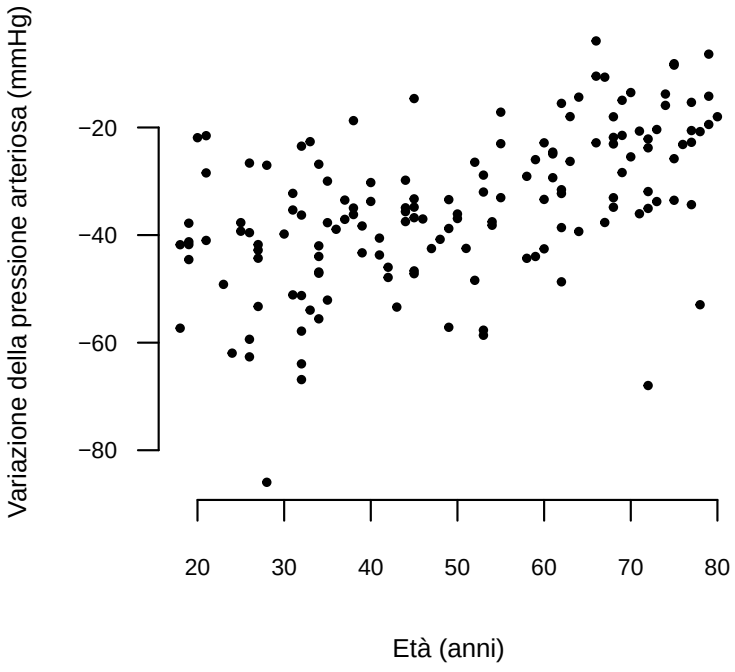
[7] Quale pensi sia il valore massimo di $R^2$? Perché?

*Regressione lineare multipla*

Il nostro set di dati contiene altri due descrittori: `Age` e `Sex`. È biologicamente plausibile che questi fattori possano influenzare la pressione sanguigna, quindi dovremmo aggiungerli al nostro modello [8]. Per semplificare le cose, inizieremo con l'età e considereremo il genere in un secondo momento.

[8] Si noti che, sebbene per semplicità stiamo aggiungendo questi descrittori uno alla volta, in pratica dovremmo probabilmente iniziare da un completo, compresi tutti i descrittori che abbiamo misurato. Questo è il motivo per cui li abbiamo misurati, non è vero?

È utile, a questo punto, tracciare il cambiamento della pressione sanguigna contro l'età.




Vediamo una possibile relazione nella risposta al farmaco a seconda dell'età. Incorporiamo l'età nel nostro modello.

```
model.2 <- lm(Response ~ Weight + Age, data = pressure)
```

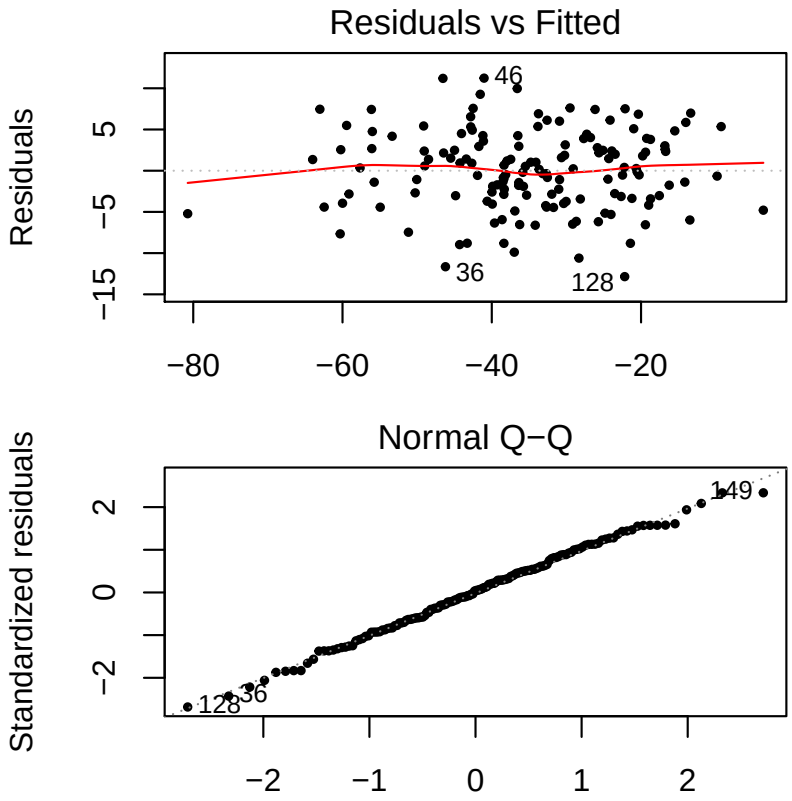
Questo genererà un modello che considera l'effetto del peso e l'effetto dell'età, indipendentemente l'uno dall'altro [9]

