Anova, Ancova, Interazione, Paradossi

Vers. 1.0.1

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

Outline

Regressione con Interazione e Fattori I

In una qualsiasi regressione multivariata, possiamo scrivere il modello come

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

con $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Concentriamoci sulla "matrice del disegno" $\mathbf{X}$, che è rappresentabile come

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdot & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \cdot & x_{2,p-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n,1} & x_{n,2} & \cdot & x_{n,p-1} \end{pmatrix}$$

e i suoi elementi devono rappresentare le covariate del modello.

Per fare degli esempio (vedete anche il file R collegato), prendiamo un classico dataset dei pinguini

Regressione con Interazione e Fattori II

Sono state osservati individui di tre specie di pinguini, e sono state misurate alcune variabili, tra cui l'isola dove sono stati osservati, la lunghezza delle pinne, BMI, sesso, e due informazioni sul becco

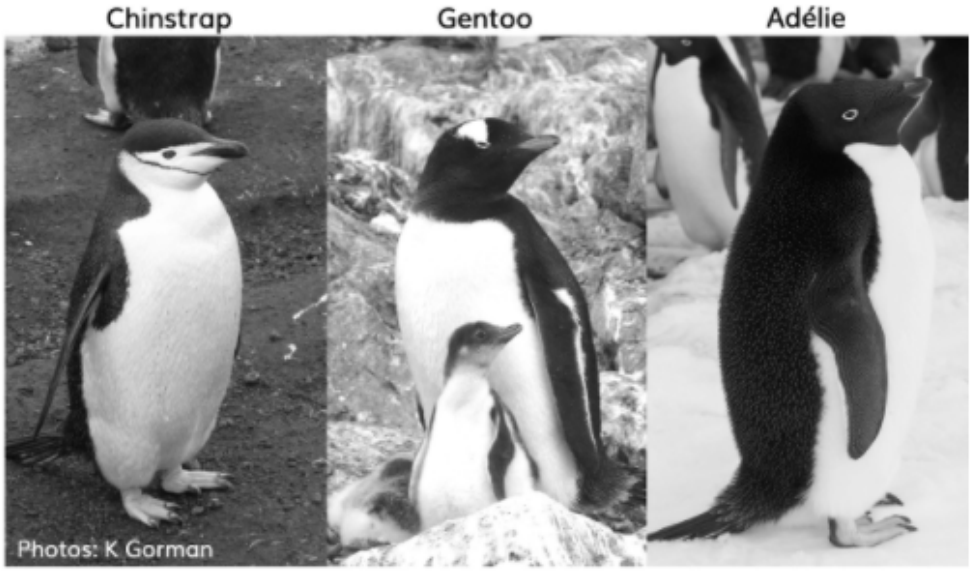


Figure: I pinguini

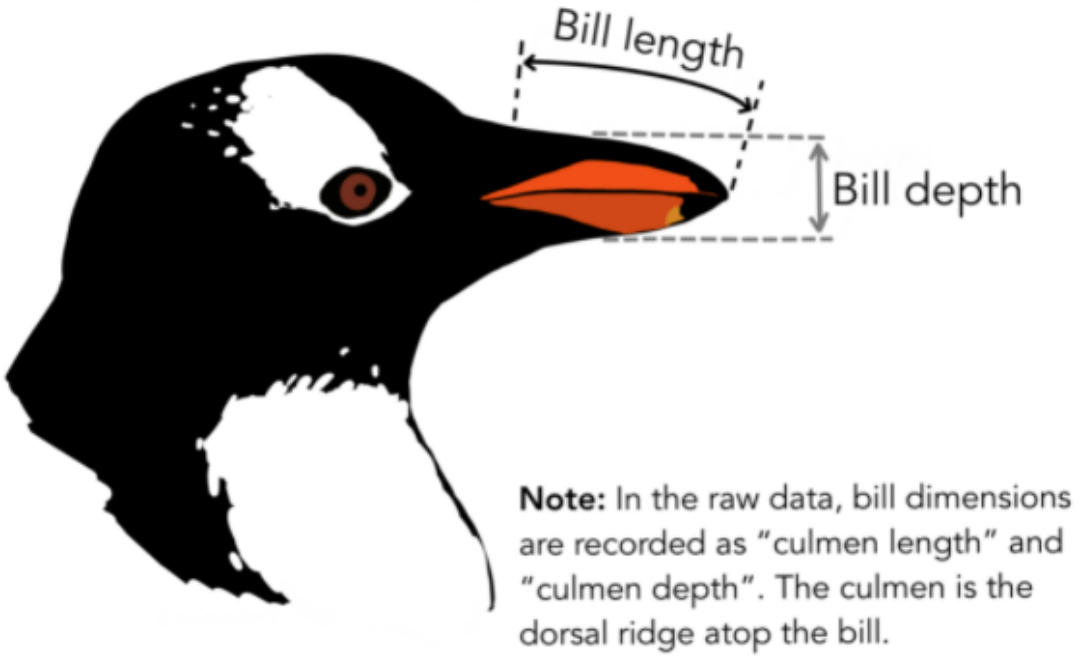


Figure: Il becco

Regressione con Interazione e Fattori IV

Qui sotto trovate una descrizione dei dati.

```
      species          island    bill_length_mm  bill_depth_mm
 Adelie    :152   Biscoe    :168   Min.   :32.10   Min.   :13.10
 Chinstrap: 68   Dream     :124   1st Qu.:39.23   1st Qu.:15.60
 Gentoo    :124   Torgersen: 52   Median :44.45   Median :17.30
