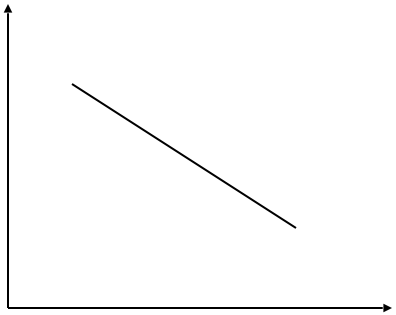
Funzioni di Regressioni Non Lineari

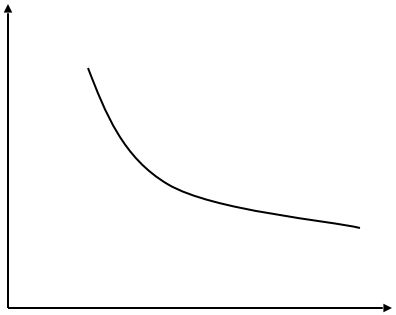
- Nelle precedenti lezioni abbiamo assunto che le funzioni di regressione della popolazione siano lineari.
  - Cioè che l'effetto su Y di una variazione unitaria di X non dipenda dal valore di X e che la pendenza della regressione sia costante.

- Quando ciò non è verificato, abbiamo il caso delle funzioni di regressione della popolazione non lineare

- Useremo due gruppi di metodi per individuare e modellare tali funzioni
- I metodi del primo gruppo sono utilizzabili nel caso in cui l'effetto su Y della variazione di X_1 dipende dal valore di X_1.

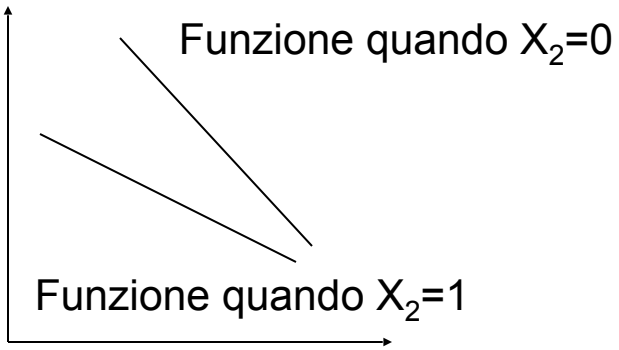


Pendenza costante



Pendenza funzione del valore di X_1

- I metodi del secondo gruppo sono utilizzabili quando l'effetto su Y di una variazione in X_1 dipende dal valore di un'altra variabile indipendente, X_2.