Quali sono le ipotesi nulle [10] che questo modello sta testando?

Di nuovo, vogliamo verificare le ipotesi del modello usando i grafici diagnostici.

[9] Questo significa che il modello guarderà l'effetto del peso dell'individuo sulla sua risposta al farmaco, indipendentemente dalla sua età e viceversa.

[10] Ce n'è più d'una !!!




Dal momento che i grafici diagnostici sembrano buoni (varianza uniforme e residui distribuiti normalmente), possiamo continuare con questo modello.

```
summary(model.2)
```

```
##
## Call:
## lm(formula = Response ~ Weight + Age, data = pressure)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8545  -3.3021   0.1972   3.1070  11.2320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.60214    2.04592  -4.205 4.53e-05 ***
## Weight      -0.65779    0.02392 -27.501  < 2e-16 ***
## Age          0.42390    0.02169  19.541  < 2e-16 ***
## ---
```

```
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.832 on 147 degrees of freedom
## Multiple R-squared:  0.8837, Adjusted R-squared:  0.8821
## F-statistic: 558.6 on 2 and 147 DF,  p-value: < 2.2e-16
```

Puoi interpretare il risultato di questo modello proprio come hai fatto per il precedente. Ci dice che c'è un effetto statisticamente significativo del peso ($p < 2 * 10^{-16}$) e dell'età ($p < 2 * 10^{-16}$) sulla risposta al farmaco ($F_{2,147} = 558.6$ [11]). Nota che ora il modello spiega l'88% della variabilità!

[11] Si noti che la statistica F riportata da `summary` si riferisce all'intero modello. Se si desidera conoscere la statistica F per componenti specifici del modello, è possibile eseguire `anova(model.2)`, che ti dà F e DF per i vari descrittori del tuo modello.)

### *Predittori qualitativi e variabili fittizie*

Consideriamo ora il sesso e aggiungiamolo al nostro modello. Traccia i dati, pensi che il sesso influenzi la risposta al farmaco?

In questo caso, abbiamo a che fare con una variabile qualitativa discreta, con due livelli, F e M. Tutto ciò che abbiamo detto finora vale ancora e `lm` è in grado di gestire questo tipo di variabili senza alcun problema. Tuttavia, il modo in cui affrontiamo questo tipo di variabili è leggermente diverso.

```
model.3 <- lm(Response ~ Weight + Age + Sex, data = pressure)
```

Questo è il modello:

$\text{Risposta} = \beta_0 + \beta_1 * \text{Peso} + \beta_2 * \text{Età} + \beta_3 * D$

Introduciamo una nuova variabile $D$, chiamata "variabile fittizia", che codifica Sesso in questo modo:

$$D = \begin{cases} 1, & \text{se Sesso} = M \\ 0, & \text{altrimenti} \end{cases}$$

Di default R assegna 0 al primo livello della variabile (il *livello di riferimento*, in questo caso F) e 1 al secondo [12].

[12] I livelli sono ordinati alfabeticamente; vedere la Lezione 1 per come modificare l'ordinamento dei livelli.

Pertanto, per le osservazioni al livello di riferimento (quindi soggetti di sesso femminile), il terzo termine $\beta_3 * D$ sarà 0; per soggetti di sesso maschile sarà $\beta_3 * 1 = \beta_3$. Quindi $\beta_3$ rappresenta la **differenza tra la risposta di un maschio e una femmina**, mantenendo tutti gli altri fattori costanti.

