CAPITOLO 6 – Misura delle relazioni tra variabili per le decisioni aziendali
Paragrafo 6.3 – Analisi e impiego della correlazione multipla e del modello di regressione lineare multipla

6

Paragrafo 6.3
Analisi e impiego della correlazione multipla e del modello di regressione lineare multipla

Argomenti

- Analisi e misura della correlazione multipla

- Modello di regressione lineare multipla

- Problemi da affrontare in presenza di correlazione tra le variabili indipendenti

L'analisi e la misura della correlazione multipla

Nella realtà aziendale come pure in contesti economici e sociali, data la natura complessa dei fenomeni, l'analisi della relazione tra variabili non si limita quasi mai allo **studio** di una singola coppia, ma piuttosto a **due o più variabili** insieme.

Al manager interessa l'analisi simultanea della relazione tra la variabile di interesse a due o più variabili che la influenzano, in modo che i risultati ottenuti siano uno strumento per prendere delle decisioni e intraprendere le azioni strategiche per modificare la variabile di interesse.

Quando l'analisi della correlazione interessa più di due variabili, l'indice di correlazione può essere calcolato per tutte le possibili coppie di variabili, ad esempio siano X, Y , e Z, avremo la correlazione di x_1 vs y; x_2 vs y; x_1 vs x_2, e il risultato ottenuto viene riportato in una matrice detta matrice di correlazione.

Alternativamente sarà possibile utilizzare l'indice di correlazione multipla: il **coefficiente di correlazione multipla** detto anche **coefficiente di determinazione R^2**, misura la proporzione di variabilità totale che viene spiegata dall'insieme di predittori.

Il coefficiente di correlazione multipla è definito come **rapporto** tra **devianza** dovuta alla **regressione** e **devianza totale**, e l'informazione che fornisce è una misura della relazione lineare che intercorre tra più di due variabili considerate insieme.

Esempio

Nel caso dell'azienda alimentare, supponiamo che il manager si ponga il seguente quesito: se prendo in esame i vari tipi di costi variabili qual è l'influenza che essi hanno sui costi totali? E quindi in sostanza la domanda è come si determinano in questo caso le inclinazioni delle rette del ricavo o dei costi totali e il punto di partenza della retta dei costi, punto che rappresenta i costi fissi?

Tabella 6.3 Volume di produzione e voci di costi variabili

Unità	Produzioni annue in centinaia di tonnellate	Salari	Contibuti sociali	Pensioni	**Totale costi personale**	Imposte	Amm.ti	Interessi	Altre spese	**Totale altri costi**	**Costi totali (costi personale + altri costi)**
Ancona	**35,58**	9,85	2,34	1,4	**13,59**	0,85	11,78	1,66	1,73	**16,02**	29,61
Aosta	**32,96**	7,11	2,2	1,4	**10,71**	0,89	13,44	1,66	1,13	**17,12**	27,83
Bari	**34,37**	7,7	2,4	1,4	**11,5**	0,94	12,43	1,66	1,86	**16,89**	28,39
Bologna	**33,37**	9,66	1,88	1,4	**12,94**	0,83	11,72	1,66	2,02	**16,22**	29,17
Cagliari	**36,18**	6,61	2,64	1,4	**10,65**	0,72	9,46	1,66	7,68	**19,52**	30,17
Campobasso	**33,77**	3,87	1,54	1,4	**6,81**	0,64	8,06	1,66	13,11	**23,47**	30,28
Catanzaro	**31,96**	4,87	1,94	1,4	**8,21**	0,67	8,72	1,66	8,8	**19,85**	28,06
Firenze	**40,6**	5,31	2,12	1,4	**8,83**	0,65	8,92	1,66	13,21	**24,45**	33,28
Genova	**38,59**	6,08	2,44	1,4	**9,92**	0,69	9,56	1,66	7,44	**19,35**	29,28
L'Aquila	**36,58**	8,03	3,22	1,4	**12,65**	0,69	11,1	1,66	3,42	**16,86**	29,51
Milano Nord	**36,78**	8,9	3,56	1,4	**13,86**	0,9	11,92	1,66	2,93	**17,42**	31,28
Milano Sud	**38,25**	5,51	2,21	1,4	**9,12**	0,69	9,7	1,66	9,89	**21,94**	31,06
Napoli	**33,97**	7,71	2,52	1,4	**11,63**	0,78	9,67	1,66	6,09	**18,2**	29,83
Palermo	**34,97**	8,14	3,78	1,4	**13,32**	0,73	9,01	1,66	3,66	**15,07**	28,39
Perugia	**32,96**	3,81	1,89	1,4	**7,1**	0,71	9,26	1,66	9,44	**21,07**	28,17
Potenza	**36,38**	5,41	2,56	1,4	**9,37**	0,81	9,97	1,66	7,46	**19,9**	29,28
Roma Nord	**38,79**	10,54	3,59	1,4	**15,53**	0,9	11,11	1,66	1,86	**15,53**	31,06
Roma Sud	**45,02**	10,99	3,72	1,4	**16,11**	0,83	12,01	1,66	4,66	**19,16**	35,27
Torino	**33,97**	6,59	3,01	1,4	**11**	0,76	10,58	1,66	3,28	**16,27**	27,28
Trento	**34,57**	5	1,78	1,4	**8,18**	0,74	9,64	1,66	8,51	**20,54**	28,72
Venezia	**33,57**	5,13	2,06	1,4	**8,59**	0,86	8,58	1,66	9,26	**20,36**	28,94

