# 8. ANALISI DELLA COVARIANZA (ANCOVA)

**L'analisi della covarianza è un metodo statistico che risulta dalla combinazione dell'analisi di regressione con l'analisi della varianza**. È utile quando **all'analisi dei dati di una certa variabile**, posta sotto controllo in base ad un determinato disegno sperimentale, **risulta associata una covariata** per cui sia difficile creare dei gruppi omogenei da sottoporre alla sperimentazione. Per esempio, se si vuole studiare l'effetto di una determinata dieta alimentare, il peso iniziale degli animali sottoposti alla sperimentazione può influire sul dato finale portando a conclusioni errate. In teoria potremmo ovviare a questo inconveniente formando gruppi sperimentali di animali aventi tutti lo stesso peso, ma questo non è spesso possibile e non sarebbe neanche molto utile. Un esperimento così programmato, non rispecchiando una situazione realistica dove esiste una normale eterogeneità nel peso degli animali, ci consentirebbe di stabilire l'effetto della dieta alimentare solo su animali che hanno un determinato peso di partenza. In altre parole, il risultato non sarebbe generalizzabile e quindi di scarso interesse. Questi problemi possono essere risolti qualora il fattore di eterogeneità sia misurabile, direttamente o indirettamente, attraverso una variabile concomitante (variabile indipendente) utilizzando l'analisi della regressione. Questo approccio statistico ci consentirebbe di ridurre la varianza dell'errore, dovuta all'eterogeneità di partenza, e correggere le medie dei gruppi per ottenere una corretta stima degli effetti dei trattamenti. Ovviamente non è detto che questo metodo sia sempre applicabile, occorre infatti come **presupposto che la regressione della variabile dipendente** (Y, variabile **regressa**) sulla variabile indipendente (X, regressore) **sia la stessa in tutti i gruppi sperimentali**. Se questa assunzione è soddisfatta allora è possibile trovare un **coefficiente di regressione (b) comune** che ci consenta appunto di **aggiustare le medie della variabile oggetto di studio ad un valore iniziale identico per tutti i gruppi** eliminando così l'eterogeneità intrinseca nei dati.

Contrariamente a quanto si potrebbe pensare, nell'ancova non è necessario che siano soddisfatte le assunzioni relative alla regressione ( distribuzione normale degli errori, omogeneità della varianza)

Esempio (da Lison): si vuole studiare l'effetto di un determinato farmaco sul peso del fegato in due gruppi di sei topolini (lotto 1= gruppo controllo; lotto 2 = gruppo trattato). Per valutare correttamente l'effetto si misura il peso corporeo iniziale di ciascun animale in quanto vogliamo evitare che l'effetto del farmaco sul fegato venga confuso con quello eventualmente dovuto al peso corporeo.

(Il peso corporeo è qui utilizzato come misura indiretta dell'eterogeneità del peso del fegato negli animali dei due gruppi)

| pes_anim | pes_org | lotto |
|---|---|---|
| 111 | 11,2 | 1 |
| 126 | 12,6 | 1 |
| 103 | 10,5 | 1 |
| 118 | 12,6 | 1 |
| 148 | 14,4 | 1 |
| 134 | 14 | 1 |
| 121 | 13,3 | 2 |
| 110 | 12,1 | 2 |
| 125 | 14 | 2 |
| 108 | 11,4 | 2 |
| 126 | 13,5 | 2 |
| 104 | 11,6 | 2 |

RIEPILOGO **peso animale X**

| *Gruppi* | *Media* | *Varianza* | *Dev(SS)* |
|---|---|---|---|
| Lotto 1 | **123,33** | 264,667 | 1323,333 |
| Lotto 2 | **115,67** | 89,867 | 449,3333 |
| totale | **119,5*** | 161,152 | 1772,667 |

(Si noti come i due gruppi hanno medie differenti)

* gran media

RIEPILOGO **di peso organo Y**

| *Gruppi* | *Media* | *Varianza* | *Dev (SS)* |
|---|---|---|---|
| Lotto 1 | 12,55 | 2,311 | **11,555** |
| Lotto 2 | 12,65 | 1,187 | **5,935** |
| totale | 12,60 | 1,593 | 17,520 |

Verifichiamo ora se il peso dell'organo è influenzato dal peso corporeo.