Diamo un'occhiata al sommario del modello per chiarirlo meglio.

```
summary(model.3)
```

```
##
## Call:
## lm(formula = Response ~ Weight + Age + Sex, data = pressure)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6250  -3.2145   0.1914   3.2024  11.0588
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.36394    2.11449  -3.956 0.000119 ***
## Weight      -0.66298    0.02645 -25.062  < 2e-16 ***
## Age          0.42224    0.02204  19.154  < 2e-16 ***
## SexM         0.41103    0.88467   0.465 0.642898
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.845 on 146 degrees of freedom
## Multiple R-squared:  0.8839, Adjusted R-squared:  0.8815
## F-statistic: 370.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

L'output non è molto diverso da quello che avevamo prima. Ha il sesso un effetto statisticamente significativo sulla risposta? Quale percentuale della varianza viene spiegata da questo modello? [13].

Considera le stime

$\beta_1 = -0.66;\ \beta_2 = 0.42;\ \beta_3 = 0.41$

Ciò significa che:

- Per ogni aumento di 1 kg di peso, la risposta diminuisce di 0,66 mmHg (mantenendo uguale età e sesso)
- Per ogni aumento di 1 anno di età, la risposta aumenta di 0,42 mmHg (mantenendo il peso e il sesso uguali)
- Se il paziente è di sesso maschile, la risposta aumenta di 0,41 mmHg

Quindi, quale prevedi che sia la risposta di un uomo di 50 anni che pesa 82 kg?

### *Variabili fittizie per più livelli*

Potresti aver capito a questo punto, che le variabili fittizie sono ciò che R usa per codificare i gruppi o altri fattori discreti quando eseguiamo un ANOVA!

In alcuni casi, tuttavia, avrai più di due livelli; il ragionamento è lo stesso, tuttavia saranno usate più variabili fittizie per definire i diversi livelli.

[13] Si noti come $R^2$ sia aumentato, sebbene di poco, poiché abbiamo aggiunto un parametro extra; invece $R_{adj}^2$ è diminuito!

Ad esempio, supponiamo di aver misurato i livelli dell'ormone luteinizzante ($LH$) in tre diverse specie di pesci: sgombro, salmone e tonno.

Puoi codificare la variabile *Specie* con due (numero di livelli - 1) variabili fittizie $D_1$ e $D_2$ così:

$$D_1 = \begin{cases} 1, & \text{se Specie = "Salmone"} \\ 0, & \text{altrimenti} \end{cases}$$

$$D_2 = \begin{cases} 1, & \text{se Specie = "Tonno"} \\ 0, & \text{altrimenti} \end{cases}$$

Quindi:

| Specie | $D_1$ | $D_2$ |
|---|---|---|
| Sgombro | 0 | 0 |
| Salmone | 1 | 0 |
| Tonno | 0 | 1 |

Il nostro modello potrebbe essere qualcosa del tipo:
$\text{LH} = \beta_0 + \beta_1 * D_1 + \beta_2 * D_2 + \epsilon$
dove $\beta 1$ rappresenta la differenza tra i livelli di LH nel salmone e nello sgombro e $\beta 2$ la differenza tra i livelli di LH nel tonno e nello sgombro.

## *Sezione 2: scelta di un modello*

Tornando al nostro esempio iniziale, abbiamo tre modelli:

1. $\text{Risposta} = \beta_0 + \beta_1 * \text{Peso} + \epsilon$
2. $\text{Risposta} = \beta_0 + \beta_1 * \text{Peso} + \beta_2 * \text{Età} + \epsilon$
3. $\text{Risposta} = \beta_0 + \beta_1 * \text{Peso} + \beta_2 * \text{Età} + \beta_3 * D_{maschio} + \epsilon$

Potremmo obiettare che il modello #2 è meglio del modello #1, in quanto spiega una percentuale molto più grande della varianza (88% vs 58%), ma che dire del modello #3?

È corretto dire che: poiché il Sesso non ha un effetto statisticamente significativo sulla risposta, e dato che il valore di $R^2_{adj}$ è inferiore (anche se in minima parte), dovremmo eliminare Sesso dal modello, e solo considerare l'età e il peso come predittori? Un modo per decidere è usare la funzione `anova` per confrontare i due modelli. Questo verifica l'ipotesi nulla che il modello più complesso non si adatti ai dati meglio di quello più semplice [14].

[14] Per essere più precisi, ciò che effettivamente fa è verificare se i coefficienti extra stimati nel modello più complesso non sono diversi da 0. Nota che puoi usare questa funzione solo se tutti i parametri del modello più semplice sono presenti anche nel modello più complesso.

