16 Il Modello lineare multivariato

Nella pratica, è estremamente comune la situazione in cui si cerchi di stimare la relazione esistente tra una particolare variabile (che chiameremo genericamente Y) ed un gruppo di altre (che chiameremo genericamente $X_1, ..., X_k$), dove queste ultime sono considerate esplicative del comportamento della Y. Si tratta di un caso particolare di un ambito di ricerca al momento estremamente attivo, e che viene spesso indicato come Supervise Learning; in generale, lo studio di relazioni come $Y = f(X)$ va anche sotto il nome di Machine Learning o Statistical Learning. In questo corso, ci limiteremo a trattare il caso più comune, che è anche il più semplice; quello in cui assumiamo che $f(.)$ abbia una forma lineare in un vettore di parametri β.

Spiegazione Semplice

Cos'è il "Modello Lineare"? (Regressione)
Questo capitolo introduce la **regressione lineare**, probabilmente il metodo più usato in tutta la statistica.
L'idea è semplice: abbiamo una variabile "risposta" Y (es. il prezzo di una casa) e vogliamo "spiegarla" o "prevederla" usando altre variabili "esplicative" X (dette anche *regressori*), come la superficie (X_1), il numero di stanze (X_2), ecc.
Il modello "lineare" ipotizza che la relazione tra Y e le X sia una linea (o un "iperpiano" in più dimensioni).

In particolare, supponiamo che valga la seguente relazione, per $i = 1, 2, ..., n$:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k + \epsilon_i$$

dove le variabili ϵ_i sono considerate dei "residui" o degli "errori" e catturano tutto quello che non è spiegato dalla relazione lineare tra le y e le x_i. In forma matriciale, possiamo scrivere

$$Y = X\beta + \epsilon, \tag{1}$$

dove Y è un vettore $n \times 1$ della forma $Y = (y_1, ..., y_n)^T$, X è una matrice $n \times k$ le cui colonne sono costituite da $(x_{j1}, ..., x_{jn})^T$ e ϵ è un vettore di residui; iniziamo supponendo che si tratti di residui Gaussiani, con valor medio nullo e matrice di varianza/covarianza $E[\epsilon\epsilon^T] = \Omega$. Comunemente la prima colonna di X viene scelta come il vettore di costanti $(1, 1, ..., 1)$ così ad esempio per $k = 2$ abbiamo

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i.$$

Spiegazione Semplice

Le due equazioni (Singola e Matriciale)

- **Equazione per y_i:** Questa è l'equazione per una *singola* osservazione (es. la i-esima casa). Dice che il suo prezzo y_i è una somma pesata dei suoi regressori (es. $\beta_1 \times \text{superficie}_i + \beta_2 \times \text{stanze}_i \dots$) più un termine di errore ϵ_i. I β sono i "pesi" (parametri) che vogliamo stimare.
- **Equazione Matriciale $Y = X\beta + \epsilon$:** Questa è solo un modo compatto per scrivere *tutte* le n equazioni (per tutte le n case) in un colpo solo.
  - Y: Un vettore colonna con tutte le risposte $(y_1, \dots, y_n)$.
  - X: Una matrice (la "matrice di design") dove ogni riga è un'osservazione (una casa) e ogni colonna è un regressore (superficie, stanze, ecc.).
  - β: Un vettore colonna con tutti i pesi $(\beta_1, \dots, \beta_k)$ che vogliamo trovare.
  - ϵ: Un vettore colonna con tutti gli errori $(\epsilon_1, \dots, \epsilon_n)$.

L'equazione $y_i = \beta_1 + \beta_2 x_i + \epsilon_i$ è il caso più semplice: una retta, dove β_1 è l'intercetta (il valore quando $x = 0$) e β_2 è la pendenza.

Abbiamo bisogno di una condizione ulteriore sulle variabili X:

Condition 128. *Le variabili X sono deterministiche ed il rango della matrice è esattamente pari a k.*

Ambedue le condizioni possono essere rimosse ed effettivamente lo sono nella ricerca più recente; si tratta però di ipotesi semplificatrici che costituiscono un punto di partenza naturale. Si noti che questa condizione implica immediatamente $k \leq n$; questa ipotesi in particolare viene abbandonata negli studi più recenti (high-dimensional statistics/big data).