**Regressione: peso organo su peso animale nel lotto 1**

| | *Coeff.* | *Errore std* | *Stat t* | *Sign.* | *L. inf. 95%* | *L. sup. 95%* |
|---|---|---|---|---|---|---|
| Intercetta | 1,43136 | 1,527139 | 0,937282 | 0,401678 | -2,80866 | 5,671378 |
| Variabile X 1 | 0,090151 | 0,012293 | 7,333293 | 0,001841 | 0,056019 | 0,124283 |

ANALISI VARIANZA di regressione R al quadrato = 0,930769

| | *gdl* | *SQ* | *MQ* | *F* | *Significatività F* |
|---|---|---|---|---|---|
| **Regressione** | 1 | **10,75503** | 10,75503 | 53,77718 | 0,001841 |
| **Residuo** | 4 | **0,79997** | 0,199992 | | |
| Totale | 5 | **11,555** | | | |

**Regressione: peso organo su peso animale nel lotto 2**

| | *Coefficienti* | *Errore standard* | *Stat t* | *Valore di significatività* | *Inferiore 95%* | *Superiore 95%* |
|---|---|---|---|---|---|---|
| Intercetta | -0,16944 | 1,763827 | -0,09606 | 0,928092 | -5,0666 | 4,727731 |
| Variabile X2 | 0,110831 | 0,015207 | 7,288282 | 0,001884 | 0,06861 | 0,153051 |

ANALISI VARIANZA di regressione R al quadrato = 0,929971

| | *gdl* | *SQ* | *MQ* | *F* | *Significatività F* |
|---|---|---|---|---|---|
| Regressione | 1 | 5,519377 | 5,519377 | 53,11905 | 0,001884 |
| Residuo | 4 | 0,415623 | 0,103906 | | |
| Totale | 5 | 5,935 | | | |

L'analisi della regressione, condotta separatamente per i due gruppi, ci indica che c'è una regressione significativa del peso del fegato sul peso dell'animale in entrambi i gruppi. Se avessimo eseguito una semplice ANOVA sul peso dell'organo per valutare l'effetto del trattamento avremmo ottenuto i seguenti risultati:

| pes_orgL1 | pes_orgL2 |
|---|---|
| 11,2 | 13,3 |
| 12,6 | 12,1 |
| 10,5 | 14 |
| 12,6 | 11,4 |
| 14,4 | 13,5 |
| 14 | 11,6 |
| 12,55 | 12,65 |

**Analisi varianza: ad un fattore (peso organo)** $r^2$= 0,002

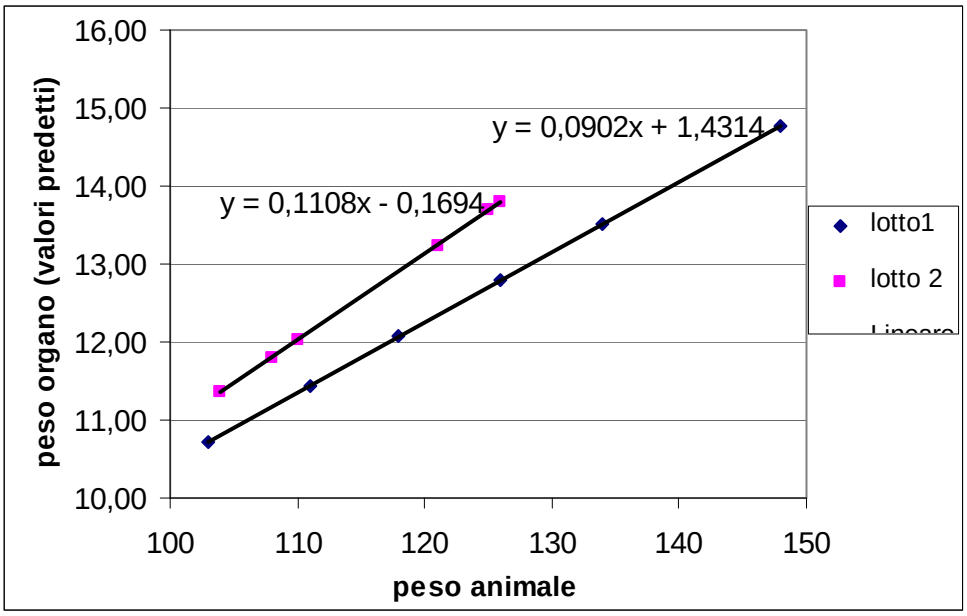
| *Variazione* | *SQ o SS* | *gdl* | *MQ o MS* | *F* | *Sign.* | *F crit* |
|---|---|---|---|---|---|---|
| Tra gruppi | 0,03 | 1 | 0,03 | 0,017 | 0,898 | 4,9646027 |
| **In gruppi** | **17,49** | 10 | **1,749** | | | |
| Totale | 17,52 | 11 | | | | |