Per effettuare l'analisi suddetta viene innanzitutto calcolata la matrice di correlazione:

	Produzioni annue in centinaia di tonnellate	*Totale costi personale*	*Totale altri costi*
Produzioni annue in centinaia di tonnellate	1		
Totale costi personale	0,5 (p-value=0,017)	1	
Totale altri costi	0,13 (p-value=0,56)	-0,73 (p-value=0,000)	1

Come risulta dalla tabella i costi totali hanno una relazione negativa forte (-0,73) con il volume della produzione, e una più bassa (0,50) e positiva con il costo del lavoro; mentre vi è un legame molto scarso e non significativo con il volume della produzione.

Il modello di regressione lineare multipla

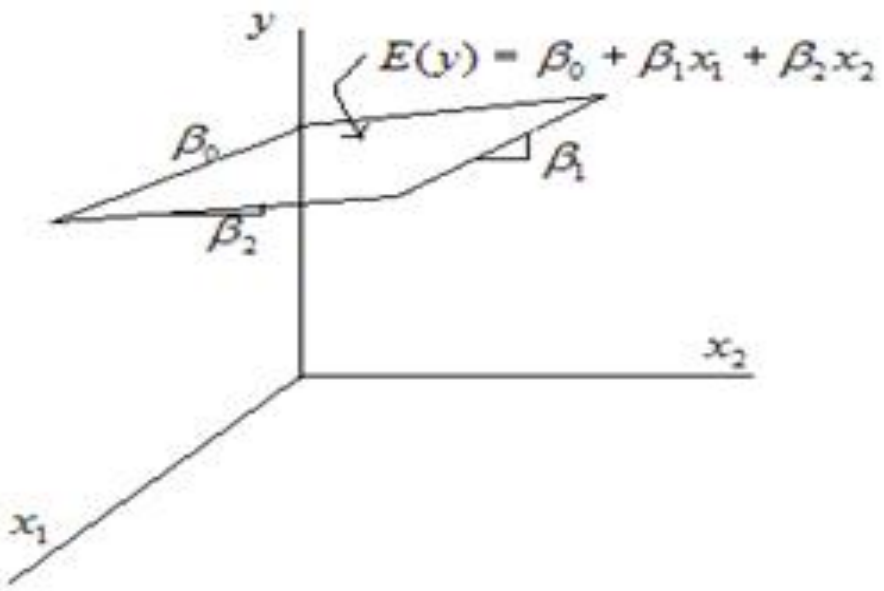
La regressione multipla è usata per **scopi predittivi**, cioè sulla base delle correlazioni tra le variabili si cerca di prevedere la risposta sulle unità statistiche; per descrivere il comportamento di un determinato fenomeno; per individuare modelli causali, cioè forme funzionali che includendo una serie di variabili spieghino la risposta osservata su un campione di soggetti, spiegando pertanto il comportamento di un determinato fenomeno oggetto di studio.

Occorrerà stimare una funzione lineare che individui un piano, se le variabili sono tre compresa la dipendente, o un iper-piano in uno spazio (k + 1) dimensionale se le variabili sono più di tre, che sia il più vicino possibile a tutti i punti dello spazio tridimensionale o k+1 dimensionale.