Spiegazione Semplice

Condizioni Tecniche

- **"X sono deterministiche":** Questa è un'ipotesi semplificativa. Significa che non consideriamo le X (es. la superficie delle case) come variabili casuali, ma come numeri fissi, "dati".
- **"Rango della matrice pari a k":** Questo è fondamentale. Significa che $k \leq n$ (abbiamo più osservazioni che parametri da stimare) e che **non ci sono regressori ridondanti**. Non possiamo, ad esempio, includere la "superficie in m^2" (X_1) e la "superficie in $piedi^2$" (X_2) come due regressori separati, perché uno è un multiplo dell'altro e l'informazione è la stessa.

Proposition 129. *Lo stimatore di massima verosimiglianza nel modello 1 ha la seguente forma:*

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T Y$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}(\hat{\epsilon}^T \hat{\epsilon}) \quad \hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta}$$

Si ha inoltre che

$$\hat{\beta}_{MLE} \triangleq N(\beta, \sigma^2 (X^T X)^{-1}),$$

$$n \times \frac{\hat{\sigma}^2_{MLE}}{\sigma^2} \sim \chi^2_{n-k}$$

Spiegazione Semplice

La Soluzione: Lo Stimatore OLS/MLE
Questa proposizione ci dà la "risposta" al problema della regressione.

- **Come stimare β?** La "ricetta" (stimatore) per β è $\hat{\beta} = (X^T X)^{-1} X^T Y$.
- **Perché questa formula?**
  1. **(OLS) Minimi Quadrati Ordinari:** Questa formula è quella che *minimizza la somma dei quadrati degli errori* ($\sum \epsilon_i^2$). È la "migliore" linea che passa attraverso i dati. (Remark 130)
  2. **(MLE) Massima Verosimiglianza:** Se, e solo se, assumiamo che gli errori ϵ_i seguano una distribuzione *Gaussiana*, allora questa formula OLS è *anche* lo stimatore di Massima Verosimiglianza. (La dimostrazione lo mostra derivando la log-verosimiglianza Gaussiana).
- **La "Pagella" dello Stimatore $\hat{\beta}$:** Il teorema ci dice anche le proprietà di $\hat{\beta}$ (la sua "pagella"):
  - È **non-distorto**: la sua media è $N(\beta, \dots)$, quindi è centrato sul valore vero β.
  - È **Gaussiano**: la sua distribuzione campionaria segue una curva a campana (multivariata).
  - La sua **Varianza** (precisione) è $\sigma^2 (X^T X)^{-1}$.
- **Stimatore della Varianza $\hat{\sigma}^2$:** Ci dice che la varianza degli errori σ^2 (normalizzata) segue una distribuzione Chi-quadro (χ^2).

Dimostrazione. Notiamo innanzitutto che la funzione di verosimiglianza prende la forma

$$L(\beta, \sigma^2; Y, X) = \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} exp\{-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\},$$

$$\log L(\beta, \sigma^2; Y, X) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta).$$

Il calcolo del gradiente ci dà facilmente

$$\nabla_\beta \log L(\beta, \sigma^2; Y, X) = -\frac{1}{2\sigma^2}(-2X^TY + 2X^TX\beta)$$

$$\frac{\partial}{\partial\sigma^2}\log L(\beta, \sigma^2; Y, X) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T(Y - X\beta)$$

and setting these quantities to zero leads to the previous expressions for $\hat{\beta}_{MLE}$ $\hat{\sigma}^2_{MLE}$. It is also immediate to see that

$$\begin{aligned}\hat{\beta}_{MLE} &= (X^TX)^{-1}X^TY \\ &= (X^TX)^{-1}X^T(X\beta + \epsilon) \\ &= \beta + (X^TX)^{-1}X^T\epsilon\end{aligned}$$

whence

$$\begin{aligned}E[(\hat{\beta}_{MLE} - \beta)(\hat{\beta}_{MLE} - \beta)^T] &= E[(X^TX)^{-1}X^T\epsilon((X^TX)^{-1}X^T\epsilon)^T] \\ &= E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}] \\ &= (X^TX)^{-1}X^TE[\epsilon\epsilon^T]X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\sigma^2I_nX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}.\end{aligned}$$

La dimostrazione del risultato sullo stimatore della varianza richiede alcune ulteriori discussioni e sarà riportata più avanti. □

Remark 130. *Lo stimatore $\hat{\beta}_{MLE}$ è anche noto come OLS (Ordinary Least Squares).*

