Capitolo 1

Regressione

1.1 Premessa

Per elaborare delle teorie economiche occorre raggruppare le relazioni tra variabili in modo da formare un modello. Un modello statistico è una rappresentazione parsimoniosa, fedele e necessaria della realtà derivata dall'evidenza empirica e da deduzioni logiche. La realtà è osservata, si formulano delle ipotesi, si assumono delle relazioni di causa ed effetto tra le variabili di interesse, ci si avvale delle conoscenze a-priori derivanti da teorie. Ciò si traduce nella formalizzazione di un modello statistico, basato su una struttura probabilistica, che viene sottoposto ad inferenza mediante un'indagine campionaria.

La costruzione di un modello statistico si concretizza in tre fasi successive: *specificazione*, *stima dei parametri*, *verifica*. La fase più delicata è la ricerca di una corretta specificazione del modello. Sulla base di conoscenze a-priori derivanti da teorie, assunzioni, ipotesi, risultati sperimentali, si formula una relazione funzionale tra le variabili di interesse individuando la funzione $f(\cdot)$ che lega la variabile dipendente Y e le variabili esplicative o predittori $(X_1, \ldots, X_K)$. Lo statistico intro-

durrà un elemento stocastico nella specificazione del modello affinchè questo sia parsimonioso così da costituire un'approssimazione fedele della realtà, di sua natura sarà un modello non deterministico. La relazione funzionale più semplice tra due variabili è data dall'equazione di una retta così da ottenere:

$$Y = \beta_1 + \beta_2 X + u \tag{1.1}$$

dove i parametri sono β_1 e β_2, rispettivamente intercetta e coefficiente angolare della retta, mentre u è la componente aleatoria o stocastica che riassume il non spiegato teoricamente (le variabili omesse) così come l'errore di misurazione. L'obiettivo sarà quello di pervenire a delle stime dei parametri del modello e di verificare la bontà di adattamento del modello ai dati per una possibile generalizzazione delle affermazioni teoriche suffragate dall'evidenza empirica.

1.2 Il modello classico di regressione lineare semplice

1.2.1 Il modello e le ipotesi

Il modello di cui si tratta nel seguito è detto modello classico di regressione lineare semplice. Esso è detto *semplice* poiché considera la relazione tra due sole variabili a differenza di quello multiplo che include più predittori. Il termine *lineare* sarà riferito ai parametri indipendentemente dalle variabili che possono essere opportunamente trasformate. Il modello è poi basato su ipotesi dette *classiche*, che fanno di questo modello il punto di riferimento per altri modelli basati sulla rimozione di talune delle ipotesi classiche. Il modello, infine, è detto di *regressione* poiché con esso si intende stimare o predire il valor medio della variabile dipendente sulla base di valori prefissati della va-

riabile esplicativa, per cui si dice che la variabile dipendente regredisce verso la media al variare dei valori della variabile esplicativa.

Si supponga di studiare la spesa per consumo settimanale di un dato prodotto (i.e., la variabile dipendente Y) in funzione di diversi livelli di reddito (i.e., la variabile esplicativa X). Lo statistico dovrà scegliere la relazione che spieghi il valore atteso della distribuzione condizionata di Y dato il livello i-esimo di X distinguendo il caso discreto:

$$E(Y|X = x_i) = \sum yp(y|x_i) \tag{1.2}$$

dove $p(y|x_i)$ descrive la distribuzione di probabilità condizionata al livello i-esimo del reddito, dal caso continuo

$$E(Y|X = x_i) = \int yf(y|x_i)dy \tag{1.3}$$

dove $f(y|x_i)$ descrive la funzione di densità di probabilità condizionata al livello i-esimo del reddito.

Si può ipotizzare che nella popolazione la spesa media settimanale sia funzione lineare del reddito. Ciò si traduce nell'assumere che la rappresentazione cartesiana dei punti di coordinate date dal livello di reddito x_i e valore atteso della spesa $E(Y|X = x_i)$ sia descritta esattamente da una retta, detta di regressione, definita come

$$E(Y|x_i) = f(x_i) = \beta_1 + \beta_2 x_i \tag{1.4}$$

dove β_1 è l'intercetta e β_2 è il coefficiente di regressione che descrive anche la pendenza della retta. Invero, se si osserva un individuo con reddito pari a x_i e spesa per consumo pari a y_i, sarà naturale ritenere che questa spesa non coincida esattamente con il valore atteso del gruppo, ovvero sarà maggiore o minore del valore atteso, e tale scostamento sarà descritto da una variabile casuale denominata errore:

$$u_i = y_i - E(Y|x_i) \tag{1.5}$$

in quanto è strettamente legato al processo di estrazione casuale dell'individuo dalla popolazione. Pertanto, se si osserva un campione di n individui per i quali si hanno le osservazioni (x_i, y_i), il modello sarà definito come

$$y_i = E(Y|x_i) + u_i \tag{1.6}$$

dove $E(Y|x_i)$ costituisce la componente deterministica del modello e u_i la componente stocastica del modello che rende y_i realizzazione anch'essa di una variabile aleatoria. Assumendo la linearità rispetto a X il modello diventa:

$$y_i = \beta_1 + \beta_2 x_i + u_i \tag{1.7}$$

La v.c. u_i è detta errore e rappresenta non solo tutte le variabili omesse dal modello, ma anche un elemento di casualità fondamentale e non prevedibile del fenomeno stesso, oltre agli errori di misura che si sono potuti commettere all'atto della rilevazione dei dati. È opportuno considerare alcune ipotesi, dette classiche, sulla distribuzione di probabilità di questa pertubazione e sul modello in generale:

