Analisi di Regressione Multipla

Stima OLS della relazione *Test Score/STR* :

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \quad R^2 = .05, \; SER = 18.6$$

$$(10.4) \quad (0.52)$$

E’ una stima credibile dell’effetto causale sul rendimento nei test di un cambio del rapporto studenti-insegnanti?

No: ci sono fattori omessi che confondono l’effetto (reddito familiare; se lo studente parla l’inglese come lingua madre) questi rendono le stime OLS distorte: *STR* cattura anche l’effetto dei fattori omessi.

Analisi di Regressione Multipla

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$$

1. Stima

Similarità con la regressione semplice

- β_0 è la costante (intercetta della regressione)
- da β_1 a β_k sono tutti chiamati coefficienti angolari
- u è l'errore stocastico (o disturbo)
- Abbiamo bisogno dell'assunzione sulla media condizionata pari a zero, ovvero
- $E(u|x_1,x_2, \ldots,x_k) = 0$
- Dobbiamo sempre minimizzare la somma dei quadrati dei residui, così avremo k+1 condizioni del primo ordine

Interpretazione della regressione multipla

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$, da cui

$\Delta\hat{y} = \Delta\hat{\beta}_1 x_1 + \Delta\hat{\beta}_2 x_2 + \ldots + \Delta\hat{\beta}_k x_k$,

perciò mantenendo fisse $x_2, \ldots, x_k$ si ha che

$\Delta\hat{y} = \Delta\hat{\beta}_1 x_1$, cioè ogni β ha

una interpretazione *ceteris paribus*

Interpretazione della regressione multipla

Consideriamo il caso $k = 2$, perciò

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, quindi

$\hat{\beta}_1 = \left(\sum \hat{r}_{i1} y_i\right) / \sum \hat{r}_{i1}^2$, dove $\hat{r}_{i1}$ sono

i residui della stima della regressione

$\hat{x}_1 = \hat{\gamma}_0 + \hat{\gamma}_2 \hat{x}_2$

Interpretazione della regressione multipla

- L'equazione precedente implica che regredendo y su x_1 e x_2 otteniamo lo stesso effetto per x_1 che otterremmo regredendo y su i residui della regressione di x_1 su x_2
- Questo significa che solo la parte di x_{i1} che è non correlata con x_{i2} serve a spiegare la y_i perciò stiamo stimando l'effetto di x_1 su y dopo aver depurato l'effetto di x_2 su x_1

Regressione semplice e regressione multipla

Confrontiamo la regressione semplice $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$

con la regressione multipla $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

In genere, $\tilde{\beta}_1 \neq \hat{\beta}_1$ tranne il caso in cui:

$\hat{\beta}_2 = 0$ (perciò l'effetto parziale di x_2 *è nullo*) O

x_1 e x_2 non sono correlate nel campione di dati

Bontà del modello

Possiamo pensare ad ogni osservazione come composta da una parte spiegata, ed una non spiegata,

$y_i = \hat{y}_i + \hat{u}_i$ definiamo:

$\sum\left(y_i - \overline{y}\right)^2$ somma totale degli scostamenti dalla media al quadrato (SST)

$\sum\left(\hat{y}_i - \overline{y}\right)^2$ somma spiegata degli scostamenti dalla media al quadrato (SSE)

$\sum \hat{u}_i^2$ somma dei residui al quadrato (SSR)

Quindi SST = SSE + SSR

Bontà del modello

- Quanto bene la nostra retta di regressione si adatta ai dati?

- Possiamo calcolare la proporzione della somma totale degli scarti al quadrato (SST) che è spiegata dal modello, chiamiamo questa misura R-quadro della regressione

- $R^2 = SSE/SST = 1 – SSR/SST$

Bontà del modello

Si può pensare a R^2 come uguale al quadrato del coefficiente di correlazione tra il valore osservato y_i e il valore stimato $\hat{y}_i$

$$R^2 = \frac{\left(\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum (y_i - \bar{y})^2\right)\left(\sum (\hat{y}_i - \bar{\hat{y}})^2\right)}$$

R-quadro

- R^2 non può mai diminuire quando una nuova variabile indipendente è aggiunta alla regressione, anzi, è molto probabile che crescerà

- Poiché R^2 tende a crescere con il numero di variabili indipendenti incluse, non è una misura adeguata per confrontare diversi modelli di regressione

Assunzioni per Correttezza

- Modello di regressione della popolazione è lineare nei parametri: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

Le assunzioni sono quelle della regressione semplice più quella relativa alla non collinearità perfetta tra le variabili indipendenti:

- $E(u|x_1, x_2, \ldots x_k) = 0$, implica che tutte le variabili esplicative sono esogene ;
- Utilizziamo un campione casuale di dimensione n, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i=1, 2, \ldots, n\}$, dal modello di popolazione, in modo tale che il modello campionario è $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$;
- X e u hanno quattro momenti, cioé: $E(X^4) < +\infty$ e $E(u^4) < +\infty$.
- Nessuna delle x è costante, e <u>non</u> esiste una relazione <u>lineare perfetta</u> tra loro (collinearità)

Troppe o Poche Variabili

- Cosa succede se nel modello di regressione si inseriscono variabili non rilevanti?
- Non c’è effetto sulle stime dei parametri, e le stime OLS restano corrette
- Cosa succede se escludiamo variabili rilevanti?
- Le stime OLS saranno distorte

Errore per Variabili Rilevanti Omesse

Supponiamo che il modello vero sia

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, ma stimiamo

$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$, quindi

$$\tilde{\beta}_1 = \frac{\sum \left(x_{i1} - \overline{x}_1 \right) y_i}{\sum \left(x_{i1} - \overline{x}_1 \right)^2}$$