16.1 Il rapporto con i teoremi di proiezione

Lo stimatore di massima verosimiglianza nel modello lineare multivariato si presta ad una importante interpretazione usando gli strumenti dell'algebra lineare. In particolare, notiamo che il valor medio di Y dato X è dato da $E[Y] = X\beta$; il valore di β è ignoto, ma è naturale definire il valore $\hat{Y} := X\hat{\beta}$ come il valore "previsto" per il vettore Y sulla base delle stime $\hat{\beta}$ e del valore dei regressori X. Analogamente, il vettore degli "errori" $\epsilon = Y - X\beta$ si può stimare

$$\hat{\epsilon} = Y - X\hat{\beta}.$$

Notiamo ora che

$$\hat{Y} := X\hat{\beta} = X(X^TX)^{-1}X^TY = P_XY,$$
$$P_X := X(X^TX)^{-1}X^T.$$

Spiegazione Semplice

La "Hat Matrix" (Matrice Cappello)
Questa è un'interpretazione geometrica molto potente.
Il vettore Y (i nostri dati osservati) vive in uno spazio a n dimensioni. Lo spazio "permesso" per le nostre previsioni (lo spazio "generato dalle colonne di X") è un sottospazio più piccolo (a k dimensioni).
La formula $\hat{Y} = X(X^TX)^{-1}X^TY$ può essere riscritta come $\hat{Y} = P_XY$.

La matrice $P_X = X(X^TX)^{-1}X^T$ è chiamata la "Hat Matrix" (matrice cappello) perché, se moltiplicata per il vettore Y, "gli mette il cappello" ($\hat{Y}$).
Geumetricamente, P_X è una **matrice di proiezione**: "schiaccia" il vettore Y (dati veri) sul sottospazio X (lo spazio delle previsioni possibili), trovando il punto $\hat{Y}$ in quello spazio che è *più vicino* a Y. Questo è esattamente ciò che fa la minimizzazione dei quadrati.

E' immediato verificare che P_X è una matrice di proiezione, cioè è simmetrica ed idempotente; infatti

$$P_X^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T$$

In particolare, l'azione della matrice P_X (che ha dimensioni $n \times n$) corrisponde a proiettare il vettore Y sullo spazio vettoriale di dimensione k generato dalle colonne di X (ricordiamo che queste colonne sono linearmente indipendenti per ipotesi). Analogamente abbiamo

$$\hat{\epsilon} = Y - X\hat{\beta} = M_X Y$$

$$M_X = I - P_X = I - X(X^TX)^{-1}X^T.$$

Anche M_X è simmetrica ed idempotente:

$$M_X^2 = (I - P_X)^2 = I - 2P_X + P_X^2 = I - 2P_X + P_X = M_X$$

In particolare, l'azione di M_X consiste nel proiettare Y nello spazio ortogonale a quello generato dalle colonne di X. Questo ha alcune conseguenze importanti; si ha infatti

$$M_X P_X = P_X M_X = 0,$$

dove intendiamo con 0 la matrice $n \times n$ costituita da tutti zeri. Come ulteriore conseguenza, notiamo che

$$X^T\hat{\epsilon} = 0,$$

ed in particolare il vettore dei residui stimati è ortogonale a qualsiasi vettore che giaccia nello spazio generato dalle colonne di X.

Spiegazione Semplice

I Residui e l'Ortogonalità

- $\hat{\epsilon} = Y - \hat{Y} = (I - P_X)Y = M_X Y$**:** Se P_X proietta Y sullo spazio X per ottenere $\hat{Y}$ (la previsione), allora $M_X = I - P_X$ fa l'opposto: calcola ciò che rimane, $\hat{\epsilon}$ (il residuo).
- $X^T\hat{\epsilon} = 0$ **(Ortogonalità):** Questo è un risultato cruciale. Dice che il vettore dei residui $\hat{\epsilon}$ è "ortogonale" (perpendicolare, in senso geometrico) allo spazio dei regressori X.