                                  Mean   :43.92   Mean   :17.15
                                  3rd Qu.:48.50   3rd Qu.:18.70
                                  Max.   :59.60   Max.   :21.50
                                  NA's   :2       NA's   :2
 flipper_length_mm  body_mass_g       sex           year
 Min.   :172.0     Min.   :2700   female:165   Min.   :2007
 1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
 Median :197.0     Median :4050   NA's  : 11   Median :2008
 Mean   :200.9     Mean   :4202                Mean   :2008
 3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
 Max.   :231.0     Max.   :6300                Max.   :2009
 NA's   :2         NA's   :2
```

Regressione con Interazione e Fattori V

Ipotizziamo di essere interessati alla bill depth, e di vedere come questa cambia in funzione dele altre variabili. Assumiamo che sia distribuita, almeno approssimativamente come una normale, e vediamo che relazione ha con il sesso. Se guardassimo solo la distribuzione dei dati, senza far eun test, non è facile capire se c'è una relazione o no

Regressione con Interazione e Fattori VI

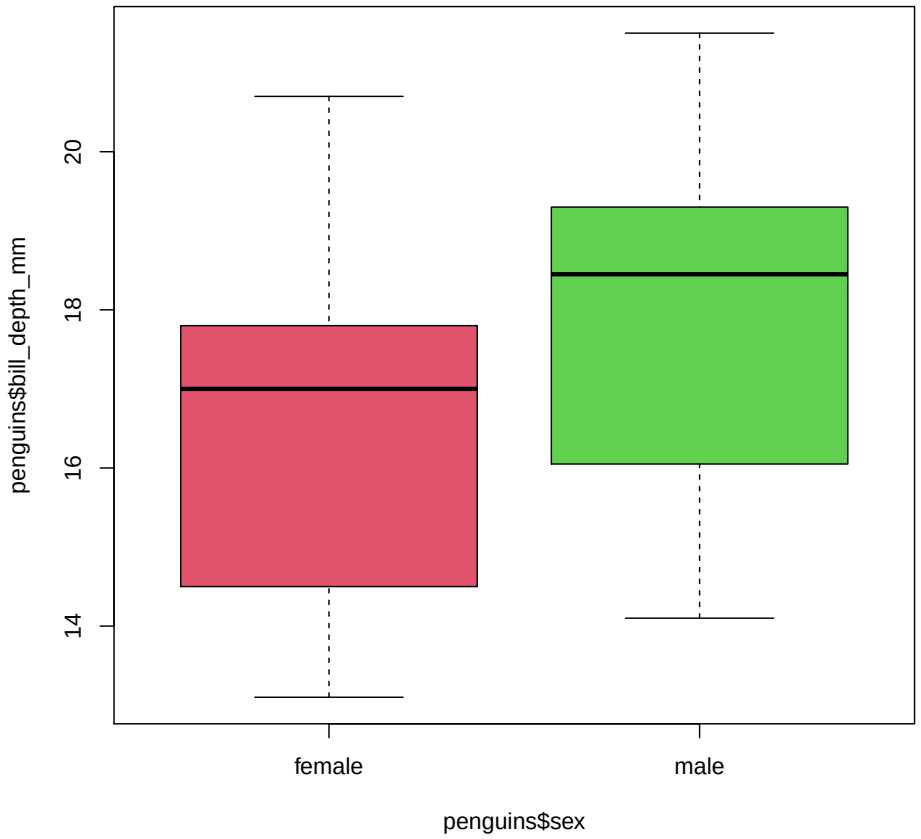


Figure: boxplot

Potremmo assumere quindi che $Y_{i,m} \sim N(\mu_m, \sigma_m^2)$ e $Y_{i,f} \sim N(\mu_f, \sigma_f^2)$, con m e f he indicano maschio e femmina che hanno, rispettivamente n_m e n_f osservazioni, e chiederci se $\mu_m = \mu_f$, dove, per semplicità assumiamo che $\sigma_m^2 = \sigma_f^2 = \sigma^2$.
Possiamo fare un test t, testando l'ipotesi