1. Il valore atteso di ciascuna v.c. errore è uguale a zero:

$$E(u_i) = 0 \Rightarrow E(y_i) = E[E(Y|x_i)] + E(u_i) = \beta_1 + \beta_2 x_i \tag{1.8}$$

il ché significa che non c'è errore sistematico. Questa ipotesi non è restrittiva in quanto un eventuale errore sistematico verrebbe incorporato nell'intercetta del modello;

2. La varianza dell'errore è costante:

$$var(u_i) = \sigma^2 \Rightarrow var(y_i) = \sigma^2, \forall i \tag{1.9}$$

per cui si dice che c'è omoschedasticità degli errori. Questa ipotesi è restrittiva per dati di tipo sezionale (*cross-section*) (n individui osservati al tempo t) ed è più realistica per le serie temporali (un individuo osservato n volte dal tempo t al tempo $t + n$). Infatti, se ad esempio si considera la spesa per consumo in funzione del reddito è lecito supporre che la variabilità della spesa sia crescente con il livello del reddito (eteroschedasticità);

3. La covarianza degli errori è uguale a zero:

$$cov(u_i, u_j) = E(u_i u_j) - E(u_i)E(u_j) = 0, \forall i \neq j \tag{1.10}$$

per cui gli errori sono incorrelati, ma non necessariamente indipendenti (salvo nel caso di normalità delle variabili). Questa ipotesi è scarsamente realistica per le serie temporali per le quali si osserva il fenomeno dell'autocorrelazione degli errori;

4. La variabile esplicativa X non è aleatoria, ovvero non è correlata con l'errore:

$$cov(x_i, u_i) = 0, \forall i \tag{1.11}$$

per cui si intende che il campione sia stato estratto dalle distribuzioni condizionate di Y dati i livelli della variabile X;

5. Il modello è correttamente specificato. Questa è un'ipotesi implicita del modello la cui plausibilità dipende fortemente dalle conoscenze a-priori del ricercatore. Se ad esempio si vuole stimare la relazione tra salario monetario e tasso di disoccupazione

come illustrata dalla ben nota curva di Phillips, e si sceglie erroneamente la retta si determinerebbero delle predizioni errate nel senso di sovrastimare in taluni casi e sottostimare in altri. Il problema è che nella pratica non si conoscono, come per la curva di Phillips, le variabili esatte da includere nel modello e la forma funzionale corretta che leghi tali variabili. Si formulano delle ipotesi sulla natura stocastica del modello e sulle variabili in esso incluse;

6. La varianza di X, supposta diversa da zero, non deve essere eccessivamente elevata, altrimenti un'analisi lineare condurrebbe a soluzioni non informative. Si immagini una rappresentazione cartesiana delle osservazioni per le quali il campo di variazione della X sia molto ampio: ciò significa che la nube di punti si disperde rispetto la direzione dell'asse delle ascisse e la retta di regressione avrà presumibilmente una pendenza pressocché nulla.

1.2.2 La stima dei parametri

Il modello di regressione (1.7) dovrà essere stimato al fine di pervenire ad una stima del valore atteso (1.4) indicata come:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{1.12}$$

dove $\hat{\beta}_1$ e $\hat{\beta}_2$ saranno le stime dei parametri. In tal modo, il dato osservato potrà esprimersi come somma del modello stimato e del residuo del modello:

$$y_i = \hat{y}_i + e_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + e_i \tag{1.13}$$

da cui si evince che il residuo $e_i = y_i - \hat{y}_i$ potrà interpretarsi come stima dell'errore.

La stima dei parametri è ottenuta attraverso il metodo dei minimi quadrati:

$$\min Q(\beta_1, \beta_2) = \sum_i (y_i - \beta_1 - \beta_2 x_i)^2 \tag{1.14}$$

ossia minimizzando la somma dei quadrati degli errori. Ciò si traduce nella risoluzione di un sistema di equazioni normali, eguagliando a zero le derivate prime della funzione $Q(\cdot)$ rispetto ai parametri:

$$\sum_i y_i = n\beta_1 + \beta_2 \sum_i x_i \tag{1.15}$$

$$\sum_i x_i y_i = \beta_1 \sum_i x_i + \beta_2 \sum_i {x_i}^2 \tag{1.16}$$

e controllando le condizioni del secondo ordine. Le stime dei minimi quadrati saranno date dalle seguenti espressioni:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \tag{1.17}$$

$$\hat{\beta}_2 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i {x_i}^2 - n\bar{x}^2} = \frac{Cod(X,Y)}{Dev(X)} = \frac{s_{xy}}{{s_x}^2} \tag{1.18}$$

dove $Cod(X,Y)$ e $Dev(X)$ sono rispettivamente la codevianza e la devianza, mentre s_{xy} e ${s_x}^2$ sono rispettivamente la covarianza campionaria tra X e Y e la varianza campionaria della X. Nel seguito, si utilizzerà la notazione $\hat{\beta}_1$ e $\hat{\beta}_2$ sia per le stime che per gli stimatori dei parametri β_1 e β_2, quali funzioni delle statistiche campionarie.