Il modello di regressione lineare quando esprime la dipendenza di una variabile risposta rispetto a due (o più) variabili indipendenti è detto modello di regressione multipla.

In questo paragrafo sarà affrontata l'analisi delle regressione multipla nel caso di due covariate.

Figura 6.21 Grafico di un modello regressione lineare multipla con due regressori



Vediamo le specificazioni di questo modello in generale. L'espressione della regressione è la stessa come avevamo già visto $y= f(x_1,x_2)+\varepsilon$, ma ovviamente con incluse la variabile x_1 e x_2, e con l'errore ε_i componente casuale non osservabile.
L'errore rappresenta tutti quegli elementi di cui non è possibile valutare gli effetti e che hanno, nel complesso, un'azione di disturbo.

Alcuni esempi di funzioni lineari nei parametri sono

1) $\beta_0 +\beta_1 x_1+\beta_2 x_2$

2) $\beta_0 +\beta_1 x_1+\beta_2 x_{12}$

3) $\beta_0 +\beta_1 x_1+\beta_2 x_{12} = \beta_0 +\beta_1 x_1+\beta_2 x_2$ *con* $x_2=x_{12}$

4) $\beta_0+\beta_1 \log(x_1)+\beta_2 \log(x_2)=\log[A (x_1)^{\beta_1} (x_2)^{\beta_2}]$ *e* $\beta_0=\log(A)$

Specificazione della componente deterministica

Per un campione di *n* unità statistiche e sia *i* lo i-esimo elemento del campione, la specificazione del modello è:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Le assunzioni sulla componente di disturbo sono riportate qui di seguito.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad \text{per } i=1,\dots,n$$

$$E(\varepsilon_i) = 0 \rightarrow \quad E(Y_i / x_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad \text{Per } \forall\ i=1,\dots,n$$

$$V(\varepsilon_i) = \sigma^2 \quad \rightarrow \quad V(Y_i) = \sigma^2 \quad \text{per } \forall\ i=1,\dots,n$$

$$\varepsilon_i \sim N(0,\sigma^2) \quad \rightarrow \quad Y_i \sim N((\beta_0 + \beta_1 x_i), \sigma^2))$$

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall\ i \neq j$$

Le variabili x_{1i} e x_{2i} sono valori prefissati e y_i essendo una combinazione lineare di una parte deterministica (variabili prefissate) e una componente stocastica costituita dall'errore, è essa stessa una variabile casuale (v.c.) che per effetto delle assunzioni fatte sulla componente di disturbo (errore), ha le seguenti caratteristiche:

a) il suo valore atteso è uguale alla componente deterministica
$E(y_i)= \beta_0 +\beta_1 x_{1i}+\beta_2 x_{2i}$

b) la sua varianza è omoschedastica $V(y_i))=\sigma^2$, è la σ^2:
variabilità dei valori di y intorno alla media $E(y_i)$

c) i suoi valori sono tra loro incorrelati $Cov(y_i, y_j)=0$ per la sua distribuzione è normale, in particolare $y_i \sim N(\beta_0 +\beta_1 x_{1i}+\beta_2 x_{2i} ; \sigma^2)$

Che cosa rappresentano β_1 e β_2?

Sono i coefficienti che misurano gli effetti netti vale a dire parziali delle variabili esplicative sulla variabile dipendente.

In particolare, β_1 misura la variazione di $E(y_i)$ dovuta ad una variazione unitaria di x_1 tenuta costante la variabile x_2; mentre β_2 misura la variazione di $E(y_i)$ dovuta ad una variazione unitaria di x_2 tenuta costante la variabile x_1. Il termine noto β_0 misura $E(y)$ in corrispondenza di $x_1=0$ e $x_2=0$.

Il piano dei Minimi Quadrati Ordinari (MQO) e la stima dei parametri

I coefficienti del modello di regressione multipla sono definiti seguendo il criterio del metodo dei Minimi Quadrati Ordinari (MQO) utilizzato nella regressione semplice. Pertanto, le stime dei coefficienti $\beta_0, \beta_1, \beta_2$, definite come b_0, b_1 e b_2, sono funzione dei valori osservati di y, x_1 e x_2, e tale per cui è minima la somma dei quadrati dei residui $e_i = y_i - \hat{y}_i$: min *sse (sum of squared errors)*

Dove $$sse = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left[y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2})\right]^2$$