$$H_0 : \mu_m - \mu_f = 0 \qquad H_1 : \mu_m - \mu_f \neq 0$$

dobbiamo trovare una statistica test che dipende dai dati e dai parametri incogniti, di cui conosciamo al distribuzione, almeno sotto H_0, per poter fare il test. La cosa più naturale, è di prendere $\bar{y}_m$ e $\bar{y}_f$, i.e., le medie campionarie del solo campione di maschie femmine, e prenderne la differenza

$$\bar{y}_m - \bar{y}_f \sim N(\mu_m - \mu_f, \sigma^2/n_m + \sigma^2/n_f)$$

Se H_0 è vera, allora la medie è zero e possiamo costruire un test a livello $\alpha = 0.05$, assumendo σ^2 noto, con la seguente regione di accettazione

$$-z_{0.025} < \frac{\bar{y}_m - \bar{y}_f}{\sqrt{\sigma^2 \left(\frac{n_m + n_f}{n_m n_f}\right)}} < z_{0.025}$$

Se non conosciamo σ^2, dobbiamo usare la solita trasformazione che, definindo

$$S_m^2 = \frac{\sum_{i=1}^{n_m}(y_{i,m} - \bar{y}_m)^2}{n_m - 1} \qquad S_f^2 = \frac{\sum_{i=1}^{n_f}(y_{i,f} - \bar{y}_f)^2}{n_f - 1}$$

e

$$S^2 = \frac{(n_m - 1)S_m^2 + (n_f - 1)S_f^2}{(n - 2)}$$

$(n = n_m + n_f)$ ci porta a dire che

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}$$

e quindi, se non conosciamo la varianza, possiamo usare il test **two-samples t-test**

$$-t_{n-2,0.025} < \frac{\bar{y}_m - \bar{y}_f}{\sqrt{S^2\left(\frac{n_m+n_f}{n_m n_f}\right)}} < t_{n-2,0.025}$$

P.S In teoria potremmo definire anche

$$S^2 = \frac{\sum_{i=1}^{n_m}(y_{i,m} - \bar{y})^2 + \sum_{i=1}^{n_f}(y_{i,f} - \bar{y})^2}{(n-1)}$$

che però è non distorto solo se H_0 è effettivamente vera e non quando è falsa. Questo porta ad avere un test con una probabilità di errore di seconda specie più alto, e quindi si preferisce la forma messa sopra.
Se lo facessimo con R darebbe il seguente risultato

```
Two Sample t-test

data:  penguins$bill_depth_mm[penguins$sex == "female"] and penguins$bill_depth_mm[penguins$s
t = -7.3065, df = 331, p-value = 2.066e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.860208 -1.071025
sample estimates:
mean of x mean of y
 16.42545  17.89107
```

Potremmo anche assumere che le varianze siano diverse, ma in questo caso il test è più complicato, perchè ci sono problemi con il calcolo dei gradi di libertà e serve il **t-test con la correzione di Welch**.

Regressione con Interazione e Fattori X

Quindi le medie sono effettivamente diverse. E se invece volessimo controllare se cambia con la specie? e con le combinazioni specie/sesso? La cosa diventa complicata molto velocemente, ma, per semplificare le cose, possiamo usare la regressione. La domanda è come rappresentiamo il sesso, o un fattore qualsiasi, nella matrice $\mathbf{X}$.

Prendiamo il caso semplice in cui ci sono solo due fattori: il sesso.
Assumiamo che la y_i sia la variabile di interesse e vogliamo rappresentare il modello nella seguente forma

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Visto che $y_i \sim N(\alpha + \beta x_i, \sigma^2)$, se vogliamo replicare il t-test dobbiamo fare in modo che nei due gruppi, la media delle y_i sia diversa.
Definiamo allora x_i come 0 se i è una femmina, e 1 se è una maschio, il che ci da che

$$y_i \sim N(\alpha, \sigma^2)$$

per le femmine, e

$$y_i \sim N(\alpha + \beta, \sigma^2)$$

per i maschi. Bisogna solo fare attenzione allinterpretazizone, perchè

- α è la media delle femmine
- β è quanto dobbiamo aggiungere all'effetto delle femmine, per avere quello dei maschi

Abbiamo allora che se $\beta = 0$ i due sessi non hanno un effetto diverso, altrimenti lo hanno. QUindi, per rispondere alla domanda, possiamo direttamente guardare l'output della regressione

Regressione con Interazione e Fattori XII

```
Call:
lm(formula = penguins$bill_depth_mm ~ penguins$sex)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7911 -1.8911  0.5745  1.3745  4.2745

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       16.4255     0.1425 115.286  < 2e-16 ***
penguins$sexmale   1.4656     0.2006   7.307 2.07e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.83 on 331 degrees of freedom
Multiple R-squared:  0.1389,    Adjusted R-squared:  0.1363
F-statistic: 53.39 on 1 and 331 DF,  p-value: 2.066e-12
```

Come vedete, R ha automaticamente assegnato 0 a “female”.