```
anova(model, model.2)
```

```
## Analysis of Variance Table
##
## Model 1: Response ~ Weight
## Model 2: Response ~ Weight + Age
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    148 12346.1
## 2    147  3431.7  1    8914.4 381.85 < 2.2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Come previsto, il valore di $p$ è molto basso, il che indica che il secondo modello, più complesso, si adatta meglio ai dati rispetto al primo e dovrebbe essere preferito ad esso. Vedi anche quanto è diminuita la somma residua dei quadrati (RSS), indicando che il modello è molto più vicino ai dati reali (quindi i residui, e la loro somma quadratica, sono minori).

al contrario

```
anova(model.2, model.3)
```

```
## Analysis of Variance Table
##
## Model 1: Response ~ Weight + Age
## Model 2: Response ~ Weight + Age + Sex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    147 3431.7
## 2    146 3426.7  1    5.0665 0.2159 0.6429
```

Il valore di $p$ è 0.64, a indicare che non possiamo confutare l'ipotesi nulla, il che significa che il nostro terzo modello (con età, peso e sesso) non è migliore del modello più semplice.

R fornisce anche una comoda funzione, chiamata `drop1`, che rimuove un predittore alla volta da un modello più grande. Potete vedere che conferma ciò che abbiamo visto sopra. [15]

```
drop1(model.3, test = "F")
```

```
## Single term deletions
##
## Model:
## Response ~ Weight + Age + Sex
##        Df Sum of Sq     RSS    AIC  F value Pr(>F)
## <none>               3426.7 477.31
## Weight  1   14741.9 18168.5 725.52 628.1081 <2e-16 ***
## Age     1    8610.5 12037.2 663.77 366.8706 <2e-16 ***
## Sex     1       5.1  3431.7 475.53   0.2159 0.6429
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[15] Si prega di notare che queste sono solo linee guida molto generali. La scelta di cosa includere nel modello non è facile ed esistono pareri contrastanti sul se sia meglio avere sempre un modello più semplice o più completo e spesso la risposta non è semplice. **La cosa importante è che ogni scelta sia basata su una solida motivazione.** Non fermarti solo al valore di $p$ ... pensa quale domanda ti stai chiedendo e cosa ti dice il tuo modello.

## *Sezione 3 - Interazioni tra fattori*

Nelle lezioni hai imparato a conoscere le interazioni tra fattori nella regressione multipla. Vedremo ora come analizzare le interazioni in R.

Considereremo ora i dati in `fox_workshop2.csv`. Questo set di dati [16] contiene la dimensione della cucciolata, come misura del successo riproduttivo, in due diverse popolazioni di volpi artiche, in relazione alla loro età e posizione, così come la disponibilità dei roditori [17] (che è stato classificata come bassa, media o alta).

[16] I dati per questo esempio non sono reali, ma basati sul lavoro di Tannerfeldt e Angerbjörn (Oikos, 1998)

[17] I roditori sono una fonte di cibo per le volpi artiche, ma il loro numero varia di anno in anno in molte regioni

Figure 1: Sleepy arctic fox - Eric Kilby - CC BY-SA 2.0

Come al solito, leggiamo i dati e iniziamo a esplorarli