Se non si considera il peso del corpo (variabile concomitante[1]), la differenza tra le medie relative al peso dell'organo nei due gruppi non risulta significativa. In effetti la varianza attribuibile alle cause accidentali (in gruppi) include anche la quota attribuibile al fattore eterogeneità (peso animale). La varianza d'errore risulta così sovrastimata e di conseguenza la possibilità di mettere in evidenza effetti significativi dovuti alla dieta (potenza del test) è ridotta. In questo esempio il peso degli animali controllo (lotto 1) è maggiore di quello dei trattati e questo si ripercuote sul peso del fegato, mascherando l'effetto dovuto al trattamento presente nel gruppo dei trattati.

[1] I valori della variabile peso del corpo (X) fanno riferimento alle stesse unità sperimentali su cui viene misurata la variabile peso dell'organo (Y) e per questo viene detta concomitante.

L'analisi della covarianza si basa proprio sulla possibilità di scorporare gli effetti di una variabile concomitante (covariata) e quindi di ridurre la varianza d'errore e di correggere le medie dei gruppi per avere una stima non viziata degli effetti dovuti ai trattamenti.
Una volta evidenziato che in tutti i gruppi esiste una **influenza significativa** (regressione) della variabile concomitante sulla variabile casuale, oggetto della analisi, occorre verificare che **la relazione lineare tra Y ed X possa essere adeguatamente rappresentata da un coefficiente, di regressione comune,** ovvero che l'influenza della variabile concomitante sia uguale in tutti i gruppi.
In altre parole i coefficienti di regressione dei vari gruppi devono essere più o meno uguali e quindi produrre **rette tra loro parallele**. **Se questo presupposto non è soddisfatto non è possibile fare alcuna analisi congiunta** ed i gruppi possono essere analizzati solo separatamente.

La rappresentazione grafica dei dati ci suggerisce l'esistenza di questo parallelismo.




| lotto 1 | lotto 2 |
|---|---|
| Y predicted | |
| 11,44 | 13,24 |
| 12,79 | 12,02 |
| 10,72 | 13,68 |
| 12,07 | 11,80 |
| 14,77 | 13,80 |
| 13,51 | 11,36 |
| 12,55 | 12,65 |

(medie)

(la graficazione è stata ottenuta utilizzando i coefficienti risultanti da ciascuna regressione

$\hat{y}$ **= $a_i$ + $b_i$*x**

E' comunque necessario avvalorare questa assunzione con un test statistico: **test di parallelismo**.

Innanzi tutto dobbiamo **trovare il $b_c$ comune** che può essere calcolato sia in base alla formula generale:
**$b_c = \Sigma SS(X_iY_i) / \Sigma SS(X_i)$**

sia come media ponderata (dato che le devianze non sono identiche nei due gruppi) dei due coefficienti di regressione:
**$b_c = \Sigma b_i * SS(X_i) / \Sigma SS(X_i)$**

Se utilizziamo il primo metodo di calcolo abbiamo:

**Peso animale**

| *Gruppi* | ***Dev(SS)*** |
|---|---|
| Lotto 1 | 1323,3333 |
| Lotto 2 | 449,33333 |
| **Totale** | 1772,6667 |

**p. anim - p. organo**

| *Gruppi* | *Covarianza* | ***Codevianza*** |
|---|---|---|
| Lotto 1 | 23,860 | 119,300 |
| Lotto 2 | 9,96 | 49,8 |
| **Totale** | 33,82 | 169,100 |

**$b_c$** = (119,3 + 49,8) / (1323,33 + 449,33) = **0,0954**

Se utilizziamo il secondo metodo di calcolo abbiamo:
**$b_c$** = (0,0901*1323,33 + 0,1108*449,33) / 1772,66 = **0,0954**

Occorre ora impostare i **parametri delle regressione comune.** Ricordando che la **Devianza di regressione = quadrato della codevianza / devianza della variabile indipendente,** abbiamo:
**Devianza della regressione comune, $SS_c = [\Sigma SS(X_iY_i)]^2/\Sigma SS(X_i)$** = $169,1^2$ / 1772,66 = **16,131**

**Devianza della regressione cumulativa, $SS_{cu} = \Sigma SS(b_i) =$** 10,755 + 5,519 **= 16,274**
(somma delle devianze delle due regressioni)

**Residuo** (somma dei residui delle due regressioni) = **$\Sigma SS(e_i) = 1,2156$**

**STATISTICA della REGRESSIONE COMUNE**

| | *gdl* | ***SS*** | *MS* | *F* | *Significatività F* | |
|---|---|---|---|---|---|---|
| Regr. com | 1 | 16,131 | 16,131 | | | |
| Reg.cum | 2 | 16,274 | 8,137 | 53,552 | **2,33E-05** | **(1)** |
| Differenza | 1 | 0,143 | 0,143 | 0,944 | **0,359** | **(2)** |
| Residuo | 8 | 1,2156 | 0,152 | | | |

Il denominatore di F è il residuo.