Le condizioni del primo ordine sono tre e producono tre equazioni in tre incognite,

$$\frac{\delta sse}{\delta b_0} = 0 \qquad \frac{\delta sse}{\delta b_1} = 0 \qquad \frac{\delta sse}{\delta b_2} = 0$$

Dalla risoluzione del sistema di 3 equazioni si ottengono così le tre espressioni di b_0, b_1 e b_2

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

$$b_1 = \frac{Cov(x_1, y)Var(x_2) - Cov(x_2, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - \left[Cov(x_1, x_2)\right]^2} \qquad b_2 = \frac{Cov(x_2, y)Var(x_1) - Cov(x_1, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - \left[Cov(x_1, x_2)\right]^2}$$

Le espressioni degli stimatori dei parametri β_0 β_1 e β_2 sono definite come:

$$B_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$B_2 = \frac{Cov(X_2, Y)Var(X_1) - Cov(X_1, Y)Cov(X_1, X_2)}{Var(X_1)Var(X_2) - \left[Cov(X_1, X_2)\right]^2}$$

$$B_1 = \frac{Cov(X_1, Y)Var(X_2) - Cov(X_2, y)Cov(X_1, X_2)}{Var(X_1)Var(X_2) - Cov(X_1, X_2)^2}$$

La stima corretta della varianza σ^2 viene ricavata attraverso i residui dei MQO e precisamente

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-3}$$

La somma dei quadrati dei residui viene divisa per la grandezza [(numero di osservazioni-numero dei parametri-1)]. Quindi la grandezza

$$S^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-3}$$

Box - Proprietà degli stimatori dei Minimi Quadrati Ordinari

Gli stimatori B_0, B_1, B_2 dei parametri β_0, β_1, β_2, funzioni della v.c. Y e dell'errore ε_i , sono definiti come stimatori BLUE, cioè hanno la varianza minima nella classe degli stimatori Lineari e Corretti:

B: Best ➡ varianza minima fra i Linear Unbiased Estimators (LUE) (Teorema di Gauss- Markov)
L: Linear ➡ funzioni lineari della componenti di errore
U: Unbiased (non distorto) ➡ $E(B_0)=\beta_0$ $E(B_1)=\beta_1$ $E(B_2)=\beta_2$
E: Estimator (stimatore)

Nell'ipotesi che $\varepsilon_i \sim N(0, \sigma^2)$, le distribuzioni degli stimatori B_0, B_1, B_2 e S_e^2 sono

$$B_0 \sim N(\beta_0, \sigma^2 h_0)$$

$$B_1 \sim N(\beta_1, \sigma^2 h_1)$$

$$B_2 \sim N(\beta_2, \sigma^2 h_2)$$

$$(n-3) S_e^2 / \sigma^2 \sim \chi^2_{n-3}$$

La bontà di adattamento del modello

Nell'analisi della regressione semplice abbiamo visto come l'indice di determinazione R^2 sia una misura della bontà di adattamento del modello ai dati.

Quando si confrontano modelli di regressione lineare con un diverso numero di variabili esplicative, l'indice R^2 deve utilizzato con cautela. Accade, infatti, che l'inclusione di un'ulteriore variabile esplicativa nel modello incrementa il valore della devianza di regressione (o spiegata, ESS), anche quando la variabile inclusa non è significativa per spiegare le variazioni di Y, mentre non altera il valore della variabilità totale TSS.

Per mitigare questo effetto si introduce l'indice di determinazione multipla corretto, detto corretto, che tiene conto del numero delle variabili indipendenti presenti nel modello, includendo nell'espressione di gradi di libertà rispettivamente della devianza di regressione e della devianza totale, come segue:

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}$$

La scelta di un modello tra diversi è legata al valore di corrispondente: maggiore è l'indice, migliore sarà la bontà del modello.

La verifica dell'utilità del modello

Una volta stimato il modello, è importante verificare l'esistenza o meno di un legame lineare tra la variabile dipendente Y e le variabili indipendenti X che sono dentro al modello; in altri termini, si intende valutare la significatività dei parametri considerati *congiuntamente*, verificando l'ipotesi di indipendenza lineare

H_0: $\beta_i=0$

contro l'ipotesi alternativa

H_1: almeno un'uguaglianza in H_0 è vera.