Sostituendo le (1.17) e (1.18) nella (1.14) si ottiene il valore minimo della funzione da ottimizzare:

$$Q(\hat{\beta}_1, \hat{\beta}_2) = \sum_i {e_i}^2 \tag{1.19}$$

da cui si evince che nel metodo dei minimi quadrati i residui maggiori, essendo i residui elevati al quadrato, contribuiscono in misura maggiore a determinare il valore minimo di questa funzione. Il metodo dei minimi quadrati gode delle seguenti proprietà:

1) La retta passa per il punto di coordinate $(\bar{x}, \bar{y})$, che si verifica sostituendo $\bar{x}$ nella (1.12) e tenendo conto della (1.17);

2) $E(y_i) = E(\hat{y}_i)$, $E(e_i) = 0$, $\sum_i e_i = 0$, che si dimostra sostituendo le stime $\hat{\beta}_1$ e $\hat{\beta}_2$ nella prima equazione (1.15) del sistema;

3) $\sum_i e_i x_i = 0$, che si deduce dopo aver sostituito le stime $\hat{\beta}_1$ e $\hat{\beta}_2$ nella seconda equazione (1.16).

Per valutare la precisione delle stime e in generale per l'inferenza sui parametri del modello occorre conoscere la varianza degli stimatori:

$$var(\hat{\beta}_1) = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{Dev(X)}] \tag{1.20}$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{Dev(X)} \tag{1.21}$$

la cui radice quadrata rappresenta l'errore standard della stima. Se la varianza dell'errore non è nota allora la sua stima corretta è data dalla seguente espressione:

$$\hat{\sigma}^2 = \frac{\sum_i {e_i}^2}{n-2} \tag{1.22}$$

Per quanto riguarda le proprietà degli stimatori dei minimi quadrati, si dimostra, con il teorema di Gauss-Markov, che essi sono lineari, non distorti e a varianza minima (*BLUE: Best, Linear, Unbiased Estimators*).

1.2.3 La bontà di adattamento lineare

La bontà di adattamento lineare del modello ai dati si potrebbe valutare osservando il valore minimo (1.19), in quanto tanto minore sarà tale valore tanto migliore sarà l'adattamento della retta ai dati osservati. Invero, poiché tale minimo potrà variare da caso a caso, dipendendo dall'unità di misura del fenomeno, occorre definire una misura relativa o quanto meno normalizzata per consentire anche eventuali confronti tra diversi esempi di applicazione del modello ai dati. In effetti, si osserva che il minimo della funzione di ottimo è parte della seguente decomposizione della devianza totale di Y:

$$Dev(Y) = Dev(R) + Dev(E) \tag{1.23}$$

dove $Dev(R) = \sum_i (\hat{y}_i - \bar{y})^2$ è detta *devianza di regressione*, ossia la parte di devianza totale spiegata dalla retta di regressione, mentre $Dev(E) = \sum_i {e_i}^2$ è detta *devianza dei residui*. Infatti, dalla (1.13) aggiungendo e sottraendo la media $\bar{y}$,

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i \tag{1.24}$$

elevando ambo i membri al quadrato e sommando per gli n individui:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i {e_i}^2 + 2\sum_i (\hat{y}_i - \bar{y})e_i \tag{1.25}$$

si perviene alla (1.23) in quanto, utilizzando le (1.15) e (1.16), si dimostra che il doppio prodotto si annulla.

Si potrà definire l'indice di determinazione lineare per valutare la bontà di adattamento del modello lineare ai dati osservati considerando quanta parte della devianza totale è spiegata dalla retta di regressione:

$$R^2 = \frac{Dev(R)}{Dev(Y)} = 1 - \frac{Dev(E)}{Dev(Y)} \tag{1.26}$$

che per costruzione, quale rapporto di composizione, varierà da zero ad uno, esprimendo un buon grado di adattamento lineare qualora il suo valore è prossimo ad uno.

1.2.4 L'inferenza sui parametri

A fini inferenziali, si assume che gli errori si distribuiscono normalmente:

$$u_i \sim N(0, \sigma^2) \tag{1.27}$$

Si dimostra che questa assunzione implica che gli stimatori $\hat{\beta}_1$ e $\hat{\beta}_2$ si distribuiscono normalmente:

$$\hat{\beta}_1 \sim N(\beta_1, var(\hat{\beta}_1)) \tag{1.28}$$

$$\hat{\beta}_2 \sim N(\beta_2, var(\hat{\beta}_2)) \tag{1.29}$$

e pertanto si potrà far riferimento alla normale standardizzata per la costruzione degli intervalli di confidenza e per la verifica delle ipotesi.

Si osservi che poiché la varianza degli stimatori (1.20) e (1.21) dipende dalla varianza degli errori (1.22), questa non è usualmente nota e occorre stimarla con la (1.22) pervenendo a stime corrette della varianza degli stimatori. In tal caso, si dimostra che le statistiche campionarie

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \tag{1.30}$$

$$T_2 = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}} \tag{1.31}$$

si distribuiscono come una t-Student con $(n-2)$ gradi di libertà.

Inoltre, lo stimatore corretto della varianza dell'errore è legato alla distribuzione chi-quadrato con $(n-2)$ gradi di libertà:

$$X^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim {\chi_{n-2}}^2 \tag{1.32}$$

che potrà essere impiegata per l'inferenza su σ^2.