Prendiamo il caso in cui abbiamo più di due classi, per esempio la specie. In questo caso, non possimao usare la struttura di prima perchè se usassimo

$$y_i = \alpha + \beta x_i + \epsilon_i$$

potremmo dire che $x_i = 0$ se "Adelie", $x_i = 1$ se "Chinstrap", e $x_i = 2$ se "Gentoo ". In questo caso stiamo però assumendo che la differenza tra Chinstrap e Adelie è la stessa tra Gentoo e Chinstrap, il che è un'assunzione forte.
Ipotizziamo che ognuna delle tra specie abbia solo 2 osservazioni, e definiamo la matrice del modello

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

come

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

Abbiamo quindi una regressione con 3 β

$$y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$$

Calcoliamo la distribuzione delle y nei tre gruppi

- Adelie - $y_i \sim N(\beta_1, \sigma^2)$
- Chinstrap - $y_i \sim N(\beta_1 + \beta_2, \sigma^2)$
- Gentoo - $y_i \sim N(\beta_1 + \beta_3, \sigma^2)$

quindi β_1 è l'effetto della specie "non presente" (chiamata corner point), mentre gli altri effetti sono quello che dobbiamo aggiungere al corner per avere l'effetto di quella specie.

Un modello regressivo di questo tipo si chiama **Modello Anova** (Analysis of Variance).

Vediamo l'output di R

Regressione con Interazione e Fattori XV

```
Call:
lm(formula = penguins$bill_depth_mm ~ penguins$species)

Residuals:
     Min       1Q   Median       3Q      Max
-2.84726 -0.79664  0.00336  0.75274  3.15274

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               18.34726    0.09300 197.291   <2e-16 ***
penguins$speciesChinstrap  0.07333    0.16497   0.444    0.657
penguins$speciesGentoo    -3.35062    0.13878 -24.144   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.124 on 330 degrees of freedom
Multiple R-squared:  0.6764,    Adjusted R-squared:  0.6744
F-statistic: 344.8 on 2 and 330 DF,  p-value: < 2.2e-16
```

Che ci fa vedere come specie due non sia differente da specie 1, mentre specie 3 è differente da specie 1. Ma cosa possimao dire della specie 2 VS specie 3? Abbiamo due soluzioni,

- Soluzione "matematica" - Siamo interessati a trovare il valore $\beta_3 - \beta_2$, visto che se è zero allora le due specie sono uguali. Sappiamo che

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

  e allora se definiamo un vettore

$$\mathbf{C} = \begin{pmatrix} 0 & -1 & 1 \end{pmatrix}$$

  abbiamo che

$$\mathbf{C}\hat{\boldsymbol{\beta}} = \hat{\beta}_3 - \hat{\beta}_2 \sim N(\beta_3 - \beta_2, \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')$$

  Da cui posso fare qualsiasi tipo di test. Questa procedura viene chiamata anche con il termine **contrasti**.

Regressione con Interazione e Fattori XVII

- Soluzione “pratica” - Quel che fa R è prendere un valore a caso come corner, e basta cambiarlo e rifittare la regressione

```
### code
penguins$species = relevel(penguins$species,"Chinstrap")
reg = lm(penguins$bill_depth_mm~ penguins$species)
summary(reg)

### output

Call:
lm(formula = penguins$bill_depth_mm ~ penguins$species)

Residuals:
     Min       1Q   Median       3Q      Max
-2.84726 -0.79664  0.00336  0.75274  3.15274

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             18.42059    0.13627 135.181   <2e-16 ***
penguins$speciesAdelie  -0.07333    0.16497  -0.444    0.657
penguins$speciesGentoo  -3.42395    0.17082 -20.044   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Residual standard error: 1.124 on 330 degrees of freedom
Multiple R-squared:  0.6764,    Adjusted R-squared:  0.6744
F-statistic: 344.8 on 2 and 330 DF,  p-value: < 2.2e-16
```

Ci si potrebbe chiedere perchè non usare questa matrice

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

dove abbiamo un intercetta, e una colonna per ogni specie. In qeusto caso avremo che

- Adelie - $y_i \sim N(\beta_1 + \beta_2, \sigma^2)$
- Chinstrap - $y_i \sim N(\beta_1 + \beta_3, \sigma^2)$
- Gentoo - $y_i \sim N(\beta_1 + \beta_4, \sigma^2)$

