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

- ϵ 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 β 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.

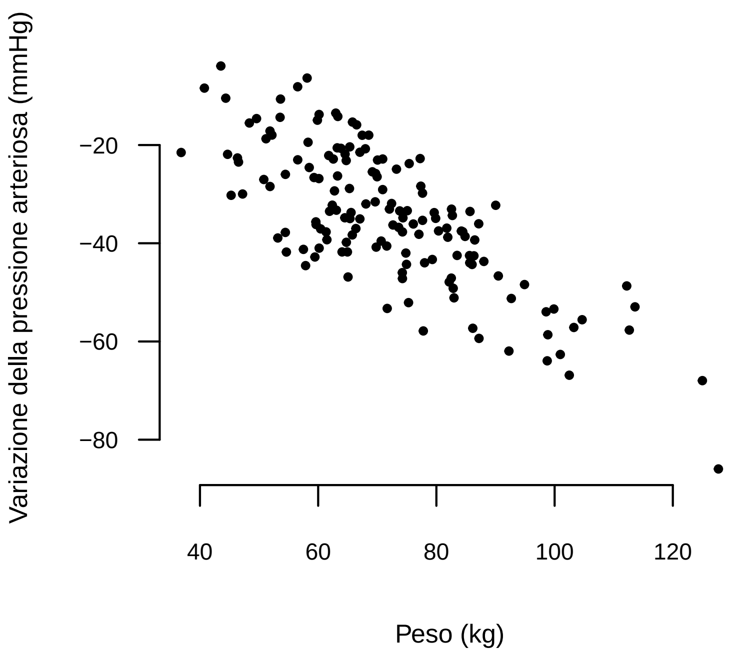
Inizia a familiarizzare con i dati. Quanti uomini e donne ci sono? Quale gamma di età e peso? Traccia le varie variabili l'una contro l'altra e vedi se emergono particolari modelli [4].

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 β 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 β (coefficienti) che sono stati stimati dal modello.

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

Quindi

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

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?

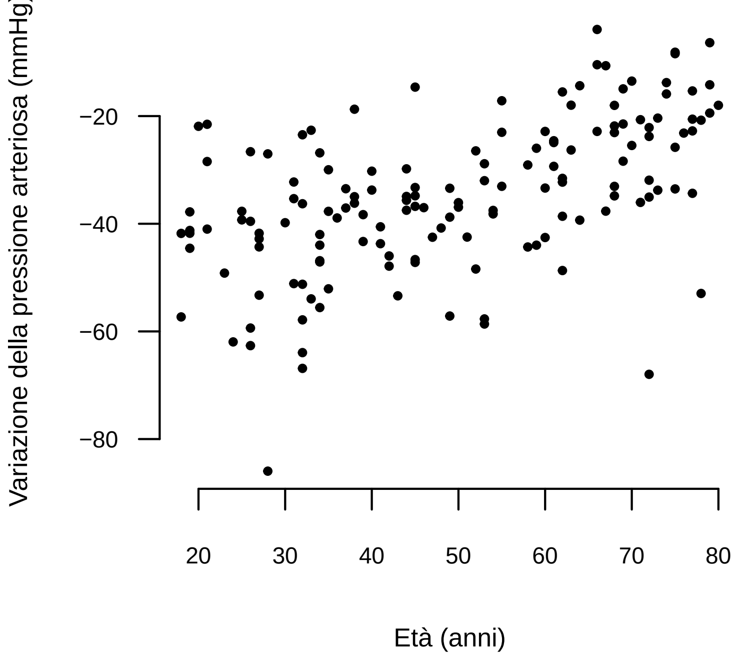
[7] Quale pensi sia il valore massimo di R^2? Perché?