1.2.5 La previsione

Il problema che viene affrontato in questo paragrafo è quello della previsione di Y dato un nuovo livello x_0 della X. La previsione viene condotta considerando la retta stimata (1.12) e distinguendo il caso della previsione media, ossia stima del valore atteso data da $\hat{y}_0$, dal caso della previsione puntuale, stima del valore osservato y_0. In entrambi i casi, si utilizzerà quale stima BLUE l'espressione $\hat{\beta}_1 + \hat{\beta}_2 x_0$. Per avere un'idea dell'errore di previsione si dovrà considerare che la previsione si distribuirà normalmente con media $\beta_1 + \beta_2 X$ e varianza, nel primo caso, pari a:

$$var(\hat{y}_0) = \sigma^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i {x_i}^2}] \tag{1.33}$$

mentre, nel secondo caso, la varianza sarà maggiore essendo:

$$var(y_0) = \sigma^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i {x_i}^2}] \tag{1.34}$$

Se si sostituisce la stima corretta alla varianza dell'errore si utilizzerà la statistica t-Student con $(n-2)$ gradi di libertà per la costruzione

degli intervalli di confidenza della previsione. In generale, nel grafico che illustra la retta di regressione stimata, si illustra l'errore di previsione attraverso le cosiddette bande di confidenza della previsione media e della previsione puntuale per avere un'idea sull'accuratezza del modello: questa sarà tanto maggiore quanto più si è nei prossimi del punto medio di coordinate $(\bar{x}, \bar{y})$, mentre le bande si ampliano quando ci si allonta da tale valore così che si incrementa l'errore di previsione.

1.2.6 La valutazione dei risultati dell'analisi di regressione

L'analisi della regressione viene condotta distinguendo la variabile dipendente da quella esplicativa sulla base della teoria che si vuole verificare empiricamente. Successivamente, si stima la retta di regressione e occorrerà prestare particolare attenzione alla coerenza dei segni delle stime ottenute per l'intercetta e per il coefficiente di regressione rispetto alla teoria e alle ipotesi formulate. Si procederà poi ad analizzare i risultati del test e degli intervalli di confidenza delle stime per comprendere la significatività delle stime ottenute rispetto alle ipotesi nulle che rispettivamente ciascun parametro sia pari a zero. Si potranno poi sottoporre a test altre ipotesi nulle se si ha sufficiente informazione per presumere altri valori di ciascuno dei parametri. Talvolta, infatti, è bene effettuare più test con differenti ipotesi nulle in quanto nell'accettare un'ipotesi nulla bisogna essere consapevoli che un'altra ipotesi nulla può essere ugualmente compatibile con i dati. Per questo motivo, è preferibile dire che *si può accettare l'ipotesi nulla* piuttosto che dire che *la si accetta*. La bontà del modello lineare di addattarsi ai dati è valutata attraverso l'indice di determinazione lineare. Si vedrà nella regressione multipla che tale indice potrà incrementarsi se si aggiungono variabili esplicative nel modello e pertanto la rilevanza di

un suo valore alto avrà fondamento solo se accompagnata a valori del test significativi e soprattutto a bassi valori degli errori standard della stima.

1.3 Il modello classico di regressione lineare multipla

1.3.1 Il modello e l'interpretazione dei parametri

Si consideri il problema di spiegare la variabile dipendente Y attraverso $k-1$ variabili esplicative mediante il modello di regressione lineare multipla:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_j x_{ji} + \cdots + \beta_k x_{ki} + u_i \tag{1.35}$$

Il parametro β_1 è l'intercetta e rappresenta l'effetto medio di tutte le variabili escluse dal modello qualora fossero pari a zero tutti gli altri parametri. Il parametro β_j è il coefficiente di regressione parziale relativo alla variabile X_j, misurando il cambiamento in media di Y per una variazione unitaria di X_j mantenendo costanti i valori delle altre variabili.

Al fine di comprendere il significato dei coefficienti di regressione parziale, si consideri il classico esempio di spiegare la produzione Y in funzione del lavoro X_2 e del capitale X_3. Se si è interessati a valutare l'incremento della produzione dovuto all'incremento del lavoro si dovrà "controllare" l'effetto del capitale. Si procederà regredendo sia la Y che la X_2 rispetto alla variabile X_3 così da esprimere i valori osservati quali funzioni delle stime e dei residui come nella (1.13):

$$y_i = b_1 + b_{13} x_{3i} + e_{1i} \tag{1.36}$$

$$x_{2i} = b_2 + b_{23}x_{3i} + e_{2i} \tag{1.37}$$

dove b_1 è la stima dell'intercetta e b_{13} è la stima del coefficiente di regressione nella prima regressione, e analogamente b_2 e b_{23} nella seconda regressione. I residui possono essere espressi nel seguente modo:

$$e_{1i} = y_i - b_1 - b_{13}x_{3i} \tag{1.38}$$

$$e_{2i} = x_{2i} - b_2 - b_{23}x_{3i} \tag{1.39}$$

indicando, per la i-esima osservazione, il valore di Y dopo aver rimosso l'effetto lineare di X_3 ed il valore di X_2 dopo aver rimosso l'effetto lineare di X_3 rispettivamente. Se si regredisce ora il residuo della prima regressione rispetto al residuo della seconda regressione si determina l'equazione:

$$e_{1i} = c_1 + c_2 e_{2i} + e_{3i} \tag{1.40}$$

dove in particolare c_2 è la stima del coefficiente di regressione e misura l'effetto "netto" di un cambiamento unitario di X_2 su Y, ossia la produttività marginale del lavoro al netto dell'effetto capitale. In altre parole, c_2 coinciderebbe con la stima del coefficiente di regressione parziale relativo alla variabile lavoro nel modello di regressione lineare multipla.

1.3.2 Il modello in forma matriciale: le ipotesi e la stima

Il modello di regressione lineare multipla in forma matriciale si definisce nel seguente modo:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{1.41}$$

dove $\mathbf{y}$ è un vettore colonna di n osservazioni della variabile Y, $\mathbf{X}$ è una matrice di n righe e k colonne, di cui la prima è formata da tutti uno, contenente le osservazioni delle $k-1$ variabili esplicative sugli n individui, β è un vettore colonna di k parametri del modello, u è il vettore colonna di n errori.