Errore per Variabili Rilevanti Omesse

Il modelo vero è

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, sostituendo

al numeratore si ottiene

$$\sum (x_{i1} - \overline{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i) =$$

$$\beta_1 \sum (x_{i1} - \overline{x}_1)^2 + \beta_2 \sum (x_{i1} - \overline{x}_1) x_{i2} + \sum (x_{i1} - \overline{x}_1) u_i$$

Errore per Variabili Rilevanti Omesse

$$\tilde{\beta} = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \overline{x}_1) x_{i2}}{\sum \left((x_{i1} - \overline{x}_1)^2 \right)} + \frac{\sum (x_{i1} - \overline{x}_1) u_i}{\sum \left((x_{i1} - \overline{x}_1)^2 \right)}$$

poichè $\mathrm{E}(u_i) = 0$, prendendo il valore atteso

$$E\left(\tilde{\beta}_1\right) = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \overline{x}_1) x_{i2}}{\sum \left((x_{i1} - \overline{x}_1)^2 \right)}$$

Errore per Variabili Rilevanti Omesse

Consideriamo la regressione di x_2 su x_1

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 \text{ da cui } \tilde{\delta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$$

quindi $E\left(\tilde{\beta}_1\right) = \beta_1 + \beta_2 \tilde{\delta}_1$

Sintesi sulla Direzione dell'Errore

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	errore positivo	errore negativo
$\beta_2 < 0$	errore negativo	errore positivo

Sintesi sull'Errore dovuto a Variabili rilevanti Omesse

- Due casi in cui l'errore è uguale a zero

  $\beta_2 = 0$, cioè x_2 non appartiene al modello

  e/o x_1 e x_2 non sono correlate nel campione.

- Se la correlazione tra x_2 , x_1 e x_2 , y è nella stessa direzione, l'errore sarà positivo
- Se la correlazione tra x_2 , x_1 e x_2 , y è in direzione opposto, l'errore sarà negativo

Il Caso Generale

- Tecnicamente, possiamo stabilire il segno dell'errore solo nel caso generale che tutte le variabili *x* sono non correlate

- Assumiamo che le variabili *x* non sono correlate

Varianza degli stimatori OLS

- Sappiamo che la distribuzione della nostra stima è centrata intorno al valore vero del parametro.
- Quanto dispersa è questa distribuzione?
- Misuriamo la dispersione con la varianza della distribuzione,
- Assumendo $\mathrm{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$ (Omoschedasticità)

Varianza degli stimatori OLS

- Sia $\boldsymbol{x}$ un vettore di variabili $(x_1, x_2, \ldots x_k)$
- Assumiamo che $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2$ che implica $\mathrm{Var}(y|\,\boldsymbol{x}) = \sigma^2$

- Le 4 assunzioni di correttezza, più quella di omoschedasticità sono conosciute come assunzioni di Gauss-Markov

Varianza degli stimatori OLS

Date le assunzioni Gauss-Markov

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{SST_j\left(1 - R_j^2\right)}, \text{ dove}$$

$$SST_j = \sum\left(x_{ij} - \overline{x}_j\right)^2 \text{ e } R_j^2 \text{ è } R^2$$

della regressione di x_j su tutte le altre x

Componenti della Varianza degli stimatori OLS

- La varianza degli errori: una misura grande di σ^2 implica una varianza grande degli stimatori OLS
- La varianza campionaria complessiva della variabile j: una più grande SST_j implica una varianza degli stimatori OLS più piccola
- Relazione lineare tra le variabili indipendenti: un valore più grande di R_j^2 implica una maggiore varianza degli stimatori OLS

Modelli di Regressione non Correttamente Specificati

Consideriamo un modello non correttamente specificato

$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$, tale che $Var\left(\tilde{\beta}_1\right) = \dfrac{\sigma^2}{SST_1}$

Perciò, $Var\left(\tilde{\beta}_1\right) < Var\left(\hat{\beta}_1\right)$ tranne il caso in cui

x_1 e x_2 non sono correlate,

Modelli di Regressione non Correttamente Specificati

- Mentre la varianza dello stimatore del modello non correttamente specificato è minore del modello corretto, tranne il caso in cui $\beta_2 = 0$ lo stimatore del modello non correttamente specificato è distorto
- All’aumentare della dimensione campionaria, la varianza per ogni stimatore converge a zero, e la differenza delle varianze diventa meno importante

Stima della Varianza dell' Errore

- Non conosciamo la varianza dell' errore, σ^2, perché non osserviamo l' errore, u_i

- Cosa osserviamo è il residuo, $\hat{u}_i$

- Possiamo utilizzare i residui per stimare la varianza dell' errore

Stima della Varianza dell'Errore

$$\hat{\sigma}^2 = \left(\sum \hat{u}_i^2\right) \Big/ \left(n - k - 1\right) \equiv SSR/df$$

$$\text{perciò, } se\left(\hat{\beta}_j\right) = \hat{\sigma} \Big/ \left[SST_j\left(1 - R_j^2\right)\right]^{1/2}$$

- *df* = *n* – (*k* + 1), or *df* = *n* – *k* – 1
- *df* ("*degrees of freedom*", *gradi di libertà*) pari al numero di osservazioni – numero di parametri

Il Teorema Gauss-Markov

- Date le 5 assunzioni Gauss-Markov si può dimostrare che lo stimatore OLS è "BLUE"
- *Best*
- *Linear*
- *Unbiased*
- *Estimator*
- Perciò, se le assunzioni sono valide, usa lo stimatore OLS