Il modello sembrerebbe equivalente, ma abbiamo due problemi. Il primo **interpretativo**: se la media di Adelie fosse 0, esistono infiniti valori di β_1 e β_2 che danno 0, basta assumere $\beta_1 = -\beta_2$, e il modello si dice quindi **non identificabile**. Il secondo è che la matrice X non ha rango pieno, i.e., le righe sono linearmente dipendenti.
Un'altra possibile scelta per la matrice del disegno è

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

che ha rango 3, e abbiamo

- Adelie - $y_i \sim N(\beta_1, \sigma^2)$
- Chinstrap - $y_i \sim N(\beta_2, \sigma^2)$

- Gentoo - $y_i \sim N(\beta_3, \sigma^2)$

Questa può essere utilizzata, ma i software preferiscono l'altra formalizzazione (qui i test del tipo $H_0 : \beta_1 = 0$ non hanno nessuna utilità).

Introduziono il concetto di **interazione** con degli esempi. Con lo stesso dataset, ipotizziamo di voler spiegare la stessa variabile di prima (la profondità del becco) con il BMI e la lunghezza delle pinne

$$y_i = \beta_1 + \beta_2 x_{i,BMI} + \beta_3 x_{i,length} + \epsilon_i$$

Facciamo girare il modello e otteniamo

```
Call:
lm(formula = bill_depth_mm ~ flipper_length_mm + I(log(body_mass_g)),
data = penguins)

Residuals:
Min       1Q   Median       3Q      Max
-3.0932 -1.1698 -0.0643   1.1651   4.1865

Coefficients:
                     Estimate  Std. Error t value Pr(>|t|)
(Intercept)          22.53170    5.68482   3.963 9.06e-05 ***
flipper_length_mm    -0.10238    0.01243  -8.234 4.27e-15 ***
I(log(body_mass_g))   1.82652    0.92470   1.975   0.0491 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.603 on 330 degrees of freedom
Multiple R-squared:  0.3416,    Adjusted R-squared:  0.3376
F-statistic: 85.62 on 2 and 330 DF,  p-value: < 2.2e-16
```

Ci possiamo però chiedere se la relazione tra la lunghezza delle pinne e y cambi, in base al valore di BMI, e questo si può fare introducendo un effetto di interazione che nel modello si scrive come

$$y_i = \beta_1 + \beta_2 x_{i,BMI} + \beta_3 x_{i,length} + \beta_4 x_{i,BMI} x_{i,length} + \epsilon_i$$

vediamo adesso me si interpreta β_4 usando il trucco visto precedentemente, di aumentare di uno una specifica variabile, lasciando le altre fisse. Per esempio, supponiamo che il soggetto i' abbia le stesse variabili di i, ma con una lunghezza delle pinne maggiore di 1, abbiamo che

$$y_{i'} = \beta_1 + \beta_2 x_{i,BMI} + \beta_3 (x_{i,length} + 1) + \beta_4 x_{i,BMI} (x_{i,length} + 1) + \epsilon_{i'} =$$

$$y_i + \beta_2 + \beta_4 x_{i,BMI}$$

quindi possiamo vedere β_2 come l'effetto "marginale", mentre l'aumento/diminuzione dovuta a β_4 dipende dal valore di $x_{i,BMI}$.
Quindi, se $\beta_4 = 0$, la relazione tra la lunghezza delle pinne e y non cambia al variare si BMI. Possiamo utilizzare R che ci da il seguente risultato

```
reg = lm(bill_depth_mm~ flipper_length_mm+ I(log(body_mass_g))+
flipper_length_mm: I(log(body_mass_g)) , data = penguins)
summary(reg)
```

```
Call:
lm(formula = bill_depth_mm ~ flipper_length_mm + I(log(body_mass_g)) +
    flipper_length_mm:I(log(body_mass_g)), data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3450 -1.0469 -0.0476  1.0983  4.0123

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                             -170.71076   65.34508  -2.612  0.00940 **
flipper_length_mm                          0.86693    0.32679   2.653  0.00837 **
I(log(body_mass_g))                       24.82138    7.80065   3.182  0.00160 **
flipper_length_mm:I(log(body_mass_g))     -0.11520    0.03881  -2.968  0.00321 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.584 on 329 degrees of freedom
Multiple R-squared:  0.3588,    Adjusted R-squared:  0.353
F-statistic: 61.37 on 3 and 329 DF,  p-value: < 2.2e-16
```

è interessante vedere come l'effetto "marginale" della lunghezza delle pinne era negativo senza interazione, mentre diventa positivo con l'interazione. Per concludere questa parte, facciamo un modello in cui, oltre all'interazione, mettiamo anche il sesso (variabile fattore).