Remark 131. *Supponiamo che la matrice X contenga un termine costante, cioè una colonna della forma $e = (1, ..., 1)^T$. Allora abbiamo $e^T\hat{\epsilon} = 0$, cioè $\sum_{i=1}^n \hat{\epsilon}_i = 0$; in altre parole, la somma dei residui di regressione è sempre nulla se tra i regressori è incluso un termine costante.*

Spiegazione Semplice

La somma dei residui è 0
Questa è la conseguenza pratica più famosa dell'ortogonalità. Se il tuo modello include un'intercetta (la colonna di "1", X_1), la condizione $X^T\hat{\epsilon} = 0$ implica che la colonna "intercetta" moltiplicata per i residui fa 0.
$(1, 1, \ldots, 1) \cdot (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^T = \sum \hat{\epsilon}_i = 0$.
In qualsiasi regressione OLS con un'intercetta, la somma (e quindi la media) degli errori di previsione (residui) è *sempre* esattamente zero.

Remark 132. *Il fatto che lo stimatore di massima verosimiglianza nel modello lineare multivariato coincida con la soluzione di un problema puramente geometrico di proiezione su sottospazi vettoriali è assolutamente degno di nota. Si tratta dell'ennesima sorprendente proprietà della legge Gaussiana.*

Possiamo ora enunciare le seguenti ulteriori proprietà delle matrici P_X M_X

Lemma 133. *Le matrici P_X, M_X hanno tutti autovalori pari a zero o uno e rango k, $n-k$ rispettivamente.*

Dimostrazione. Poichè le matrici sono reali e simmetriche, possiamo diagonalizzarle come

$$P_X = Q\Lambda Q^T, \text{ con } QQ^T = Q^TQ = I_n$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & 0 & ... \\ ... & 0 & ... & 0 \\ 0 & ... & 0 & \lambda_n \end{pmatrix}$$

Si ha allora

$$P_X^2 = Q\Lambda Q^T Q\Lambda Q^T = Q\Lambda^2 Q^T = Q\Lambda Q^T,$$

da cui segue che necessariamente $\lambda_i = 0$ o 1, per $i = 1, 2, ..., n$; ragionamento identico si applica a M_X. Per quello che riguarda il rango, notiamo che esso eguaglia il numero di autovalori diversi da zero (con molteplicità), quindi nel caso di matrici di proiezioni la traccia (cioè la sommadi autovalori, uguale al numero di quelli che valgono 1). Ricordando che $Tr(AB) = Tr(BA)$, possiamo scrivere

$$Tr(P_X) = Tr(X(X^TX)^{-1}X^T) = Tr((X^TX)^{-1}X^TX) = Tr(I_k) = k = Rg(P_X).$$

La dimostrazione per M_X è identica. □

Spiegazione Semplice

Proprietà delle Matrici di Proiezione (Tecnico)
Questo lemma analizza le matrici P_X e M_X.

- **Idempotente ($P_X^2 = P_X$):** "Proiettare" qualcosa che è *già* stato proiettato non cambia nulla. (Come "appiattire" qualcosa che è già piatto). La prova mostra che questo implica che gli autovalori (le "forze" della trasformazione) possono essere solo 0 o 1.
- **Rango (Traccia):** La traccia di una matrice è la somma della sua diagonale (e anche dei suoi autovalori). Per una matrice di proiezione, la traccia conta quanti autovalori sono 1.
- $Tr(P_X) = k$**:** Lo spazio delle "previsioni" ha k dimensioni (una per ogni regressore).
- $Tr(M_X) = n - k$**:** Lo spazio dei "residui" ha $n - k$ dimensioni. Questo $n - k$ è il famoso numero di **"gradi di libertà"** che si usa per i test statistici (es. test t) nella regressione.

Siamo ancora in grado di concludere la dimostrazionne sul comportamento dello stimatore di massima verosimiglianza della varianza. Ricordiamo innanzitutto la densità delle variabili aleatorie $\Gamma(\alpha, \beta)$:

$$f_\Gamma(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} t^{\alpha-1} exp(-\frac{t}{\beta}) \mathbb{I}_{[0,\infty)}(t),$$

(Nota: $\beta^{\alpha-1}$ nel testo originale è stato corretto in β^α per la forma standard della Gamma)

a cui corrisponde un valor medio pari a $\alpha\beta$ ed una varianza pari a $\alpha\beta^2$. Ricordiamo che la variabile chi-quadro con p gradi di libertà corrisponde ad una Gamma di parametri $\alpha = \frac{p}{2}$, $\beta = 2$, e quindi con valor medio p e varianza 2p. La proprietà fondamentale dellle chi-quadro e che esse sono uguali in distribuzioni alla somma di p Gaussiane standard indipendneti elevate al quadrato:

$$\chi_p^2 = Z_1^2 + ... + Z_p^2, \quad Z_i \triangleq N(0, 1).$$

Poniamo ora senza perdita di generalita $\sigma^2 = 1$, e notiamo che

$$\hat{\epsilon}^T\hat{\epsilon} = (M_X\epsilon)^T M_X\epsilon = \epsilon^T M_X^2\epsilon = \epsilon^T M_X\epsilon.$$