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)
```

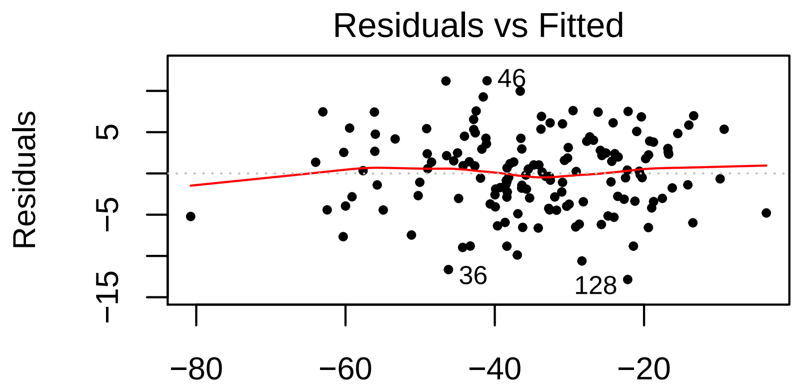
Questo genererà un modello che considera l'effetto del peso e l'effetto dell'età, indipendentemente l'uno dall'altro [9]

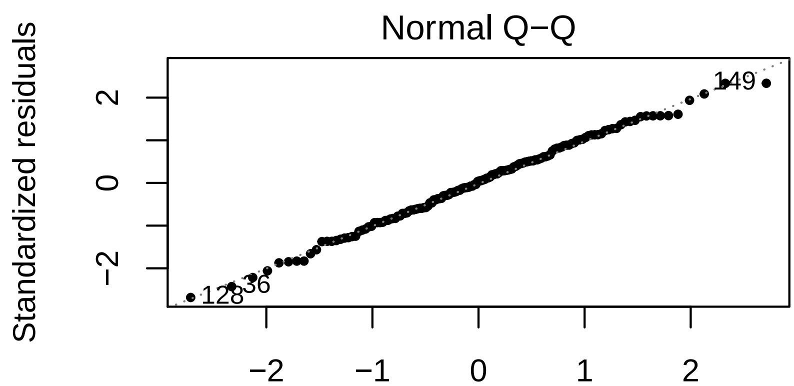
Quali sono le ipotesi nulle [10] che questo modello sta testando?

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.)

Qualitative predictors and dummy variables

Let's now consider gender and add it to our model. Plot the data, do you think gender affects the response to the drug?

In this case, we are dealing with a discrete qualitative variable, with two levels, F and M. All that we said so far still applies, and `lm` is able to deal with this type of variables with no issue. However, the way we deal with these type of variables is slightly different.

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

This is modelling the following:

$\text{Response} = \beta_0 + \beta_1 * \text{Weight} + \beta_2 * \text{Height} + \beta_3 * D$

We introduce a new variable D, called a "dummy variable", that codes Sex in this way:

$$D = \begin{cases} 1, & \text{if Sex} = M \\ 0, & \text{otherwise} \end{cases}$$

By default R assigns 0 to the first level of the variable (the *reference level*, in this case F), and 1 to the second[12]

[12] Levels are ordered alphabetically; see Workshop 1 for how to change level ordering.

Therefore, for observations at the reference level (so from female subjects), the third term $\beta_3 * D$ will be 0; for male subjects that will be $\beta_3 * 1 = \beta_3$. Thus, β_3 represents the **difference between the response of a male and a female**, keeping all of the other factors constant.

Let's have a look at the summary of the model to better clarify this.

```
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
```

The output is not much different from what we had before. Does Sex have a statistically significant effect on the response? What percent of variance is explained by this model?[13]

Consider the estimates

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

These mean that:

- For every increase in 1 kg of weight, the response decreases of 0.66 mmHg (keeping age and sex the same)
- For every increase in 1 year of age, the response increases of 0.42 mmHg (keeping weight and sex the same)
- If the patient is male the response increases of 0.41 mmHg

So, what would you predict the response of a 50 year old male weighing 82 kg will be? Write your response on the forum[14].

[13] Note how, although minimally, R^2 has increased, since we have added an extra parameter, however adj. R^2 has decreased!

[14] Remember to include the intercept in your calculations as well!

Dummy variables for multiple levels

You may have realised at this point, that dummy variables are what R uses to code for groups or other discrete factors when doing an ANOVA!

In some cases, however, you will have more than two levels; the reasoning is the same, however multiple dummy variables will be used to define the different levels.

For example, suppose you measured the levels of LH in three different species of fish: mackerel, salmon, and trout.

You can code the species variable with two dummy variables (so, number of levels - 1) D_1 and D_2 such as:

$$D_1 = \begin{cases} 1, & \text{if Species = "Salmon"} \\ 0, & \text{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1, & \text{if Species = "Tuna"} \\ 0, & \text{otherwise} \end{cases}$$

Therefore:

Species	D_1	D_2
Mackerel	0	0
Salmon	1	0
Tuna	0	1

Our model may be something like:

$\text{LH} = \beta_0 + \beta_1 * D_1 + \beta_2 * D_2 + \epsilon$

where $\beta 1$ represents the difference between LH levels in salmons and mackerels, and $\beta 2$ the difference between LH levels in tuna and mackerel.

Section 2 - Choosing a model

Going back to our initial example, we have fitted three models:

1. Response $= \beta_0 + \beta_1 * \text{Weight} + \epsilon$
2. Response $= \beta_0 + \beta_1 * \text{Weight} + \beta_2 * \text{Age} + \epsilon$
3. Response $= \beta_0 + \beta_1 * \text{Weight} + \beta_2 * \text{Age} + \beta_3 * D_{male} + \epsilon$

We could argue that #2 is better than #1, as it explains a much larger percentage of the variance (88% vs 58%), but what about #3?

Is it correct to say that since Sex does not have a statistically significant effect on the response, and since the value of adjusted R^2 is lower (albeit by a very small amount) we should drop Sex from the model, and only consider Age and Weight as predictors? One way of deciding this is to use the `anova` function to compare the two models. This tests the null hypothesis that the most complex model does not fit the data better than the simpler one[15].

[15] To be more precise, what this actually does is test whether the extra estimated coefficients in the more complex model are not different from 0. Note that you can use this function only if all of the parameters of the smallest model are also present in the more complex model.