**(1) = testa la dipendenza di Y su X ovvero l'esistenza di una regressione**
**(2) = testa il parallelismo ovvero se i coefficienti di regressione sono simili.**

Risulta pertanto che:

1) la **regressione comune è altamente significativa** e può essere **rappresentativa** della relazione lineare tra Y (peso dell'organo) ed X (peso corporeo) per entrambi i gruppi;
2) i **coefficienti di regressione non sono significativamente diversi tra loro,** per cui si può utilizzare un coefficiente comune per stabilire la relazione lineare che lega Y ad X.

A questo punto, **verificate le assunzioni** della covarianza, ovvero la significatività della regressione e l'esistenza del parallelismo tra le rette di regressione dei vari gruppi presi in esame, è lecito utilizzare il valore del **coefficiente di regressione comune, $b_c$ = 0,0954, per aggiustare i valori delle Y'** (peso dell'organo) **assegnando a tutti gli animali lo stesso peso corporeo**, che è quello della **gran media** = **119,5**

Eliminando dal peso del fegato quella parte di variabilità dovuta al diverso peso corporeo degli animali inseriti all'interno dei due gruppi, resta ora solo quella dovuta all'effetto del trattamento ed all'errore (variabilità individuale intrinseca, ovvero indipendente dal peso).

Questi valori vengono quindi sottoposti all'**analisi della varianza** per verificare se le medie aggiustate dei due gruppi differiscono significativamente tra loro. Se così fosse, potremmo ora imputare questa differenza al diverso trattamento a cui sono stati sottoposti i due gruppi (lotto 1 controllo, lotto 2 trattati col farmaco) e quindi all'effetto del farmaco sul fegato degli animali.

**Y' aggiustate alla gran media 119,5**

**$Y' = Y + b_c*(\bar{\bar{x}} - x)$**

| Lotto 1 | Lotto 2 |
|---|---|
| 12,01084 | 13,15691 |
| 11,97995 | 13,00623 |
| 12,07398 | 13,47534 |
| 12,74309 | 12,49702 |
| 11,6813 | 12,87995 |
| 12,6168 | 13,07859 |
| **12,18433** | **13,01567** |

[esempio, per il primo animale: $Y_1' = 11,2 + 0,0954*(119,5 – 111) = 12,01084$]

Le medie aggiustate possono essere calcolate direttamente in base alla formula:

$\bar{y}_i' = \bar{y}_i - b_c*(\bar{x}_i - \bar{\bar{x}})$ =12,55 – 0.0954*(123,33 – 119,5) = 12,18433

**ANALISI VARIANZA**[2] **su Y' aggiustate** $r^2$= 0,604061

| *Variazione* | *SQ* | *gdl* | *MQ* | *F* | *Sign.* | *F crit* |
|---|---|---|---|---|---|---|
| Tra gruppi | 2,07341 | 1 | 2,07341 | 15,2564 | **0,0029** | 4,964591 |
| **In gruppi** | **1,359043** | 10 | **0,135904** | | | |
| Totale | 3,432453 | 11 | | | | |

**Il risultato è completamente diverso da quello ottenuto utilizzando i valori non aggiustati del peso degli organi, in quanto ora l'effetto dovuto al farmaco risulta altamente significativo!**
Ciò è dovuto al fatto che **le medie degli organi si sono maggiormente diversificate** e, al contempo, si è **ridotta la varianza d'errore (in gruppi: 1,749 vs 0,1359)** perché è stata scorporata quella parte dovuta al fattore di eterogeneità rappresentato dal differente peso degli animali che influenza anche il peso iniziale dell'organo nei vari animali.

2

*La procedura qui utilizzata è quella ritenuta didatticamente più efficace per capire la logica che sta sotto questo tipo di analisi. In molti soft statistici in effetti l'ancova è incorporata come opzione dell'anova; l'interazione tra la covariata e la variabile dipendente testa il parallelismo, e pertanto non deve risultare significativa, mentre i calcoli vengono fatti utilizzando altre procedimenti.*