```
foxes <- read.csv("fox-workshop2.csv")

summary(foxes)
```

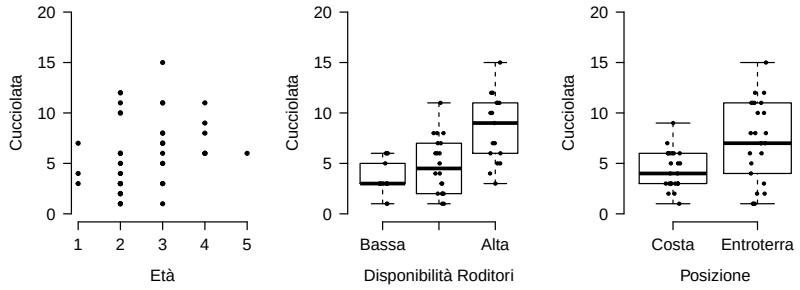
```
##       Age          Location  RodentAvail   LitterSize
##  Min.   :1.00   Coastal:25   High  :17   Min.   : 1.00
##  1st Qu.:2.00   Inland :25   Low   :13   1st Qu.: 3.00
##  Median :2.00                Medium:20   Median : 5.50
##  Mean   :2.54                            Mean   : 5.72
##  3rd Qu.:3.00                            3rd Qu.: 7.75
##  Max.   :5.00                            Max.   :15.00
```

Potresti aver notato che i livelli del fattore `RodentAvail` sono in un ordine un po' insolito [18]. Possiamo riordinarlo in modo che `Low` sia usato come riferimento.

[18] Alfabetico!

```
foxes$RodentAvail <- factor(foxes$RodentAvail,
    levels = c("Low", "Medium", "High"))
```

Possiamo quindi tracciare alcune delle relazioni tra le variabili [19].




Da una prima ispezione, sembra che il successo riproduttivo sia in qualche modo correlato a tutte le altre variabili. Possiamo usare un modello lineare per studiare queste relazioni [20].

Proprio come abbiamo fatto prima, possiamo creare un modello usando `lm`:

```
model <- lm(LitterSize ~ Age + Location + RodentAvail,
    data = foxes)
```

```
summary(model)
```

```
##
## Call:
## lm(formula = LitterSize ~ Age + Location + RodentAvail, data = foxes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.898 -1.729  0.233  1.573  4.250
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.6194     1.1996  -2.184   0.0342 *
## Age                 1.9827     0.3696   5.365 2.71e-06 ***
## LocationInland      1.5697     0.6259   2.508   0.0158 *
## RodentAvailMedium   1.3225     0.7810   1.693   0.0973 .
## RodentAvailHigh     5.8513     0.8489   6.892 1.47e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.043 on 45 degrees of freedom
## Multiple R-squared:  0.6738, Adjusted R-squared:  0.6448
## F-statistic: 23.24 on 4 and 45 DF,  p-value: 1.823e-10
```

Lascerò a te l'interpretazione punto per punto del sommario del modello [21].

Ora una domanda importante: è questo il miglior modello per descrivere i nostri dati? Il valore $R^2_{adj}$ è 0.64, il che significa che il modello spiega solo il 64% della varianza dei dati.

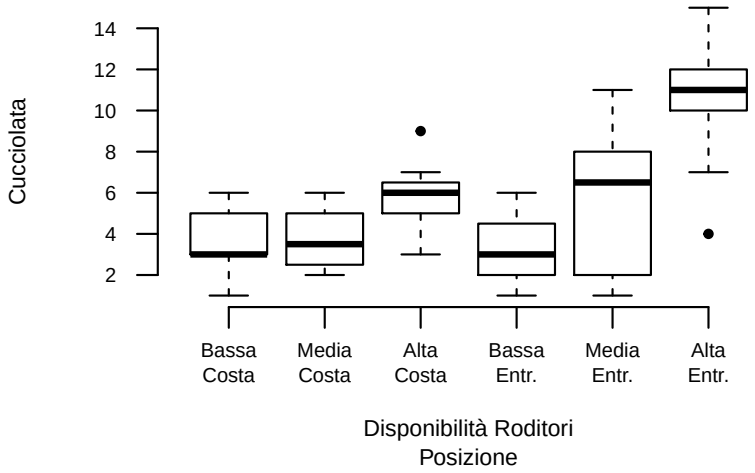
Possiamo migliorare il modello? La risposta dipende in parte dalla domanda che vogliamo affrontare. Ad esempio, una domanda inter-

[19] Sei in grado di riprodurre questi grafici?

[20] La dimensione della cucciolata è un conteggio, quindi il limite inferiore è 0; impareremo più avanti nel corso che un modello lineare non è il modo migliore per analizzare questo tipo di dati limitati, ma per il momento possiamo ignorare questo problema.

[21] Guarda l'intercetta, che cosa significa un valore negativo? Vedi qualche problema?

essante che questo modello non può attualmente rispondere è "La disponibilità dei roditori influenza le dimensioni della cucciolata di entrambe le popolazioni di volpi nello stesso modo?". Questa è una domanda più specifica e potenzialmente più interessante da chiedere, ma leggermente più complessa da rispondere. Possiamo iniziare tornando ai nostri grafici. Consideriamo questo




Questo è molto interessante! Sembra che le due popolazioni non siano uguali quando considerano l'effetto della disponibilità di roditori sul successo riproduttivo! In altre parole, esiste un'interazione tra la posizione e la disponibilità dei roditori nel determinare le dimensioni della cucciolata. Questo non solo è interessante perché ci dà l'opportunità di imparare come analizzare le interazioni in R ... ma anche perché porta altre domande come "perché c'è questa differenza?" [22].

Possiamo modificare il nostro modello per tenerne conto