Le ipotesi del modello classico potranno essere così scritte:

1) $E(\mathbf{u}) = \mathbf{0} \Rightarrow \mathbf{E}(\mathbf{y}) = \mathbf{X}\beta$;

2) $\Sigma_u = \sigma^2\mathbf{I} \Rightarrow \Sigma_y = \sigma^2\mathbf{I}$, dove Σ_u e Σ_y sono le matrici di varianze e covarianze degli errori e della variabile dipendente rispettivamente;

3) X non è stocastica;

4) $rango(\mathbf{X}) = k < n$, ossia la matrice $\mathbf{X}$ ha rango pieno, nel senso che non si può dedurre una variabile quale combinazione lineare delle altre variabili, altrimenti si dice che c'è multicollinearità;

5) il modello è correttamente specificato;

6) le varianze dei predittori non devono essere eccessivamente alte.

Il modello di regressione lineare multipla si stimerà con il metodo dei minimi quadrati:

$$Q(\beta) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{1.42}$$

per cui derivando rispetto al vettore dei parametri si otterrà la seguente stima:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.43}$$

che rappresenta una soluzione univocamente determinata se e solo se l'inversa di $\mathbf{X}'\mathbf{X}$ esiste, ossia le variabili sono indipendenti. Inoltre, si

dimostra che la matrice di varianze e covarianze di $\hat{\beta}$ è pari a $\Sigma_\beta = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Si potrà definire il vettore dei residui come

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{1.44}$$

dove $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$. Il vettore dei residui risulta essere un trasformazione lineare del vettore $\mathbf{y}$:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = [\mathbf{I} - \mathbf{H}]\mathbf{y} = \mathbf{M}\mathbf{y} \tag{1.45}$$

dove $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ e $\mathbf{M} = \mathbf{I} - \mathbf{H}$, ed è inoltre trasformazione lineare anche del vettore $\mathbf{u}$:

$$\mathbf{e} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{X}\beta + \mathbf{M}\mathbf{u} = \mathbf{M}\mathbf{u} \tag{1.46}$$

essendo $\mathbf{MX} = \mathbf{0}$. Da un punto di vista geometrico, il vettore dei residui è ortogonale al piano generato dalle colonne di $\mathbf{X}$ poiché $\mathbf{X}'\mathbf{e} = \mathbf{0}$, mentre la stima $\hat{\mathbf{y}}$ rappresenta la proiezione del vettore $\mathbf{y}$ su tale piano dove l'operatore di proiezione è $\mathbf{H}$, ossia $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. I residui hanno media pari a zero $E(\mathbf{e}) = \mathbf{0}$ e matrice di varianze e covarianze piena pari a $\Sigma_e = \sigma^2\mathbf{M}$, ossia i residui possono essere autocorrelati. La stima corretta della varianza degli errori è nuovamente data da:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} \tag{1.47}$$

che sostituita nell'espressione della matrice di varianze e covarianze di $\hat{\beta}$ determina la stima $S_\beta = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.

Analogamente al caso semplice, si potrà definire l'indice di determinazione lineare per il modello multiplo come rapporto tra devianza di regressione e devianza totale, ossia, in forma matriciale, si ha:

$$Dev(Y) = \mathbf{y}'\mathbf{y} - n\bar{y}^2 \tag{1.48}$$

$$Dev(R) = \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{y}^2 = \hat{\beta}'\mathbf{X}'\mathbf{y} - n\bar{y}^2 \tag{1.49}$$

da cui si deriva l'indice R^2 come:

$$R^2 = \frac{\hat{\beta}\mathbf{X}'\mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2} \tag{1.50}$$

che varierà tra zero ed uno. La radice quadrata dell'indice di determinazione lineare è noto come coefficiente di correlazione multipla ed è equivalente al coefficiente di correlazione lineare tra i valori osservati y_i ed i valori stimati $\hat{y}_i$ per $i = 1, \ldots, n$.

1.3.3 L'indice corretto di determinazione lineare

Quando si considera un modello multiplo, l'indice di determinazione lineare (1.26) aumenta (o quanto meno non diminuisce) al crescere del numero di variabili esplicative incluse nel modello. Infatti, il valore minimo (1.19) della funzione da ottimizzare in una regressione con k predittori sarà dato dalla seguente espressione:

$$Q(\hat{\beta}_1, \hat{\beta}_2)_k = \sum_i \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki}\right)^2 \tag{1.51}$$

mentre lo stesso valore minimo in una regressione con $(k+1)$ predittori sarà dato da

$$Q(\hat{\beta}_1, \hat{\beta}_2)_{(k+1)} = \sum_i \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki} - \hat{\beta}_{(k+1)} x_{(k+1)i}\right)^2 \tag{1.52}$$

Si osserva che $Q(\hat{\beta}_1, \hat{\beta}_2)_k \leq Q(\hat{\beta}_1, \hat{\beta}_2)_{(k+1)}$ potendo raggiungere lo stesso minimo se $\hat{\beta}_{(k+1)} = 0$. In altre parole, la devianza dei residui diminuisce al crescere del numero delle variabili e pertanto l'indice di

determinazione lineare (1.50) aumenta. In definitiva, un alto valore dell'indice R^2 non è indicatore di buon adattamento in quanto esso dipende anche dal numero di predittori inclusi nel modello.