```
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
```

As expected, the p-value is very low, indicating that the second, more complex, model fits the data better than the first one, and should be preferred to it. See also how much the residual sum of squares (RSS) has decreased, indicating that the model is much closer to the real data (hence the residuals (and their squared sum) are smaller).

Conversely

```
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
```

The p-value is 0.64, indicating that we cannot refute the null hypothesis, meaning that our third model (with Age, Weight, and Sex) is not better than the simpler model.

R also provides a convenient function, called `drop1`, that removes one predictor at a time from a larger model. You can see that this confirms what we have seen above.[16]

```
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
```

[16] Please, note that these are only very general guidelines. The choice of what to include in your model is not an easy one and different school of thoughts exist on whether it is better to always have a simpler model or a more complete one and often the answer is not straightforward. **The important thing is that any choice you make is based on a solid motivation.** Don't just stop at the p-value... think of what question you are asking and what your model tells you.

Section 3 - Interactions between factors

In the lectures you have learnt about interactions amongst factor in multiple regression. We will now see how to analyse interactions in R.

We will now consider the data in `fox_workshop2.csv`. This dataset[17] contains the litter size, as a measure of reproductive success, in two different population of Artic foxes, in relations to their age and location, as well as rodent availability[18] (which has been classified into low, medium, or high).

[17] Data for this example is fictional, but quite loosely based on work by Tannerfeldt and Angerbjörn (Oikos, 1998)

[18] Rodents are one source of food for Arctic foxes, but their number is fluctuating year-by-year in many regions

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

As usual, we read the data and begin to explore it

```
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
```

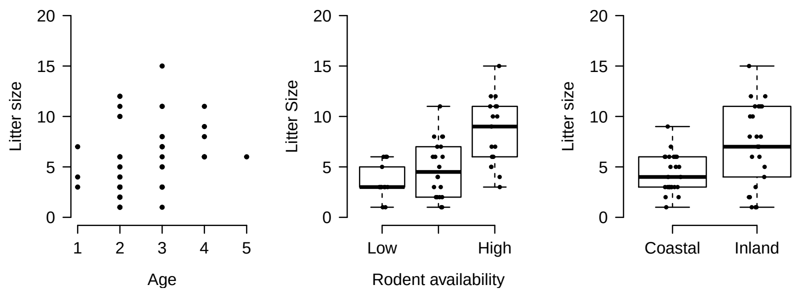
You may have noticed that the levels of the RodentAvail factor are in a slightly unusual order[19]. We can reorder it so Low is used as reference.

[19] Alphabetical!

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

We can then plot some of the relationships between variables[20].

[20] Can you reproduce these plots?



From a first inspection, it seems that reproductive success is somehow correlated to all of the other variables. We can use a linear model to study these relationships[21].

[21] Litter size is a count, therefore it is bounded to 0; we will learn later in the course that a linear model is not the best way to analyse these type of bounded data, but for the moment we can ignore this problem.

Just we did before, we can create a model using `lm`:

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

summary(model)
```