Un modello di questo tipo si chiama **ANCOVA**

Regressione con Interazione e Fattori XXV

```
Call:
lm(formula = bill_depth_mm ~ flipper_length_mm + I(log(body_mass_g)) +
    flipper_length_mm:I(log(body_mass_g)) + sex, data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6698 -0.8669 -0.0936  0.7795  3.8785

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -105.93068   47.68560  -2.221 0.027003
flipper_length_mm                    0.76498    0.23780   3.217 0.001425
I(log(body_mass_g))                 15.92919    5.69829   2.795 0.005488
sexmale                              2.48677    0.14510  17.138  < 2e-16
flipper_length_mm:I(log(body_mass_g))  -0.09819    0.02825  -3.476 0.000578

(Intercept)                       *
flipper_length_mm                 **
I(log(body_mass_g))               **
sexmale                           ***
flipper_length_mm:I(log(body_mass_g)) ***
---
```

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

Residual standard error: 1.152 on 328 degrees of freedom
Multiple R-squared:  0.6617,    Adjusted R-squared:  0.6576
F-statistic: 160.4 on 4 and 328 DF,  p-value: < 2.2e-16
```

Potremmo anche introdurre interazione tra una variabile continua, e un fattore. In questo modo avremmo un coefficiente regressivo per ognuno dei gruppi, oppure interazione tra fattori. Però, prima di vederlo, mettiamoci in una situazione più facile e introduziamo il **paradosso di Simpson**.

Paradosso di Simpson I

Prendiamo degli esempio reali.

Esempio 1 - il numero di persone condannate a morte in Florida, tra il 76 e il 77, in base al colore della pelle

Imputato	yes	no
Bianco	19 (13%)	141
Nero	17 (11%)	149

Dalla tabella sembrerebbe che o il colore della pelle è ininfluente, oppure i Bianchi venivano condannati a morte più dei Neri. Non stiamo tenendo da conto una variabile **"confondente"**, che in questo caso è il colore della pelle della vittima

Vittima	Imputato	yes	no
Bianco	Bianco	19 (14%)	132
Bianco	Nero	11 (21%)	52
Nero	Bianco	0 (0%)	9
Nero	Nero	6 (6%)	97

Paradosso di Simpson II

Notate come, anche se a parità di vittima, chi ha la pelle nera viene condannato a morte più spesso, quando non si tiene conto di questa informazione, deduciamo il contrario. Da qui il paradosso.

Per capire meglio il paradosso, riprendiamo il dataset dei pinguini, e vediamo come le analisi possono essere influenzate da variabili confondenti. Proviamo a spiegare il depth con il length del becco con la lunghezza. Usiamo R e otteniamo i seguenti risultati.

Paradosso di Simpson III

```
Call:
lm(formula = bill_depth_mm ~ bill_length_mm, data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1548 -1.4291  0.0122  1.3994  4.5004

Coefficients:
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)    20.78665    0.85417  24.335   < 2e-16 ***
bill_length_mm -0.08233    0.01927  -4.273 0.0000253 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.92 on 331 degrees of freedom
Multiple R-squared:  0.05227,   Adjusted R-squared:  0.04941
F-statistic: 18.26 on 1 and 331 DF,  p-value: 0.00002528
```

Paradosso di Simpson IV

che ci dice che c'è una relazione, significativa, di tipo negativo.
Ipotizziamo adesso di voler vedere se questa relazione cambia con la specie. Quindi ogni specie deve avere il suo regressore. Ipotizzando, per semplicità. che ogni specie abbia solo due osservazioni, possiamo scrivere la matrice come

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,length} & 0 & 0 \\ 1 & x_{2,length} & 0 & 0 \\ 1 & x_{3,length} & x_{3,length} & 0 \\ 1 & x_{4,length} & x_{4,length} & 0 \\ 1 & x_{5,length} & 0 & x_{5,length} \\ 1 & x_{6,length} & 0 & x_{6,length} \end{pmatrix}$$

che ci da quindi

- Adelie - $y_i \sim N(\beta_1 + \beta_2 x_{i,length}, \sigma^2)$
- Chinstrap - $y_i \sim N(\beta_1 + (\beta_2 + \beta_3) x_{i,length}, \sigma^2)$
- Gentoo - $y_i \sim N(\beta_1 + (\beta_2 + \beta_4) x_{i,length}, \sigma^2)$

Paradosso di Simpson V

dove β_1 è l'intercetta, β_2 è l'effetto regressivo del primo gruppo, mentre β_3 e β_4 sono i valori che dobbiamo aggiungere a β_2 per ottenere l'effetto regressivo del gruppo 2 e 3. Anche qui, testare $\beta_3 = 0$ e $\beta_4 = 0$ si può usare per verificare se la relazione cambia nei gruppi. Usiamo R per ottenere i risultati, e abbiamo

