REGRESSIONE LINEARE MULTIPLA

Il modello della regressione lineare multipla

Il modello della regressione lineare multipla descrive la i-esima osservazione y_i della variabile dipendente come fuzione lineare delle k variabili $x_{i1}, x_{i2}, ..., x_{ik}$ più una v.c. $\varepsilon_i \sim N\left(0;\sigma^2\right)$ indipendente dalla v.c. $\varepsilon_j \left(i \neq j\right)$.

Possiamo quindi esprimere un campione di n osservazioni come

$$
\begin{aligned}
y_1 &= x_{11}\beta_1 + x_{12}\beta_2 + ... + x_{1k}\beta_k + \varepsilon_1 \\
y_2 &= x_{21}\beta_1 + x_{22}\beta_2 + ... + x_{2k}\beta_k + \varepsilon_2 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
y_n &= x_{n1}\beta_1 + x_{n2}\beta_2 + ... + x_{nk}\beta_k + \varepsilon_n
\end{aligned}
$$

oppure, con le notazioni matriciali,

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

oppure, in modo compatto,

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

dove $\mathbf{y}$ è un vettore $n \times 1$, $\mathbf{X}$ è una matrice $n \times k$, $\boldsymbol{\beta}$ è un vettore $k \times 1$, $\boldsymbol{\varepsilon}$ è un vettore $n \times 1$. La matrice di covarianza del vettore $\boldsymbol{\varepsilon}$ è

$$\mathrm{Cov}\,\boldsymbol{\varepsilon} = E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \begin{bmatrix} E\varepsilon_1^2 & E\varepsilon_1\varepsilon_2 & \dots & E\varepsilon_1\varepsilon_n \\ E\varepsilon_2\varepsilon_1 & E\varepsilon_2^2 & \dots & E\varepsilon_2\varepsilon_n \\ \dots & \dots & \dots & \dots \\ E\varepsilon_n\varepsilon_1 & E\varepsilon_n\varepsilon_2 & \dots & E\varepsilon_n^2 \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

dove I_n è la matrice identità $n \times n$.
I parametri del modello sono $\boldsymbol{\beta}$ e σ^2.

Stimatori di $\boldsymbol{\beta}$ e σ^2 ai minimi quadrati

Con il metodo dei minimi quadrati il valore b stimato del vettore $\boldsymbol{\beta}$ minimizza la quantità

$$\sum_{i=1}^{n}\left(y_i - x_{i1}b_1 - \ldots - x_{ik}b_k\right)^2 = \left(y - Xb\right)^T\left(y - Xb\right) =$$

$$= y^T y - 2b^T X^T y + b^T X^T Xb. \qquad (1)$$

Ricordiamo le regole di derivazione

$$\frac{\partial b^T a}{\partial b} = a, \quad \frac{\partial\left(b^T Ab\right)}{\partial b} = Ab + A^T b$$

dove b e a sono due vettori colonna, A è una matrice quadrata, e la derivata $\partial f/\partial b$ di una funzione scalare f rispetto al vettore b è definita come il vettore delle derivate di f rispetto alle componenti di b.
Minimizziamo la (1) ponendo il vettore delle derivate rispetto a b uguale al vettore nullo:

$$-2X^T y + X^T Xb + X^T Xb =$$
$$= -2X^T y + 2X^T Xb = 0$$

da cui

$$b = \left(X^T X\right)^{-1} X^T y = \left(X^T X\right)^{-1} X^T \left(X\beta + \varepsilon\right) =$$
$$= \left(X^T X\right)^{-1} X^T X\beta + \left(X^T X\right)^{-1} X^T \varepsilon =$$
$$= \beta + \left(X^T X\right)^{-1} X^T \varepsilon. \quad (2)$$

(stimatore ai minimi quadrati)

Affinché esista l'inversa della matrice $k \times k$ $X^T X$, la matrice X deve avere rango k.

La media di b è

$$E[b] = \beta + \left(X^T X\right)^{-1} X^T E[\varepsilon] = \beta,$$

e quindi *b è uno stimatore corretto di* β.

La matrice di covarianza di b è

$$\mathrm{Cov}b = E\left[(b - Eb)(b - Eb)^T\right] =$$

$$= E\left[\left(X^T X\right)^{-1} X^T \varepsilon\varepsilon^T X \left(X^T X\right)^{-1}\right] =$$

$$= \left(X^T X\right)^{-1} X^T \left(E\varepsilon\varepsilon^T\right) X \left(X^T X\right)^{-1} =$$

$$= \left(X^T X\right)^{-1} X^T X \left(X^T X\right)^{-1} \sigma^2 =$$

$$= \left(X^T X\right)^{-1} \sigma^2,$$

dove abbiamo sostituito $E\varepsilon\varepsilon^T = I\sigma^2$.