Se l'ipotesi nulla è verificata, cioè H_0 non è rifiutata, il test conduce alla individuazione ed esclusione di quelle variabili che non spiegano la dipendenza lineare di Y. Viceversa, Se l'ipotesi nulla viene rifiutata, significa che almeno uno dei regressori contribuisce a spiegare, in termini di relazione lineare, la variabilità della variabile dipendente Y. Occorrerà pertanto fare dei test di significatività separati sui coefficienti dei singoli regressori.

Partendo dalla definizione della devianza totale come somma della devianza di regressione con la devianza residua, TSS (Total Sum of Squares)= ESS (Explained Sum of Squares) + RSS (residual Sum of Squares):

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

possiamo ottenere le varianze dividendo le devianze per i rispettivi gradi di libertà:

$$\mathrm{var}(Y)=\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1}=s_Y^2 \qquad \mathrm{var}(Y)_{reg}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{k-1}=s_{reg}^2 \qquad \mathrm{var}(Y)_e=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-k}=s_e^2$$

Se sono soddisfatte le condizioni di normalità distributiva della Y, di omoschedasticità e di indipendenza delle osservazioni, posta l'ipotesi di indipendenza lineare, le due variabili casuali campionarie s^2_{reg} e s^2_e sono entrambi stimatori corretti della varianza σ^2.

$$\frac{(k-1)s^2_{reg}}{\sigma^2} \sim \chi^2_{k-1} \qquad (n-k)S^2_e/\sigma^2 \sim \chi^2_{n-k}$$

La statistica test per la verifica di H0 si basa sul rapporto tra le varianze parziali:

Statistica test: $$= \frac{s^2_{reg}}{s^2_e} \sim F_{k-1,n-k}$$

Tale statistica, detta *F*, si distribuisce come una *F* di Fisher con *k-1* e *n-k* gradi di libertà.

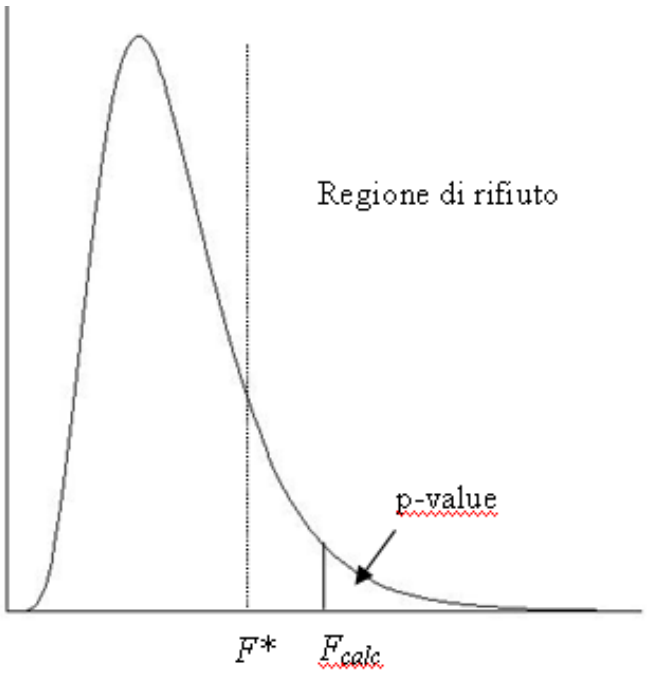
Se il valore di F^{calc} calcolato con i dati è maggiore del valore F^*, ($F^{calc}>F^*$), allora si rifiuta H_0 e si può ritenere che la variabilità spiegata dal modello sia significativamente più elevata della variabilità residua.

In sintesi:

si rifiuta H_0 se *p-value* < α

non si rifiuta H_0 se *p-value* > α, dove $p\text{-}value=P(F_{k-1,n-k}>F^{calc})$.

Figura 6.22 Distribuzione F di Fisher (un esempio)



La statistica *F* può anche essere espressa in funzione dell'indice di determinazione R^2:

$$F(k-1,n-k)=\frac{ESS/(k-1)}{RSS/(n-k)}=\frac{\frac{ESS}{TSS}\Big/(k-1)}{\frac{RSS}{TSS}\Big/(n-k)}=\frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

F è funzione monotona crescente di R^2. All'aumentare di R^2, il numeratore aumenta e il denominatore diminuisce, e quindi *F* cresce.