Affinché si possano confrontare due regressioni con la stessa variabile dipendente ma con un diverso numero di predittori si dovrà considerare il seguente indice corretto:

$$\bar{R}^2 = 1 - \frac{Dev(E)/(n-k)}{Dev(Y)/(n-1)} \tag{1.53}$$

che, in luogo delle devianze, propone le stime corrette delle varianze con gradi di libertà dati rispettivamente da $(n-k)$ e $(n-1)$. In tal modo, è pur vero che la devianza dei residui diminuisce con l'aggiunta di un predittore, ma diminuiranno anche i corrispondenti gradi di libertà. Invero, l'indice corretto non sarà necessariamente compreso tra zero ed uno, ma esso opera una correzione significativa all'indice R^2 qualora il numero di variabili esplicative è elevato in rapporto al numero di individui osservati. Infine, si dimostra che vale la seguente relazione:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{1.54}$$

che lega l'indice non corretto all'indice corretto.

1.3.4 Le correlazioni semplici e parziali

Quando si considera un modello di regressione multipla è interessante analizzare la matrice delle correlazioni semplici tra le variabili indicata con $\mathbf{R}$ di termine generico r_{lj}, tale che $r_{lj} = 1$ se $l = j$ mentre r_{1j} esprime la correlazione semplice tra la variabile dipendente Y e ciascun predittore X_j per $j = 2, \ldots, k$. Si dimostra che

$$\hat{\beta}_j = -\frac{s_1}{s_j}\frac{\Re_{1j}}{\Re_{11}} \tag{1.55}$$

dove s_1 e s_j sono le deviazioni standard della Y e della X_j rispettivamente, mentre $\Re_{1j}$ e $\Re_{11}$ sono i cofattori di r_{1j} e r_{11} rispettivamente. Il coefficiente di correlazione multipla è definito nel seguente modo:

$$R^2_{1.23\ldots k} = 1 - \frac{R}{\Re_{11}} \tag{1.56}$$

dove $R = det(\mathbf{R})$ è il determinante della matrice delle correlazioni.

Si consideri l'esempio in cui si hanno due predittori ed una variabile dipendente. Il coefficiente di correlazione parziale tra la variabile dipendente Y ed il predittore X_2, ponendo costante il livello del predittore X_3, è definito nel modo seguente:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}} \tag{1.57}$$

ed in maniera analoga si potrà esprimere la correlazione parziale tra Y e X_3 data la X_2. La correlazione parziale esprime una relazione diversa da quella espressa dalla correlazione semplice. Infatti, pur in presenza di una correlazione semplice tra Y e X_2 pari a zero, ossia $r_{12} = 0$, le due stesse variabili potrebbero risultare positivamente correlate parzialmente rispetto ad una terza variabile X_3, ossia $r_{12.3} > 0$, qualora si avesse $r_{13} > 0$ e $r_{23} < 0$.

Si dimostrano le seguenti relazioni tra l'indice di determinazione lineare (ossia il quadrato del coefficiente di correlazione multipla), i coefficienti di correlazione semplice e parziale:

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \tag{1.58}$$

$$R^2_{1.23} = r^2_{12} + (1 - r^2_{12})r^2_{13.2} \tag{1.59}$$

$$R^2_{1.23} = r^2_{13} + (1 - r^2_{13}) r^2_{12.3} \tag{1.60}$$

Si evince in tal modo che l'indice di determinazione lineare può solo aumentare (e non diminuire) con l'ingresso di un predittore nel modello. Infatti, l'indice (1.59) è costituito dalla somma di due parti: quella attribuita alla sola X_2 (ossia r_{12}) e quella non spiegata dalla X_2 (ossia $(1-r^2_{12})$) moltiplicata per la proporzione spiegata dalla X_3 dopo aver rimosso l'effetto di X_2 (ossia ${r_{13.2}}^2$). Pertanto, si avrà $R^2 > {r_{12}}^2$ fintanto ${r_{13.2}}^2 > 0$, ovvero $R^2 = {r_{12}}^2$ se e solo se ${r_{13.2}}^2 = 0$.

1.3.5 L'inferenza sui coefficienti di regressione

Per l'inferenza sui parametri si assume che il vettore degli errori segua una multinormale:

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{1.61}$$

e di conseguenza anche il vettore degli stimatori dei coefficienti di regressione seguirà una multinormale:

$$\hat{\beta} \sim N(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}) \tag{1.62}$$

Nel seguito si propone la procedura inferenziale per la seguente funzione parametrica:

$$\theta = \mathbf{c}'\hat{\beta} \tag{1.63}$$

dove $\mathbf{c}$ è un vettore colonna contenente k costanti note. Ad esempio, se si definisce un vettore formato da $k-1$ zero ed un solo valore pari ad uno in corrispondenza del j-esimo elemento, allora la funzione parametrica (1.63) corrisponderà al coefficiente di regressione β_j. Si potrà inoltre definire un test per la differenza tra due coefficienti di regressione fissando nel vettore $\mathbf{c}$ una costante pari ad uno ed un'altra

pari a meno uno mentre gli altri valori risultano pari a zero. Naturalmente, scegliendo opportune costanti per il vettore $\mathbf{c}$ la procedura inferenziale potrà tener conto di diverse ipotesi teoriche relative ad opportune combinazioni dei coefficienti di regressione.