Teorema di Gauss Markov

Nel modello $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, in cui $E\varepsilon = 0$ e $E\varepsilon\varepsilon^T = I\sigma^2$, lo stimatore b di β dei minimi quadrati è il *migliore stimatore lineare non distorto.*

Dimostrazione. Sia Ay un qualsiasi altro stimatore lineare di β, dove A è una matrice $k\times n$ di costanti. Ay è non distorto se

$$E\left(Ay\right)=E\left[A\left(X\beta+\varepsilon\right)\right]=AX\beta=\beta$$

ossia $AX=I$.

La matrice di covarianza dello stimatore Ay è

$$\begin{aligned}&E\left(Ay-\beta\right)\left(Ay-\beta\right)^T=\\&=E\left(AX\beta+A\varepsilon-\beta\right)\left(AX\beta+A\varepsilon-\beta\right)^T=\\&=E\left(\beta+A\varepsilon-\beta\right)\left(\beta+A\varepsilon-\beta\right)^T=\\&=EA\varepsilon\varepsilon^T A^T=AA^T\sigma^2.\end{aligned}$$

Poiché A può sempre essere scritta come $\left(X^TX\right)^{-1}X^T+B$, la condizione $AX=I$ implica

$$\begin{aligned}&\left[\left(X^TX\right)^{-1}X^T+B\right]X=I\\&I+BX=I\\&BX=0.\end{aligned}$$

Sostituendo A con $\left(X^T X\right)^{-1} X^T + B$ la matrice di covarianza di Ay diventa

$$\begin{aligned}
&\mathrm{Cov}\left(Ay\right) = AA^T\sigma^2 = \\
&= \left[\left(X^T X\right)^{-1} X^T + B\right]\left[X\left(X^T X\right)^{-1} + B^T\right]\sigma^2 = \\
&= \left[\left(X^T X\right)^{-1} + \left(X^T X\right)^{-1} X^T B^T + BX\left(X^T X\right)^{-1} + BB^T\right]\sigma^2 = \\
&= \left[\left(X^T X\right)^{-1} + BB^T\right]\sigma^2
\end{aligned}$$

dove abbiamo utilizzato la condizione di non distorsione BX=0.
Poiché BB^T è definita positiva (se B non è la matrice nulla) la matrice di covarianza di Ay è maggiore[1] della matrice di covarianza $\left(X^T X\right)^{-1}\sigma^2$ dello stimatore dei minimi quadrati a meno che sia B=0.

[1] Il termine *migliore stimatore* significa che la matrice di covarianza di un qualunque altro stimatore lineare si ottiene dalla matrice di covarianza dello stimatore ai minimi quadrati aggiungendo una matrice definita positiva.

Osserviamo che nella dimostrazione del teorema non abbiamo utilizzato l'ipotesi che ε_i abbiano distribuzione normale.

Il *vettore dei residui della regressione* dei minimi quadrati è

$$
\begin{aligned}
& e = y - Xb = \\
& = X\beta + \varepsilon - X\left[\beta + \left(X^T X\right)^{-1} X^T \varepsilon\right] = \\
& = \left[I_n - X\left(X^T X\right)^{-1} X^T\right]\varepsilon.
\end{aligned}
$$

Uno stimatore non distorto per σ^2 è

$$s^2 = \left(n-k\right)^{-1} e^T e.$$

Mostriamo che s^2 è non distorto calcolando la media di $e^T e$:

$$Ee^T e =$$

$$= E\left\{\varepsilon^T \left[I_n - X\left(X^T X\right)^{-1} X^T\right]\left[I_n - X\left(X^T X\right)^{-1} X^T\right]\varepsilon\right\} =$$

$$= E\left\{\varepsilon^T \left[I_n - X\left(X^T X\right)^{-1} X^T\right]\varepsilon\right\} =$$

$$= E\left\{\mathrm{tr}\left(\varepsilon^T \left[I_n - X\left(X^T X\right)^{-1} X^T\right]\varepsilon\right)\right\} =$$

$$= E\left\{\mathrm{tr}\left[I_n - X\left(X^T X\right)^{-1} X^T\right]\varepsilon\varepsilon^T\right\} =$$

$$= \mathrm{tr}\left\{\left[I_n - X\left(X^T X\right)^{-1} X^T\right]E\varepsilon\varepsilon^T\right\} =$$

$$= \sigma^2\left[\mathrm{tr}I_n - \mathrm{tr}\left(X^T X\right)^{-1} X^T X\right] =$$

$$\sigma^2\left(n-k\right),$$

dove il simbolo **tr** indica la somma degli elementi della diagonale di una matrice quadrata e $\mathrm{tr}\left(AB\right) = \mathrm{tr}\left(BA\right)$.