```
model.2 <- lm(LitterSize ~ Age + Location + RodentAvail +
    RodentAvail:Location, data = foxes)
```

La notazione `RodentAvail: Location` viene utilizzata per indicare l'interazione tra i due fattori. Un modo alternativo e completamente equivalente di indicare le interazioni è usando il segno `*`.

```
model.2 <- lm(LitterSize ~ Age + Location * RodentAvail,
    data = foxes)
```

[22] Ad esempio, una spiegazione potrebbe essere che le aree costiere offrono maggiori quantità di uccelli, che nidificano sulle scogliere della costa. Questi volatili possono essere usati dalle volpi come fonte di cibo alternativa. Quindi le volpi che vivono sulla costa hanno in media una cucciolata più piccola ogni anno, mentre le volpi che vivono nell'entroterra hanno grandi cucciolate negli anni quando c'è molto cibo a disposizione. Potresti progettare un esperimento per verificare questa ipotesi?

```
summary(model.2)
```

```
##
## Call:
## lm(formula = LitterSize ~ Age + Location * RodentAvail, data = foxes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5903 -1.0377 -0.1342  1.0894  2.8658
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -3.0511     0.9171  -3.327 0.001805 **
## Age                                2.4634     0.2883   8.544 8.17e-11 ***
## LocationInland                    -2.6479     1.0468  -2.530 0.015168 *
## RodentAvailMedium                  1.2585     0.7386   1.704 0.095597 .
## RodentAvailHigh                    3.2777     0.7648   4.286 0.000101 ***
## LocationInland:RodentAvailMedium   2.9610     1.2328   2.402 0.020707 *
## LocationInland:RodentAvailHigh     7.5482     1.3035   5.791 7.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.533 on 43 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.8002
## F-statistic:  33.7 on 6 and 43 DF,  p-value: 1.019e-14
```

Questo modello è leggermente più complesso di quello precedente, tuttavia può essere interpretato in modo molto simile. Vediamo che il modello ora spiega l'80% della variabilità nei nostri dati, un miglioramento rispetto al modello precedente [23]! Il modello ci dice anche che c'è un effetto significativo dell'età sulle dimensioni della cucciolata. $\hat{\beta}_{Age}$ ci dice che per ogni aumento di 1 anno di età c'è un aumento di 2.4 nelle dimensioni della cucciolata [24]. Ci dice anche che ci sono interazioni significative.

Possiamo visualizzare queste interazioni usando un *interaction plot*, come quello fornito dalla funzione `emmip` nel pacchetto`emmeans` [25].

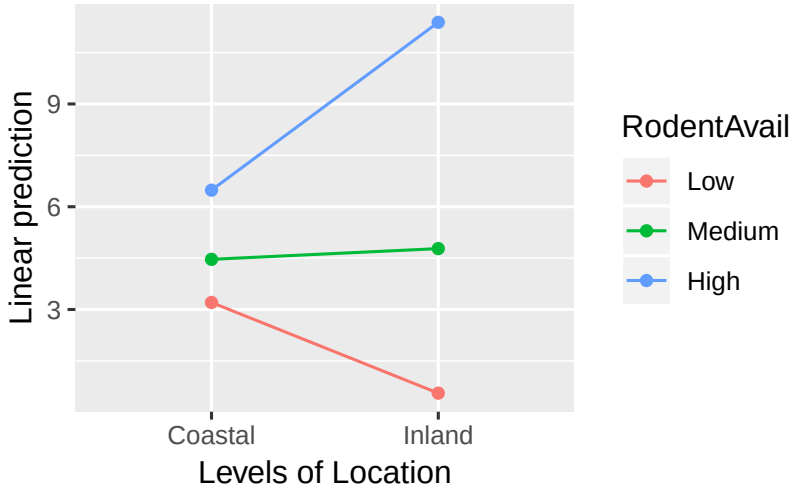
```
par(mar = c(4, 4, 1, 4), cex = 0.5, cex.lab = 1,
    cex.axis = 1)