```
## code
reg = lm(bill_depth_mm~ bill_length_mm+bill_length_mm:species, data = penguins)
summary(reg)
```

```
Call:
lm(formula = bill_depth_mm ~ bill_length_mm + bill_length_mm:species,
    data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4823 -0.7092 -0.0521  0.5394  3.8331

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     8.530080   0.795637  10.721   <2e-16 ***
bill_length_mm                  0.252510   0.020503  12.316   <2e-16 ***
bill_length_mm:speciesChinstrap -0.049885   0.005248  -9.506   <2e-16 ***
```

Paradosso di Simpson VI

```
bill_length_mm:speciesGentoo    -0.116265   0.004666 -24.919   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.9717 on 329 degrees of freedom
Multiple R-squared:  0.7587,    Adjusted R-squared:  0.7565
F-statistic: 344.8 on 3 and 329 DF,  p-value: < 2.2e-16
```

Notate come prima l'effetto era negativo, mentre adesso sono tutti positivi, visto che
$\beta_2 + \beta_3 = 0.252510 - 0.049885 = 0.202625$ e
$\beta_2 + \beta_4 = 0.252510 - 0.116265 = 0.136245$,
quindi, anche qui abbiamo il paradosso di Simpson, visto che non considerare variabili nel modello, ci porta a risposte sbagliate, e diametralmente opposte a quelle vere.
Possiamo anche cambiare il modello e non assumere che tutti abbiano la stessa intercetta, anche perchè ha poco senso, e usare la matrice

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,length} & 0 & 0 & 0 & 0 \\ 1 & x_{2,length} & 0 & 0 & 0 & 0 \\ 1 & x_{3,length} & x_{3,length} & 0 & 1 & 0 \\ 1 & x_{4,length} & x_{4,length} & 0 & 1 & 0 \\ 1 & x_{5,length} & 0 & x_{5,length} & 0 & 1 \\ 1 & x_{6,length} & 0 & x_{6,length} & 0 & 1 \end{pmatrix}$$

Paradosso di Simpson VII

che ci da quindi

- Adelie - $y_i \sim N(\beta_1 + \beta_2 x_{i,length}, \sigma^2)$
- Chinstrap - $y_i \sim N(\beta_1 + \beta_5 + (\beta_2 + \beta_3) x_{i,length}, \sigma^2)$
- Gentoo - $y_i \sim N(\beta_1 + \beta_6 + (\beta_2 + \beta_4) x_{i,length}, \sigma^2)$

quindi ogni gruppo ha la propria intercetta e proprio coefficiente angolare. Otteniamo

```
## code
reg = lm(bill_depth_mm~ bill_length_mm:species+species+bill_length_mm, data = penguins)
summary(reg)
```

```
Call:
lm(formula = bill_depth_mm ~ bill_length_mm:species + species +
    bill_length_mm, data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6574 -0.6559 -0.0483  0.5203  3.4990