Ricordiamo ora la diagonalizzazione $M_X = Q\tilde{\Lambda}Q^T$, dove $\tilde{\Lambda}$ ha $n-k$ elementi pari ad uno sulla diagonale principale ed i restanti tutti nulli. Otteniamo quindi

$$\epsilon^T M_X\epsilon = \epsilon^T Q\tilde{\Lambda}Q^T\epsilon$$

$$= u^T \tilde{\Lambda} u \quad , con\ u := Q^T \epsilon \sim N(0, I_n),$$

$$= \sum_{i=1}^{n} \tilde{\lambda}_i u_i^2$$

$$= \sum_{i=1}^{n-k} u_i^2$$

perchè la legge Gaussiana standard è invariante per rotazioni. L'ultimo termine ha una legge chi-quadro per quanto scritto sopra, il che completa la dimostrazione.

Spiegazione Semplice

Perché la varianza segue una Chi-Quadro (χ^2)?
Questa è la dimostrazione della seconda parte della Proposizione 129 ($\hat{\sigma}^2 \sim \chi^2_{n-k}$).

1. Lo stimatore $\hat{\sigma}^2$ dipende dalla "somma dei quadrati dei residui", $\hat{\epsilon}^T \hat{\epsilon}$.
2. L'algebra mostra che $\hat{\epsilon}^T \hat{\epsilon} = \epsilon^T M_X \epsilon$.
3. La matrice M_X (come visto nel Lemma 133) è una proiezione ortogonale su uno spazio di dimensione $n - k$.
4. Una proiezione ortogonale di un vettore di Gaussiane standard ϵ (qui $u = Q^T \epsilon$) è ancora un vettore di Gaussiane standard.
5. L'operazione $\epsilon^T M_X \epsilon$ "seleziona" $n - k$ di queste Gaussiane standard, le eleva al quadrato e le somma.
6. Per definizione, la **somma di $n - k$ Gaussiane standard al quadrato** è una variabile **Chi-Quadro con $n - k$ gradi di libertà**.

Example 134. *Si noti che anche il problema della stima del valor medio per variabili indipendenti $Y_i \sim N(\mu, \sigma^2)$ può essere riletta in termini di regressione. Possiamo infatti scrivere*

$$Y = \begin{pmatrix} 1 \\ 1 \\ ... \\ ... \\ 1 \end{pmatrix} \mu + \epsilon. \quad \epsilon = Y - \begin{pmatrix} 1 \\ ... \\ ... \\ 1 \end{pmatrix} \mu;$$

lo stimatore diventa

$$\hat{\mu} = (e^T e)^{-1} e^T Y = (\sum_{i=1}^{n} 1)^{-1} \times \sum_{i=1}^{n} (1 \cdot Y_i) = \overline{Y}_n$$

con

$$\hat{\mu} \sim N(\mu, \sigma^2 (e^T e)^{-1}) = N(\mu, \frac{\sigma^2}{n})$$

Nel caso $k = 2$ (con costante) il modello diviene

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

ed un poco di algebra mostra che

$$\hat{\beta}_1 = \overline{Y}_n - \hat{\beta}_2 \overline{X}_n$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2}$$

con

$$\hat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2})$$

16.2 Lo stimatore GLS (Generalized Least Squares)

Possiamo ora generalizzare il modello che abbiamo studiato sinora, immaginando che i residui ϵ abbiano una struttura di dipendenza molto più complessa di variabili indipendenti. In particolare, ipotizziamo che $E[\epsilon\epsilon^T] = \Omega$ con matrice positiva definita di rango (pieno) n. La funzione di verosimiglianza prende la forma

$$L(\beta; Y, X) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{det(\Omega)}} exp\{-\frac{1}{2}(Y - X\beta)^T \Omega^{-1}(Y - X\beta)\}$$

(Nota: il testo originale presenta $\Omega^{-1/2}$ e $(Y - X\beta)^T$ due volte, è stato corretto)