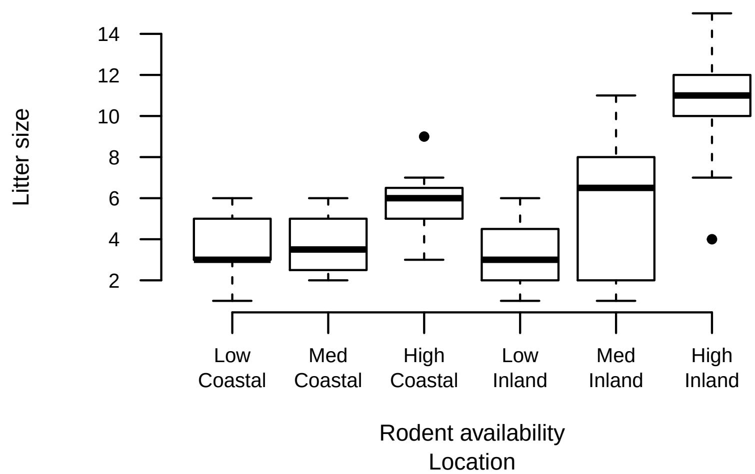
```
##
## Call:
## lm(formula = LitterSize ~ Age + Location + RodentAvail, data = foxes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.898 -1.729  0.233  1.573  4.250
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.6194     1.1996  -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
```

I will leave the point-by-point interpretation of the model's summary to you[22].

[22] Look at the intercept, what does a negative value mean there? Do you see any problems with that?

There is an important question now: is this the best model to describe our data? The adjusted R^2 value is 0.64, meaning that the model only explains about 64% of the data's variance.

Can we improve the model? The answer partly depends on the question we want to address. For example, one interesting question that this model cannot currently answer is "Does rodent availability affect the litter size of both population of foxes in the same way?". This is a more specific and potentially more interesting question to ask, but slightly more complex to answer. We can start by going back to our plots. Let's consider this



Now, that is interesting! It looks like the two populations are not equal when considering the effect of rodent availability on reproductive success! In other words, there is an interaction between location and rodent availability in deremining litter size. This is not only interesting because it gives us the opportunity to look at how to analyse interactions in R... but also because it brings about other questions such as "why there is this difference?"[23].

We can modify our model to take this into account

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

The `RodantAvail:Location` notation is used to indicate the interaction between the two factors. An alternative, and completely equivalent, way of indicating interactions is by using the `*` sign.

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

[23] For example, one explanation could be that coastal areas offer larger provisions of birds, that nest on the cliffs on the coast. These birds can be used by foxes as an alternative food source. Therefore foxes living on the coast have on average smaller litters every year, while foxes living inland have large litters in years when there is a lot of food available. Could you design an experiment to test this hypothesis?

```
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
```

This is slightly more complex from what we have seen before, however it can be interpreted in a very similar manner. We see that the model now explains 80% of the variability in our data, an improvement compared to the previous model[24]! The model also tells us that there is a significant effect of age on litter size. $\hat{\beta}_{Age}$ tells us that for each increase of 1 year in age there is an increase in 2.4 in litter size[25]. It also tells us that there are significant interactions.

We can visualise these interactions by using an *interaction plot*, such as that provided by the `emmip` function in the package `emmeans`[26].

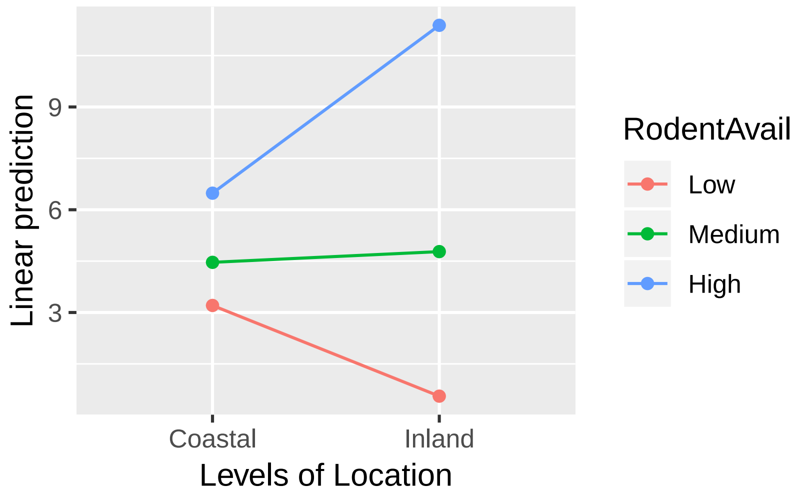
```
par(mar = c(4, 4, 1, 4), cex = 0.5, cex.lab = 1,
    cex.axis = 1)

library(emmeans)  # You need to have emmeans installed for this!
emmip(model.2, RodentAvail ~ Location)
```