Inoltre abbiamo utilizzato il fatto che la matrice $I_n - X\left(X^T X\right)^{-1} X^T$ sia idempotente, ossia *MM*=*M*.

Verifica di ipotesi su $\boldsymbol{\beta}$ *e* σ^2

Ricordiamo lo stimatore ai minimi quadrati

$$b = \beta + \left(X^T X\right)^{-1} X^T \varepsilon,$$

quindi $b - \beta$ è una funzione lineare di $\varepsilon_1, \dots, \varepsilon_n$ ed ha una distribuzione normale k-variata se $\varepsilon_1, \dots, \varepsilon_n$ hanno distribuzione normale. La matrice di covarianza di b è $\left(X^T X\right)^{-1} \sigma^2$.

Utilizziamo la distribuzione normale per costruire intervalli di confidenza e verificare ipotesi su β se σ^2 è nota. In generale σ^2 non è nota e viene stimata attraverso $s^2 = (n-k)^{-1} e^T e$.

Mostreremo che $e^T e / \sigma^2 \sim \chi^2(n-k)$, dove

$$e = \left[I_n - X\left(X^T X\right)^{-1} X^T \right] \varepsilon.$$

Diagonalizziamo la matrice

$$M = I_n - X\left(X^T X\right)^{-1} X^T,$$

ossia scriviamola come prodotto $CD_\lambda C^T$, dove D_λ è una matrice diagonale.

Ricordiamo che se A è una matrice $n \times n$, $\lambda_1,...,\lambda_n$ sono i suoi *autovalori* e $c_1,...,c_n$ i corrispondenti *autovettori-colonna* se

$$Ac_i = \lambda_i c_i \quad i = 1,...,n \quad \text{ossia}$$

$$A(c_1,...,c_n) = (\lambda_1 c_1,...,\lambda_n c_n) = (c_1,...,c_n) D_\lambda$$

dove D_λ è una matrice diagonale con gli elementi $\lambda_1,...,\lambda_n$ sulla diagonale.
Sia C la matrice le cui colonne sono gli autovettori (destri) della matrice A, ossia

$$AC = CD_\lambda$$

Se esiste l'inversa di C, moltiplichiamo la relazione precedente per C^{-1} e otteniamo

$$A = CD_\lambda C^{-1}. \qquad (3)$$

La diagonalizzazione di A può essere ottenuta anche utilizzando gli *autovettori-riga* b_i che soddisfano la relazione $b_i A_i = \lambda_i b_i$, $i = 1,...,n$ ossia, se B è la matrice le cui righe sono gli autovettori sinistri della matrice A

$$\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} A = BA = D_\lambda B,$$

$$\Rightarrow A = B^{-1} D_\lambda B.$$

Se A è simmetrica, la relazione $Ac_i = \lambda_i c_i$ può essere trasposta ottenendo $c_i^T A^T = c_i^T A = \lambda_i c_i^T$. Quindi $c_i^T = b_i$, ossia i vettori trasposti degli autovettori-colonna destri c_i sono i corrispondenti autovettori-riga sinistri b_i. Sostituendo allora B con C^T otteniamo

$$A = \left(C^T\right)^{-1} D_\lambda C^T. \quad (4)$$

Confrontando la (3) con la (4) troviamo che per una matrice simmetrica A, $C^T = C^{-1}$, che implica $C^T C = I$ e $CC^T = I$. La matrice C è in questo caso chiamata *ortogonale*: $C^T C = I$ significa che ogni colonna di C è ortogonale (non correlata) con ogni altra colonna di C, analogamente $CC^T = I$ significa che ogni riga di C è ortogonale (non correlata) con ogni altra riga di *C,* e che ogni riga e ogni colonna hanno lunghezza 1.

Diagonalizziamo *M*.