Coefficients:
```

Paradosso di Simpson VIII

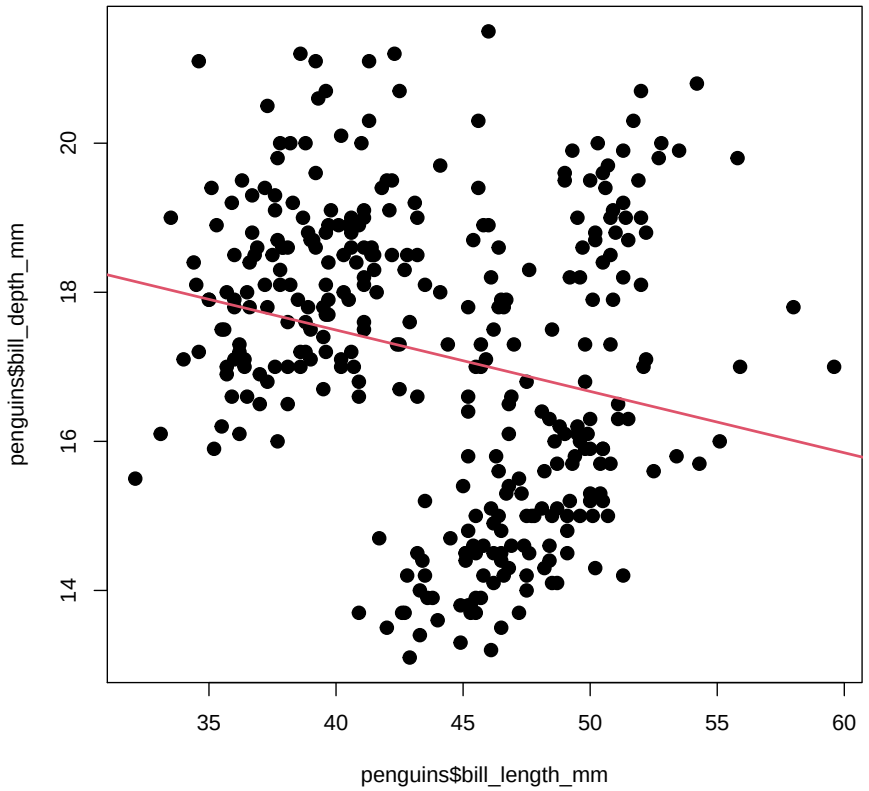
```
                                Estimate Std. Error t value      Pr(>|t|)
(Intercept)                     11.48771    1.15987   9.904       < 2e-16 ***
speciesChinstrap                -3.91857    2.06731  -1.895        0.0589 .
speciesGentoo                   -6.36675    1.77990  -3.577        0.0004 ***
bill_length_mm                   0.17668    0.02981   5.928 0.00000000779 ***
bill_length_mm:speciesChinstrap  0.04553    0.04594   0.991        0.3224
bill_length_mm:speciesGentoo     0.03093    0.04112   0.752        0.4525
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9556 on 327 degrees of freedom
Multiple R-squared:  0.7681,    Adjusted R-squared:  0.7645
F-statistic: 216.6 on 5 and 327 DF,  p-value: < 2.2e-16
```

Abbiamo anche che il segno di β_3 e β_4 è cambiato, ma non sono significativi, quindi ci sta dando una differenza tra gli effetti regressivi diversa da quella precedente. Per capire bene cosa stia succedendo, vediamo dei plot.

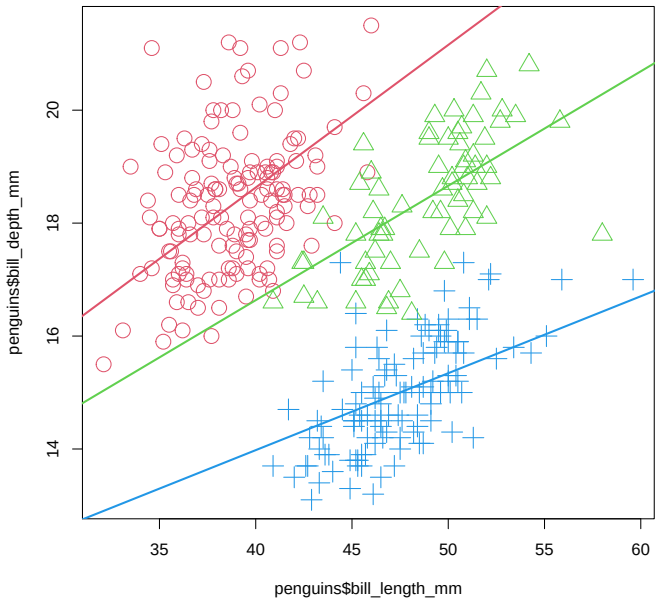
Paradosso di Simpson IX

Il modello con un solo effetto regressivo, dove la linea rosse è la regressione



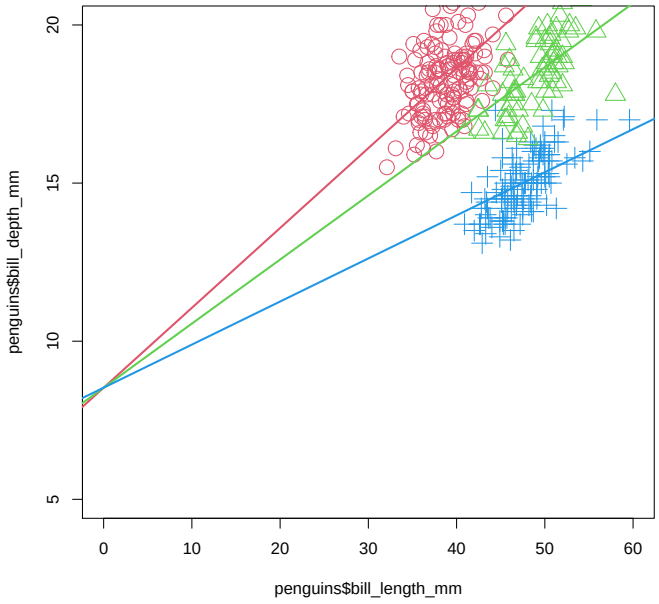
Paradosso di Simpson X

Il modello con un effetto regressivo per specie (i diversi simbolo). Ogni retta è la regressione del gruppo



le rette non fittano bene i dati proprio perchè hanno una intercetta in comune, come si vede qui

Paradosso di Simpson XI



Per concludere vediamo il modello con intercette diverse

Paradosso di Simpson XII