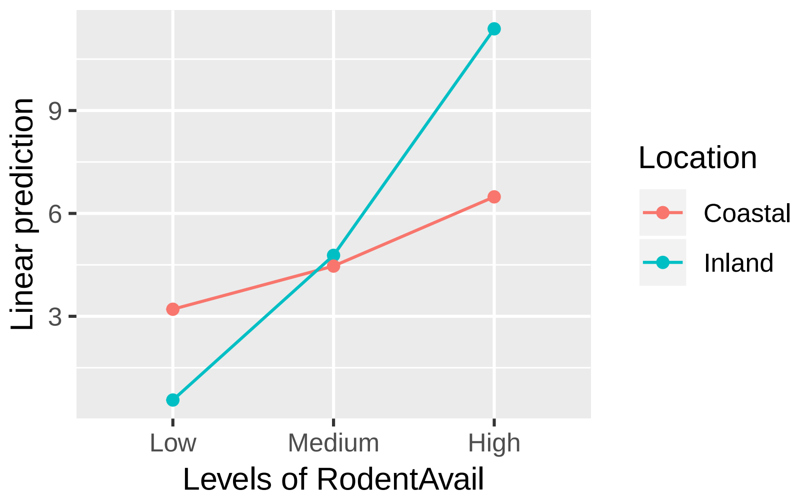
[24] You can also compare this model with the previous one using the `anova` function

[25] We need to be very careful when interpreting these coefficients. The model does not tell us why older animals have bigger litters. It may be directly because of age, or because older animals have previously had litters, and this has an effect on litter size!

[26] If you do not have the package emmeans installed, you can do so using `install.packages("emmeans")`. R also provides another function, called `interaction.plot`, that can produce the same graph)




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



These graphs show the *estimated marginal means*, that is the means estimated by our model for each level of the factors we are considering.

Both graphs show the same information; since the lines are not running parallel to each other we can say that there is an interaction between the two factors. Rodent availability is influencing the inland population more than it is the coastal population.

And what should we make of the estimates for Location and RodentAvail[27]? Because we have a significant interaction, these coefficients become slightly less useful. $\hat{\beta}_2$ is the mean difference in litter size between foxes living inland and those living on the coast, **independently of rodent availability**[28]. However, since rodent availability affects this difference... we would generally ignore these two coefficient when interpreting our model. More formally, in the presence of

[27] That is $\hat{\beta}_2 = 2.46$ and $\hat{\beta}_3 = -2.65$

[28] Can you interpret $\hat{\beta}_3$? Write your explanation in the forum

interactions, we generally ignore the main effects (so the independent effects of each of the two interacting factors across the whole sample).

Finally, let's say that we want to know whether there is a statistical differece between the coastal and inland population, at the different level of RodentAvail. We can use the **emmeans** and **pairs** functions to do so. These function can perform pairwise comparison (just like the Tukey test did), also taking into account interactions. Rather than comparing all of the possible levels, we can specify specific differences (also called contrasts) that we are interested in; this will avoid, for instance, comparisons such as Inland/High vs Coastal/Low, which do not give us any particular biological information.[29].

[29] Still, in case we wanted to look at all possible comparisons... we could use `emmeans(model.2, pairwise ~ Location * RodentAvail)`. This will return all possible comparisons, without having to call 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
```

In the calls to pairs we specify that we want to compare Location at different levels of RodentAvail. The output gives us the estimate of the difference (e.g.: for low rodent availability, the coastal foxes have on average 2.6 pups more than inland foxes), and the standard error associated with this estimate[30]. We also get a p-value for each of the contrasts. Remember, although the p-value tells us that the two conditions are different, it is probably more interesting to look at some measure of effect size (such as the estimate), which tells us about the biological significance of the result. Plot the data, look at the number and think! Would a difference in 0.01 pups/litter with a p-value of 0.02 be of any biological significance? Or, would you immediately dismiss a difference in 6 pups/litter because it was associated with a p-value of 0.09[31]?

[30] Remember, these estimates are based on the values of the $\hat{\beta}$ coefficients calculated for our model, but these are only estimations of the true population parameters

[31] Think of what that 0.09 means...

Finally, can you compare, for each location, the litter size at different level of rodent availability?

A final exercise

Finally, to consolidate what explained so far, consider the dataset `nerveConduction-workshop2.csv`. This contains measures of nerve conduction velocity in myelinated and unmyelinated fibres, in relation to their diameter.

Explore the dataset, and visually determine relationships between the various variables. Fit a linear model to explore the effect of Sex, Myelination and Diameter on conduction Velocity, exploring different interactions, and define what the various parameters estimated by your model mean. Which model best describes the data? What conclusions can you draw?