Potremmo procedere come nel Caso ordinario, ma c'è una strategia più semplice. La matrice Ω si può diagonalizzare come $\Omega = Q\Lambda_\Omega Q^T$ per qualche matrice ortonormale Q che non corrisponde a quelle che abbiamo introdotto prima. Possiamo definire quindi $\Omega^{-1/2} = Q\Lambda_\Omega^{-1/2} Q^T$ con l'ovvia proprietà che

$$\begin{aligned}
\Omega^{-1/2}\Omega\Omega^{-1/2} &= Q\Lambda_\Omega^{-1/2} Q^T Q\Lambda_\Omega Q^T Q\Lambda_\Omega^{-1/2} Q^T \\
&= Q\Lambda_\Omega^{-1/2}\Lambda_\Omega\Lambda_\Omega^{-1/2} Q^T \\
&= QQ^T \\
&= I_n
\end{aligned}$$

Possiamo dunque definire il vettore $\tilde{\epsilon} := \Omega^{-1/2}\epsilon$ che (si verifica facilmente) ha matrice di varianza/covarianza pari all'identità. Quindi

$$Y = X\beta + \epsilon \mapsto \Omega^{-1/2}Y = \Omega^{-1/2}X\beta + \Omega^{-1/2}\epsilon \mapsto \tilde{Y} = \tilde{X}\beta + \tilde{\epsilon},$$

con

$$\tilde{Y} := \Omega^{-1/2}Y, \quad \tilde{X} := \Omega^{-1/2}X \quad \tilde{\epsilon} := \Omega^{-1/2}\epsilon.$$

Lo stimatore diventa quindi

$$\tilde{\beta}_{GLS} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}Y,$$

che è uno stimatore non distorto con legge Gaussiana e matrice di varianza e covarianza

$$E[(\tilde{\beta}_{GLS} - \beta)(\tilde{\beta}_{GLS} - \beta)^T] = (X^T\Omega^{-1}X)^{-1}.$$

(Nota: il testo originale riporta Ω^{-1}, la forma corretta è $(X^T\Omega^{-1}X)^{-1}$)

Spiegazione Semplice

Cosa succede se gli errori NON sono i.i.d.? (GLS)
Il modello OLS standard (Prop. 129) assume che gli errori ϵ siano indipendenti e con la stessa varianza σ^2 (cioè $E[\epsilon\epsilon^T] = \sigma^2 I$).
Il problema: Cosa succede se gli errori sono *correlati* tra loro (es. dati finanziari) o hanno *varianze diverse* (eteroschedasticità)? In questo caso, la matrice Ω non è più I (l'identità), ma una matrice complessa.
La Soluzione (GLS - Generalized Least Squares): L'OLS $\hat{\beta} = (X^TX)^{-1}X^TY$ non è più lo stimatore migliore (non è più "efficiente").
La soluzione GLS consiste in un "trucco" algebrico:

1. Troviamo una "radice quadrata" $\Omega^{-1/2}$ della matrice Ω^{-1}.
2. **Trasformiamo i dati:** Moltiplichiamo *tutto* (sia Y che X) per questa matrice: $\tilde{Y} = \Omega^{-1/2}Y$ e $\tilde{X} = \Omega^{-1/2}X$.
3. **Magia:** Questo "sbianca" (whitens) gli errori. I nuovi errori $\tilde{\epsilon}$ *sono* i.i.d. (hanno matrice I).
4. **Conclusione:** Applichiamo la *vecchia* formula OLS ai *nuovi* dati $(\tilde{Y}, \tilde{X})$: $\hat{\beta}_{GLS} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}$ Sostituendo $\tilde{Y}$ e $\tilde{X}$, si ottiene la formula finale del GLS: $\hat{\beta}_{GLS} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}Y$.

Remark 135. *Lo stimatore GLS può essere visto come i coefficienti di proiezione su un sottospazio generato dalle colonen di X quando la proiezione sia effettuata con una metrica Riemanniana indotta dalla matrice positive definita Ω^{-1}.*

Exercise 136. *Calcolare la forma dello stimatore GLS quando la matrice Ω sia nulla al di fuori della diagonale, nel caso $k = 1$.*

Exercise 137. *Calcolare la matrice di varianza e covarianza dello stimatore OLS (non GLS) quando la matrice di varianza di ϵ sia Ω non diagonale (cioè quando lo stimatore OLS sia applicato assumendo ipotesi in realtà non soddisfatte).*