Poiché M è simmetrica, abbiamo $C^T = C^{-1}$ ossia $\left(C^T\right)^{-1} = \left(C^{-1}\right)^{-1} = C$ e dunque

$$M = CD_\lambda C^T.$$

Poiché *M* è idempotente, abbiamo

$$CD_\lambda C^T = M = MM = CD_\lambda C^T CD_\lambda C^T =$$

$$= CD_\lambda C^{-1} CD_\lambda C^T = CD_\lambda D_\lambda C^T = C\left(D_\lambda\right)^2 C^T$$

e quindi

$$D_\lambda = \left(D_\lambda\right)^2$$

che implica

$$\lambda_i = \lambda_i^2, \quad i = 1, \ldots, n.$$

Quest'ultima relazione è valida solo nel caso $\lambda_i = 0$ oppure $\lambda_i = 1$. Perciò gli autovalori di una matrice idempotente possono valere solo zero o uno. La traccia di *M* è allora

$$\mathrm{tr}M = \mathrm{tr}\left(CD_\lambda C^T\right) = \mathrm{tr}\left(D_\lambda C^T C\right) = \mathrm{tr}D_\lambda$$

che è uguale al numero di autovalori non nulli di *M*. Ma la traccia di *M* è anche uguale a *n-k*, come mostrato nel paragrafo precedente.
Il numero di autovalori non nulli di *M* è *n-k*; supponiamo, senza perdere generalità, che siano i primi *n-k* elementi della diagonale di D_λ.

La statistica di interesse può essere scritta come

$$e^T e = \varepsilon^T M \varepsilon = \varepsilon^T C D_\lambda C^T \varepsilon = \eta^T D_\lambda \eta =$$

$$= \begin{bmatrix} \eta_1^T & \eta_2^T \end{bmatrix} D_\lambda \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \eta_1^T \eta_1$$

dove abbiamo definito $\eta = C^T \varepsilon$ e η_1 come il vettore costituito dai primi *n-k* elementi di η. Poiché ε è normale, $\eta = C^T \varepsilon$ è normale multivariato. La matrice di covarianza di η è

$$E\eta\eta^T = E\left(C^T \varepsilon\varepsilon^T C\right) = C^T C \sigma^2 = I_n \sigma^2.$$

Questo significa che gli *n-k* elementi del vettore η_1 sono normali e non correlati, ciascuno con varianza σ^2. La somma dei quadrati $\eta_1^T \eta_1$ di questi *n*-*k* elementi, divisa per σ^2, è distribuita come una v.c. $\chi^2(n-k)$ e tale distribuzione può essere utilizzata per costruire intervalli di confidenza e verificare ipotesi su σ^2.

Per verificare ipotesi su β quando σ^2 non è nota utilizziamo il fatto che il vettore dei residui *e* e lo stimatore dei minimi quadrati *b* sono non correlati (e quindi indipendenti perché entrambi normali).

Quindi $E\left[(b-\beta)e^T\right] = 0$. Infatti

$$E\left[(b-\beta)e^T\right] =$$

$$= E\left\{\left(X^T X\right)^{-1} X^T \varepsilon\varepsilon^T \left[I - X\left(X^T X\right)^{-1} X^T\right]\right\} =$$

$$= \left(X^T X\right)^{-1} X^T \left(E\varepsilon\varepsilon^T\right)\left[I - X\left(X^T X\right)^{-1} X^T\right] =$$

$$= \left(X^T X\right)^{-1} X^T \left[I - X\left(X^T X\right)^{-1} X^T\right]\sigma^2 =$$

$$= \left[\left(X^T X\right)^{-1} X^T - \left(X^T X\right)^{-1} X^T X \left(X^T X\right)^{-1} X^T\right]\sigma^2 =$$

$$= 0.$$

Consideriamo l'ipotesi nulla $H_0 : \beta_i = \beta_{i0}$. Denotiamo la matrice $\left(X^T X\right)^{-1}$ con $C = \left(c_{ij}\right)$, da non confondere con la matrice C del paragrafo precedente.
Dalla (2) sappiamo che sotto l'ipotesi nulla $\left(b_i - \beta_{i0}\right) \sim N\left(0; c_{ii}\sigma^2\right)$. Il rapporto

$$\frac{\left(b_i - \beta_{i0}\right) \big/ \sqrt{c_{ii}\sigma^2}}{s/\sigma} = \frac{b_i - \beta_{i0}}{s\sqrt{c_{ii}}}$$

è un rapporto tra una v.c normale standard e la radice quadrata di una v.c. $\chi^2\left(n-k\right)$ divisa per i suoi gradi di libertà e dunque ha distribuzione $t\left(n-k\right)$ poiché numeratore e denominatore sono indipendenti. Il test t basato su questo rapporto è usato di frequente nella regressione: se rifiutiamo l'ipotesi nulla $\beta_i = 0$ concludiamo che la variabile x_i descrive in modo significativo la variabile y.
Se accettiamo l'ipotesi nulla, consideriamo x_i poco significativa e possiamo ometterla dall'equazione della regressione.