Lo stimatore BLUE di θ è dato dalla combinazione lineare degli stimatori BLUE dei coefficienti di regressione:

$$\hat{\theta} \sim N(\theta, \sigma^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}) \tag{1.64}$$

Nell'ipotesi di non conoscere la varianza dell'errore e di stimarla con la stima corretta $\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/(n-k)$, si potrà definire la statistica test

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} \tag{1.65}$$

che si distribuisce come una t-Student con $(n-k)$ gradi di libertà.

1.3.6 L'analisi della varianza: il test totale ed il test parziale

Nella regressione multipla si è interessati dapprima a verificare l'ipotesi nulla che tutti i coefficienti di regressione siano simultaneamente nulli. Il test sull'intera regressione si costruisce a partire dalla decomposizione della devianza totale in devianza di regressione e devianza dei residui. Si dimostra che considerando l'ipotesi di normalità degli errori la statistica test, nell'ipotesi nulla

$$F = \frac{SSR/(k-1)}{SSE/(n-k)} = \frac{MSR}{MSE} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{1.66}$$

si distribuisce come una F di Snedecor-Fisher con $k-1$ e $n-k$ gradi di libertà, ossia il rapporto tra due variabili casuali indipendenti e

distribuite come χ^2 rapportate ai rispettivi gradi di libertà. Pertanto, si potrà considerare il valore di significatività associato al valore empirico derivante dal rapporto (1.66).

Causa di variazione	Somma dei quadrati	Gradi di libertà	Media dei quadrati	statistica F	sign.
$X_2, \dots X_k$	SSR	$k-1$	MSR	MSR/MSE	p
Residuo	SSE	$n-k$	MSE		
Totale	SST	$n-1$			

Tabella 1.1: Analisi della varianza nella regressione: il test totale

Sulla base dell'analisi della varianza si potrà considerare una decomposizione alternativa che tenga conto di una suddivisione dei predittori in due gruppi formati rispettivamente dai primi $q-1$ predittori e dai rimanenti $k-q$ predittori. In questo caso, si è interessati a verificare l'ipotesi nulla che i coefficienti di regressione del secondo gruppo di predittori siano uguali a zero. Secondo questa ipotesi, la variabile definita come

$$F = \frac{SSR_{k-q}/(k-q)}{SSE/(n-k)} = \frac{({R_k}^2 - {R_q}^2)/(k-q)}{(1-{R_k}^2)/(n-k)} \tag{1.67}$$

si distribuisce come una F di Snedecor Fisher con $(k-q)$ e $(q-1)$ gradi di libertà, verificando il contributo del gruppo addizionale di variabili nel modello utile per la spiegazione del fenomeno.
Nel caso particolare di $q = k-1$ si considera l'effetto addizionale di una sola variabile al modello, così da valutare la significatività del relativo coefficiente di regressione.

Causa di variazione	Somma dei quadrati	Gradi di libertà	Media dei quadrati	statistica F
$X_2, \dots X_q$	SSR_q	$q-1$	MSR_q	
$X_{q+1}, \dots X_k$	$SSR_{(k-q)}$	$k-q$	$MSR_{(k-q)}$	$MSR_{(k-q)}/MSE$
$X_2, \dots X_k$	SSR	$k-1$	MSR	MSR/MSE
Residuo	SSE	$n-k$	MSE	
Totale	SST	$n-1$		

Tabella 1.2: Analisi della varianza nella regressione: il test parziale

1.3.7 Il Chow test sulla stabilità

Nel seguito si proporrà un test per verificare l'ipotesi di uguaglianza dei parametri in due regressioni indipendenti. In particolare, si considerano due campioni indipendenti di numerosità n_1 e n_2 rispettivamente estratti da popolazioni per le quali si ipotizzano due modelli classici di regressione lineare. Si vuole verificare l'ipotesi nulla che i parametri del primo modello adattato al primo campione sono uguali ai parametri del secondo modello adattato al secondo campione. Se tale ipotesi fosse rispettata, si potrebbe stimare un'unica equazione per l'insieme formato dai dati raggruppati ottenendo la somma dei quadrati SSR spiegata dalla regressione; questa non dovrebbe discostarsi troppo dalla somma delle due somme dei quadrati derivanti dalle regressioni sui due campioni indipendenti indicate con SSR_1 e SSR_2. La variabile test è definita nel seguente modo:

$$F = \frac{[SSR - (SSR_1 + SSR_2)]/k}{[SSR_1 + SSR_2]/[n_1 + n_2 - 2k]} \tag{1.68}$$

e si distribuisce come una F con k e $[n_1 + n_2 - 2k]$ gradi di libertà; se il valore empirico eccede in maniera significatica il valore critico allora non si può sostenere che le due regressioni sono uguali.

1.3.8 Le procedure di selezione delle variabili

Uno dei problemi più importanti da risolvere nella regressione multipla è la scelta di quante e quali variabili inserire nel modello. Ciò perchè occorre sempre giungere ad un compromesso tra il vantaggio di inserire quante più variabili esplicative possibili in modo da ridurre la componente erratica e lo svantaggio dovuto all'aumento dei costi e delle varianze delle stime. Ci sono varie procedure che permettono la risoluzione di questo problema:

a) la scelta a-priori delle variabili effettuata dall'analista economico aziendale in base ad assunzioni e modelli teorici (funzione di domanda o di offerta, funzione di produzione, etc.);

b) la generazione di tutte le regressioni possibili (o di un sottoinsieme ottimale) confrontate sulla base di un indice statistico (l'indice corretto di determinazione lineare $\bar{R}^2$, l'errore quandratico medio della stima, il C_p di Mallows);

c) l'applicazione di un algoritmo selettivo che iterativamente introduce variabili (regressione *forward*) o elimina variabili (regressione *backward*), ovvero introduce ed elimina variabili (regressione *stepwise*);

L'algoritmo (*backward*) consta di tre stadi:

1) regressione completa con k predittori;

2) test F parziale per valutare la significatività di ciascun predittore;

3) il predittore per il quale si ha il valore più basso del test F parziale e tale valore non è significativo viene rimosso; si ricalcola la regressione omettendo tale predittore e si ritorna al passo due.