library(emmeans)  # Devi avere installato emmeans!
emmip(model.2, RodentAvail ~ Location)
```

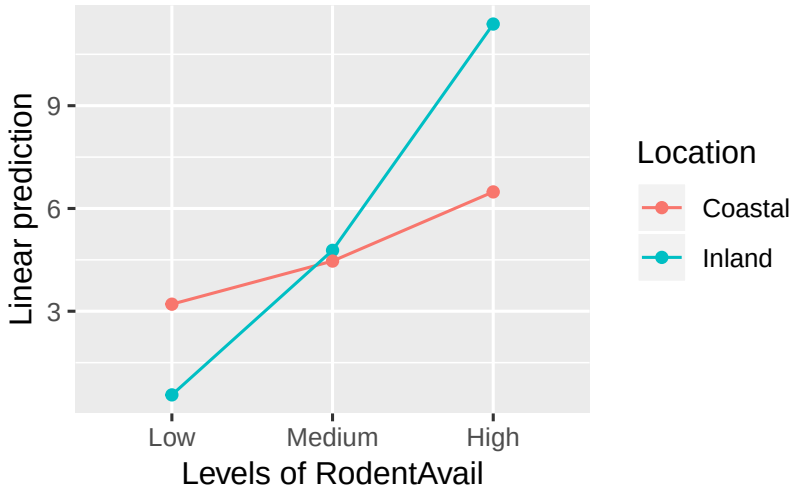
[23] Puoi anche confrontare questo modello con quello precedente usando la funzione `anova`

[24] Dobbiamo fare molta attenzione quando interpretiamo questi coefficienti. Il modello non ci dice perché gli animali più vecchi hanno cucciolate più grandi. Può essere direttamente a causa dell'età, o perché gli animali più vecchi hanno già avuto cucciolate, e questo ha un effetto sulle dimensioni della figliata!

[25] Se non hai installato il pacchetto `emmeans`, puoi farlo usando `install.packages("emmeans")`. R fornisce anche un'altra funzione, chiamata `interaction.plot`, che può produrre lo stesso grafico)