Consideriamo ora l'ipotesi nulla $r_1\beta_1 + r_2\beta_2 = r_0$ (una combinazione lineare dei coefficienti della regressione sia uguale ad una costante r_0). Sotto l'ipotesi nulla la v.c. $r_1 b_1 + r_2 b_2$ ha media $r_1\beta_1 + r_2\beta_2 = r_0$.
La sua varianza è

$$
\begin{aligned}
&\mathrm{Var}\left(r_1 b_1 + r_2 b_2\right) = \\
&= \mathrm{Var}\left(r_1 b_1\right) + \mathrm{Var}\left(r_2 b_2\right) + 2\mathrm{Cov}\left(r_1 b_1, r_2 b_2\right) = \\
&= r_1^2 \mathrm{Var} b_1 + r_2^2 \mathrm{Var} b_2 + 2 r_1 r_2 \mathrm{Cov}\left(b_1, b_2\right) = \\
&= \left(r_1^2 c_{11} + r_2^2 c_{22} + 2 r_1 r_2 c_{12}\right)\sigma^2.
\end{aligned}
$$

Il rapporto

$$
\frac{r_1 b_1 + r_2 b_2 - r_0}{\left(r_1^2 c_{11} + r_2^2 c_{22} + 2 r_1 r_2 c_{12}\right)^{1/2} s} \sim t\left(n-k\right)
$$

è un rapporto tra una v.c. normale standard e la radice quadrata di una v.c. $\chi^2\left(n-k\right)$ indipendente dal numeratore divisa per i suoi gradi di libertà e può essere utilizzato per verificare l'ipotesi nulla $r_1\beta_1 + r_2\beta_2 = r_0$.

Il test precedente di una combinazione lineare dei coefficienti della regressione può essere generalizzato a *m* combinazioni lineari.
Sia *R* una matrice $m \times k$ e sia *r* un vettore $m \times l$. Scriviamo le *m* combinazioni lineari in forma compatta $R\beta = r$.
Sotto l'ipotesi nulla *Rb* ha distribuzione normale *m*-variata con vettore media

$$E(Rb) = R(Eb) = R\beta = r$$

e matrice di covarianza

$$\mathrm{Cov}(Rb) = E\left[R(b-\beta)(b-\beta)^T R^T\right] =$$

$$= RE(b-\beta)(b-\beta)^T R^T =$$

$$= R\left(X^T X\right)^{-1} R^T \sigma^2.$$

Se un vettore normale m- variato z ha media zero e matrice di covarianza V, la forma quadratica $z^T V^{-1} z$ *ha distribuzione* $\chi^2(m)$.

Ritornando al test per l'ipotesi nulla $R\beta = r$ possiamo utilizzare la statistica

$$\frac{\left(Rb-r\right)^T \left[\mathrm{Cov}\left(Rb\right)\right]^{-1} \left(Rb-r\right)}{\sigma^2} =$$

$$= \frac{\left(Rb-r\right)^T \left[R\left(X^T X\right)^{-1} R^T\right]^{-1} \left(Rb-r\right)}{\sigma^2}$$

che ha distribuzione $\chi^2\left(m\right)$.

Se σ^2 non è nota

$$\frac{\left(Rb-r\right)^T \left[R\left(X^T X\right)^{-1} R^T\right]^{-1} \left(Rb-r\right)}{ms^2} \quad (5)$$

ha distribuzione $F\left(m; n-k\right)$.

Verifica dell'ipotesi generale lineare

Presentiamo ora un approccio diverso per la verifica dell'ipotesi $R\beta = r$ sui coefficienti β della regressione.