L'algoritmo si arresta se il valore più basso del test F parziale risulta comunque significativo e pertanto non potranno essere eliminati ulteriori predittori.
L'algoritmo (*forward*) consta di quattro stadi:

1) si considera il modello senza predittori stimando solo l'intercetta;

2) si calcolano i coefficienti di correlazione semplici tra la variabile dipendente e ciascun predittore, selezionando il predittore più correlato;

3) il predittore selezionato entra nel modello se il valore empirico del test F parziale risulta significativo passando poi allo stadio quattro; altrimenti la procedura si arresta adottando il modello in corso;

4) si calcolano i coefficienti di correlazione parziale tra la variabile dipendente e ciascun predittore non ancora inserito nel modello al netto dell'effetto dei predittori già entrati nel modello, selezionando il predittore più correlato e ritornando allo stadio tre.

La regressione *stepwise* adotta un algoritmo analogo al *forward* ma rimette in discussione i predittori già inseriti in precedenza verificando la loro significatività in ogni iterazione attraverso il test F parziale. La procedura si arresta se sia il test di ingresso che il test di rimozione risultano non significativi.

1.3.9 Gli intervalli di previsione

La previsione consiste nel determinare il valore della variabile di risposta per una nuova unità sulla base delle misurazioni dei k predittori, ossia $\mathbf{x}'_* = [1, x_{2*}, \ldots, x_{k*}]$ dove con $*$ indichiamo l'$(n+1)$-esima unità. La previsione è basata sulla stima dei parametri ottenuta considerando n unità statistiche: $\hat{y}_* = \mathbf{x}'_* \beta$. L'errore di previsione sarà definito come $e_p = y_* - \hat{y}_*$. Questo sarà uno stimatore, distribuito normalmente, non distorto (con media pari a zero) e varianza data dalla seguente espressione:

$$var(e_p) = E[\hat{y}_* - y_*]^2 = var(\hat{y}_*) + var(y_*) = var(\mathbf{x}'_* \beta) + \boldsymbol{\sigma^2} \quad (1.69)$$

in quanto la covarianza tra y_* (che dipende dall'errore $\mathbf{u}$) e $\hat{y}_*$ (che dipende dallo stimatore $\hat{\beta}$) è nulla. Considerando la varianza dello stimatore del vettore dei coefficienti di regressione si ottiene:

$$var(e_p) = \sigma^2[\mathbf{x}_*{}'(\mathbf{X}'X)-1\mathbf{x}_* + 1] \quad (1.70)$$

sulla base della quale sarà possibile costruire la banda di confidenza della previsione per y_*.

1.4 Le forme funzionali del modello

1.4.1 Il modello centrato

Si ottiene il modello centrato quando le variabili sono trasformate considerando lo scarto tra valore assunto dalla variabile e valore medio della stessa. Procedendo in tal modo si ipotizza un modello di regressione con intercetta nulla in quanto questa esprime proprio la media della variabile dipendente (regressione attraverso l'origine). La stima ottenuta con il metodo dei minimi quadrati gode comunque di

proprietà ottimali. Alcuni esempi di applicazione sono rappresentati dalla teoria del portafoglio monetario (*Capital Asset Pricing Model*) che esprime il premio del titolo in proporzione alla sua volatilità, dalla teoria del reddito permanente di Friedman che esprime il consumo quale proporzione del reddito permanente, dalla teoria dei costi variabili di produzione che postula la proporzionalità del costo variabile rispetto all'output prodotto, dalla teoria monetarista per la quale il tasso di inflazione è direttamente proporzionale all'offerta di moneta.

1.4.2 Il modello con le variabili standardizzate

Il modello di regressione definito per le variabili standardizzate presenta i coefficienti di regressione definiti nel seguente modo:

$$BETA_j = \beta_j \frac{s_j}{s_y} \tag{1.71}$$

per $j = 1, \ldots, k$, dove β_j indica il corrispondente parametro del modello con variabili non standardizzate, mentre s_j e s_y sono le deviazioni standard del j-esimo predittore e della variabile dipendente rispettivamente. Tale modello consente di confrontare i valori numerici delle stime dei coefficienti di regressione in quanto essi sono espressi in unità standard, individuando in tal modo quale dei predittori ha una maggiore incidenza sulla variazione del valore atteso della variabile dipendente. Nelle applicazioni, si effettuano entrambe le regressioni con e senza la standardizzazione, in modo da arricchire l'interpretazione dei risultati.

1.4.3 Il modello log-log

L'ipotesi di linearità del modello potrebbe essere riferita sia alle variabili che ai parametri; in generale, si fa riferimento ai parametri in quanto spesso è possibile operare delle trasformazioni delle variabili

per ricondurci ad un modello lineare. Un esempio è rappresentato da una funzione:

$$w_i = \alpha {z_i}^{\gamma} \tag{1.72}$$

per la quale operando la trasformazione logaritmica delle variabili diviene lineare.

1.4.4 Il modello semilog: log-lin e lin-log

1.4.5 Il modello a trasformazione reciproca

1.4.6 La regressione polinomiale

1.5 L'uso delle variabili dummy nella regressione