Si ricorda che quando nel modello di regressione è compreso un solo regressore, il valore della statistica test t sulla significatività del parametro è uguale alla radice quadrata del valore della statistica test F sulla significatività complessiva del modello: $t=\sqrt{F}$

La significatività delle stime

Si ricorre alla procedura del test di ipotesi per rispondere alla domanda se ciascuna variabile esplicativa ha singolarmente un effetto significativo o meno sulla variabile dipendente.

L'ipotesi per verificare la significatività di un singolo coefficiente è

$$H_0 : \beta_j = 0 \qquad \text{con} \qquad j = 1, \ldots, k$$

se questa ipotesi non viene rifiutata, la variabile x_j non ha un potere esplicativo e può essere eliminata dal modello.

Le statistiche test per verificare l'ipotesi nulla per ciascun coefficiente

$$\frac{|\beta_j|}{S\sqrt{h_j}} \sim t_{n-k-1}$$

In sintesi, data l'ipotesi $H_0 : \beta_j = 0$ contro l'ipotesi alternativa $H_1 : \beta_j \neq 0$ (test bidirezionale) con un livello di significatività , occorre individuare la regione di rifiuto ottenuta dai valori della statistica *t* che sono maggiori del valore teorico , dedotto dalla tavola delle probabilità della distribuzione t di Student.

Nel caso del test bidirezionale, il *p-value* è la probabilità che una variabile casuale t Student con $g=n-3$ gradi di libertà generi una realizzazione maggiore in valore assoluto di quella ottenuta $p\text{-}value=P(|t_{n-3}|>|t^{calc}|)$.

Problemi da affrontare se vi è correlazione tra le variabili indipendenti

L'esistenza di una correlazione fra le variabili indipendenti di un modello di regressione rappresenta una violazione alle assunzioni classiche del modello di regressione.

Questa situazione viene identificata con il termine di **multicollinearità** e riduce la capacità previsiva di ogni singola variabile indipendente in modo proporzionale alla forza della sua relazione con le altre variabili indipendenti.

Quali sono le conseguenze della multicollinearità sulla stima del modello di regressione?

a) Le stime dei Minimi Quadrati Ordinari dei parametri sono ancora corrette; ma le varianze e gli errori standard delle stime aumentano.

b) Eventuali cambiamenti nella specificazione del modello, mediante l'aggiunta o l'eliminazione di una o più variabili, modificano anche sensibilmente le stime dei parametri sono molto sensibili a cambi di specificazione

c) le stime possono non essere significative anche in presenza di un coefficiente di Determinazione R^2 elevato.

La multicollinearità si distingue in due tipologie:

1)la quasi multicollinearità

2)multicollinearità perfetta

La multicollinearità si valuta mediante l'impiego di alcune misure:

a)l'indice di Determinazione lineare R^2_{k0} del modello di regressione in cui X_k è la k-esima variabile che dipende dagli altri *m*-1 regressori. Se $R^2_{k0} > 0,9$, allora siamo in presenza di multicollinearità.

b) l'Indice di Tolleranza: $Tollerance = 1 - R^2_{i0}$

c) il Fattore di Accrescimento della Varianza: $Vif_i = \dfrac{1}{1 - R^2_{i0}}$

Quando VIF>5 siamo in presenza di un'alta multicollinearità.

Esistono alcuni modi per individuare il caso di multicollinearità nei dati:

1) Si può innanzitutto ragionare sulla matrice di correlazione delle variabili. Se esistono dei valori della correlazione prossimi a ± 9, la multicollinearità è sospetta.

2) Si possono stimare regressioni tra ogni covariata e le altre covariate. Se una di queste regressioni presenta un coefficiente di determinazione R2 prossimo a 1, allora la covariata considerata come variabili dipendente sarà casua di multicollinearità nella regressione originale.

3) Si può utilizzare il metodo basato sul calcolo degli autovalori della matrice dei dati. Si calcola la radice quadrata del rapporto tra gli autovalori massimo e minimo $\sqrt{\lambda_{\max} / \lambda_{\min}}$. La multicollinearità viene individuata quando il rapporto è maggiore di 20.

4) Quando le statistiche test t e il test F danno risultati contradditori

Il problema della multicollinearità può essere risolto eliminando dal modello il regressore (o i regressori) che genera la multicollinearità, oppure apportando una modifica al regressore ad esempio sostituendolo con la funzione lineare in cui compare l'altro regressore; o infine aumentando la dimensione campionaria in modo tale da ridurre l'errore standard delle stime.