Sia b lo stimatore dei minimi quadrati di β senza imporre le m restrizioni lineari e sia b^* lo stimatore di β soggetto alle restrizioni $Rb^* = r$. La somma dei quadrati dei residui $y - Xb^*$ sarà maggiore della somma dei quadrati dei residui $y - Xb$.
Scomponiamo $y - Xb^*$ in due componenti

$$y - Xb^* = (y - Xb) + \left(Xb - Xb^*\right). \quad (6)$$

Se $\left(y - Xb^*\right)^T \left(y - Xb^*\right)$ è maggiore di $(y - Xb)^T (y - Xb)$ allora la restrizione $Rb^* = r$ su b^* rende l'adattamento alla regressione meno efficace e quindi è da rifiutare. Premoltiplicando la (6) per la sua trasposta otteniamo la relazione (7)

$$\left(y - Xb^*\right)^T \left(y - Xb^*\right) =$$

$$= (y - Xb)^T (y - Xb) + \left(Xb - Xb^*\right)^T \left(Xb - Xb^*\right).$$

La somma dei quadrati $(y - Xb)^T (y - Xb) = e^T e$ divisa per σ^2 ha distribuzione $\chi^2(n-k)$.

La somma dei quadrati $\left(Xb - Xb^*\right)^T \left(Xb - Xb^*\right)$ divisa per σ^2 ha distribuzione $\chi^2(m)$. Poiché

$$e^T X = 0$$

$$\left(y - Xb\right)^T \left(Xb - Xb^*\right) = e^T X \left(b - b^*\right) = 0$$

allora i due addendi della (7) sono indipendenti e dunque

$$\frac{\left(Xb - Xb^*\right)^T \left(Xb - Xb^*\right) \big/ m}{e^T e \big/ (n-k)} \sim F\left(m; n-k\right)$$

e la statistica F può essere utilizzata per testare l'ipotesi ed è identica alla(5). Per mostrarlo determiniamo lo stimatore dei minimi quadrati b^* di β soggetto alla restrizione $Rb^* = r$ mediante la funzione lagrangiana

$$L = \frac{1}{2}\left(y - Xb^*\right)^T \left(y - Xb^*\right) + \lambda^T \left(Rb^* - r\right)$$

dove λ è il vettore degli m moltiplicatori di Lagrange.

Derivando L otteniamo

$$\begin{cases} \dfrac{\partial L}{\partial b^*} = -X^T\left(y - Xb^*\right) + R^T\lambda = 0 \\ \dfrac{\partial L}{\partial \lambda} = Rb^* - r = 0 \end{cases}$$

Moltiplicando la prima equazione per $R\left(X^T X\right)^{-1}$ si ha

$$-R\left(X^T X\right)^{-1} X^T\left(y - Xb^*\right) + R\left(X^T X\right)^{-1} R^T\lambda = 0$$

che, risolta in λ, dà

$$\lambda =$$

$$= \left[R\left(X^T X\right)^{-1} R^T\right]^{-1}\left[R\left(X^T X\right)^{-1} X^T y - R\left(X^T X\right)^{-1} X^T Xb^*\right] =$$

$$= \left[R\left(X^T X\right)^{-1} R^T\right]^{-1}(Rb - r)$$

dove abbiamo usato la relazione $Rb^* = r$.

Sostituendo nella prima equazione il valore di λ ottenuto si ha

$$-X^T y + X^T Xb^* + R^T\left[R\left(X^T X\right)^{-1} R^T\right]^{-1}\left(Rb - r\right) = 0$$

che possiamo risolvere in b^*

$$b^* =$$

$$= \left(X^T X\right)^{-1} X^T y - \left(X^T X\right)^{-1} R^T \left[R\left(X^T X\right)^{-1} R^T \right]^{-1} (Rb - r) =$$

$$= b - \left(X^T X\right)^{-1} R^T \left[R\left(X^T X\right)^{-1} R^T \right]^{-1} (Rb - r)$$

Grazie a quest'ultima relazione otteniamo

$$\left(Xb - Xb^*\right)^T \left(Xb - Xb^*\right) =$$

$$= \left(b - b^*\right)^T X^T X \left(b - b^*\right) =$$

$$= \left\{ \left(X^T X\right)^{-1} R^T \left[R\left(X^T X\right)^{-1} R^T \right]^{-1} (Rb - r) \right\}^T X^T X \cdot$$

$$\cdot \left\{ \left(X^T X\right)^{-1} R^T \left[R\left(X^T X\right)^{-1} R^T \right]^{-1} (Rb - r) \right\} =$$

$$= (Rb - r)^T \left[R\left(X^T X\right)^{-1} R^T \right]^{-1} R\left(X^T X\right)^{-1} R^T \cdot$$

$$\cdot \left[R\left(X^T X\right)^{-1} R^T \right]^{-1} (Rb - r) =$$

$$= (Rb - r)^T \left[R\left(X^T X\right)^{-1} R^T \right]^{-1} (Rb - r)$$

che è identica al numeratore della (5).