```
emmip(model.2, Location ~ RodentAvail)
```




Questi grafici mostrano le *medie marginali stimate*, ovvero le medie stimate dal nostro modello per ogni livello dei fattori che stiamo considerando.

Entrambi i grafici mostrano la stessa informazione; poiché le linee non sono parallele tra loro, possiamo dire che esiste un'interazione tra i due fattori. La disponibilità di roditori influenza la popolazione nell'entroterra più di quanto influenzi la popolazione costiera.

E cosa dovremmo fare delle stime per la posizione e la disponibilità dei roditori [26]? Poiché abbiamo un'interazione significativa, questi coefficienti diventano leggermente meno utili. $\hat{\beta}_2$ è la differenza media tra le dimensioni della cucciolata tra le volpi che vivono nell'entroterra e quelle che vivono sulla costa, **indipendentemente dalla disponibilità dei roditori** [27]. Tuttavia, poiché la disponibilità dei roditori influenza questa differenza ... ignoreremmo questi due coefficienti

[26] Cioé $\hat{\beta}_2 = 2.46$ e $\hat{\beta}_3 = -2.65$

[27] Come puoi interpretare $\hat{\beta}_3$?

quando interpretiamo il nostro modello. Più formalmente, in presenza di interazioni, generalmente ignoriamo gli effetti principali (quindi gli effetti indipendenti di ciascuno dei due fattori che interagiscono nell'intero campione).

Infine, diciamo che vogliamo sapere se esiste una differenza statistica tra la popolazione costiera e quella dell'entroterra, ai diversi livelli di disponibilità dei roditori. Possiamo usare le funzioni `emmeans` e `pair` per farlo. Queste funzioni possono eseguire il confronto a coppie (proprio come il test di Tukey), tenendo anche conto delle interazioni. Piuttosto che confrontare tutti i possibili livelli, possiamo specificare differenze specifiche (chiamate anche contrasti) a cui siamo interessati; questo eviterà confronti, come ad esempio Entroterra/Alta Disponibilità verso Costa/Bassa Disponibilità, che non ci forniscono particolari informazioni biologiche [28].

[28] Ancora, nel caso volessimo guardare a tutti i possibili confronti ... potremmo usare `emmeans(model.2, pairwise ~ Location * RodentAvail)`. Ciò restituirà tutti i possibili confronti, senza dover usare `pairs`

```
marginals <- emmeans(model.2, ~Location * RodentAvail)
pairs(marginals, by = "RodentAvail")
```

```
## RodentAvail = Low:
##  contrast          estimate        SE df t.ratio p.value
##  Coastal - Inland  2.647908 1.0467996 43   2.530  0.0152
##
## RodentAvail = Medium:
##  contrast          estimate        SE df t.ratio p.value
##  Coastal - Inland -0.313044 0.7195572 43  -0.435  0.6657
##
## RodentAvail = High:
##  contrast          estimate        SE df t.ratio p.value
##  Coastal - Inland -4.900337 0.7572678 43  -6.471  <.0001
```

Nelle chiamata a `pairs`, si specifica che vogliamo confrontare la posizione a diversi livelli di disponibilità dei roditori. L'output ci dà la stima della differenza (ad esempio: per basse disponibilità di roditori, le volpi costiere hanno in media 2.6 cuccioli in più rispetto alle volpi interne) e l'errore standard associato a questa stima [29]. Abbiamo anche un valore $p$ per ciascuno dei contrasti. Ricorda, anche se il valore $p$ ci dice che le due condizioni sono diverse, probabilmente è più interessante osservare una certa misura della dimensione dell'effetto (come la stima), che può spiegare il significato biologico del risultato. Traccia i dati, guarda il numero e pensa! Una differenza di 0.01 cuccioli con un valore $p$ di 0.02 sarebbe interessante dal punto di vista biologico? Oppure, elimineresti immediatamente una differenza di 6 cuccioli per cucciolata perché associata a un valore $p$ di 0.09 [30]?

[29] Ricorda che queste stime si basano sui valori dei coefficienti $\hat{\beta}$ calcolati per il nostro modello, ma queste sono solo stime dei parametri reali della popolazione

[30] Pensa a cosa significa 0.09 ...

Infine, puoi confrontare, per ogni posizione, le dimensioni della cucciolata a diversi livelli di disponibilità dei roditori?

### *Un ultimo esercizio*

Infine, per consolidare ciò che è stato spiegato fino ad ora, considera il set di dati `nerveConduction-workshop2.csv`. Questo set contiene misure della velocità di conduzione nervosa in fibre mielinizzate e non mielinizzate, in relazione al loro diametro.

Esplora il set di dati e determina visivamente le relazioni tra le varie variabili. Adatta un modello lineare per esplorare l'effetto di Sesso, Mielinizzazione e Diametro sulla Velocità di Conduzione, esplorando diverse interazioni e definisci quali sono i vari parametri stimati dal tuo modello. Quale modello descrive meglio i dati? Quali conclusioni puoi trarre?