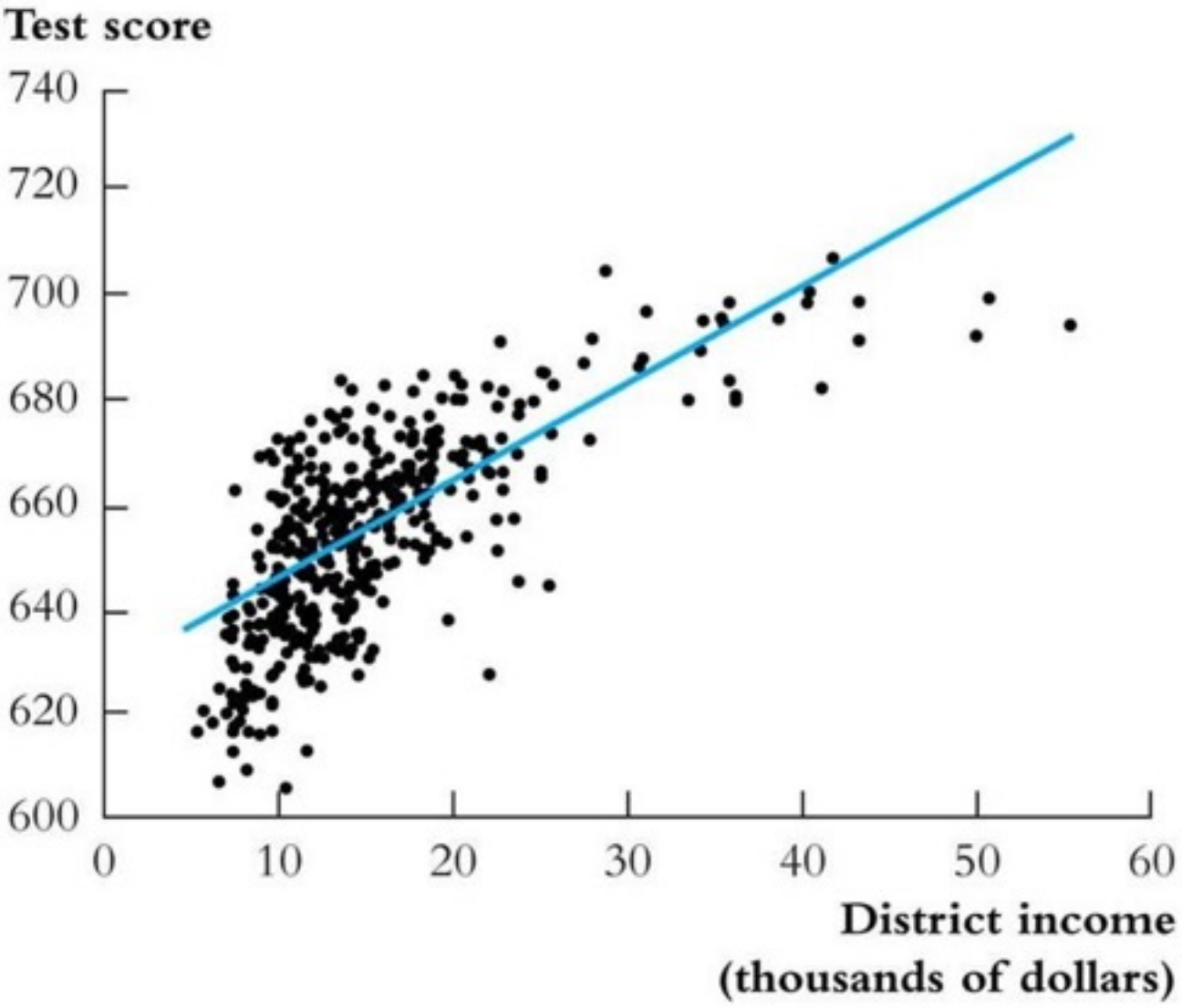
Pendenza Funzione dipende da X_2

- Anche se non lineari nelle X, questi modelli sono funzioni lineari dei coefficienti ignoti del modello di regressione e sono delle varianti del modello di regressione multipla già visto.

- Allo stesso modo, i coefficienti ignoti di tali funzioni non lineari possono essere stimati e sottoposti a verifica usando gli OLS.

Un Approccio Generale alle funzioni di regressione non lineari

- Nella regressione multipla abbiamo notato come il reddito sia un fattore rilevante per spiegare i punteggi dei test.
- In particolare possiamo prendere in considerazione il reddito medio annuo pro capite nel distretto (reddito distretto).
- Come già visto faremo riferimento al distretto della California nel 1998.


Test score
740
720
700
680
660
640
620
600
0
10
20
30
40
50
60
District income
(thousands of dollars)

- La figura mostra il grafico a nuvola dei punteggi e del reddito medio del distretto. Le due variabili sono fortemente e positivamente correlate (0.71).
- Tuttavia, i punti sono disposti in maniera particolare: al di sotto della retta per redditi<10000$ o >40000$; al di sopra della retta quando il reddito è compreso tra 15000$ e 30000$.
- La relazione quindi non appare quella di una retta, bensì di una funzione quadratica che presenta una curvatura in grado di avvicinarsi maggiormente ai punti del grafico.

- Il modello di regressione quadratico è il seguente

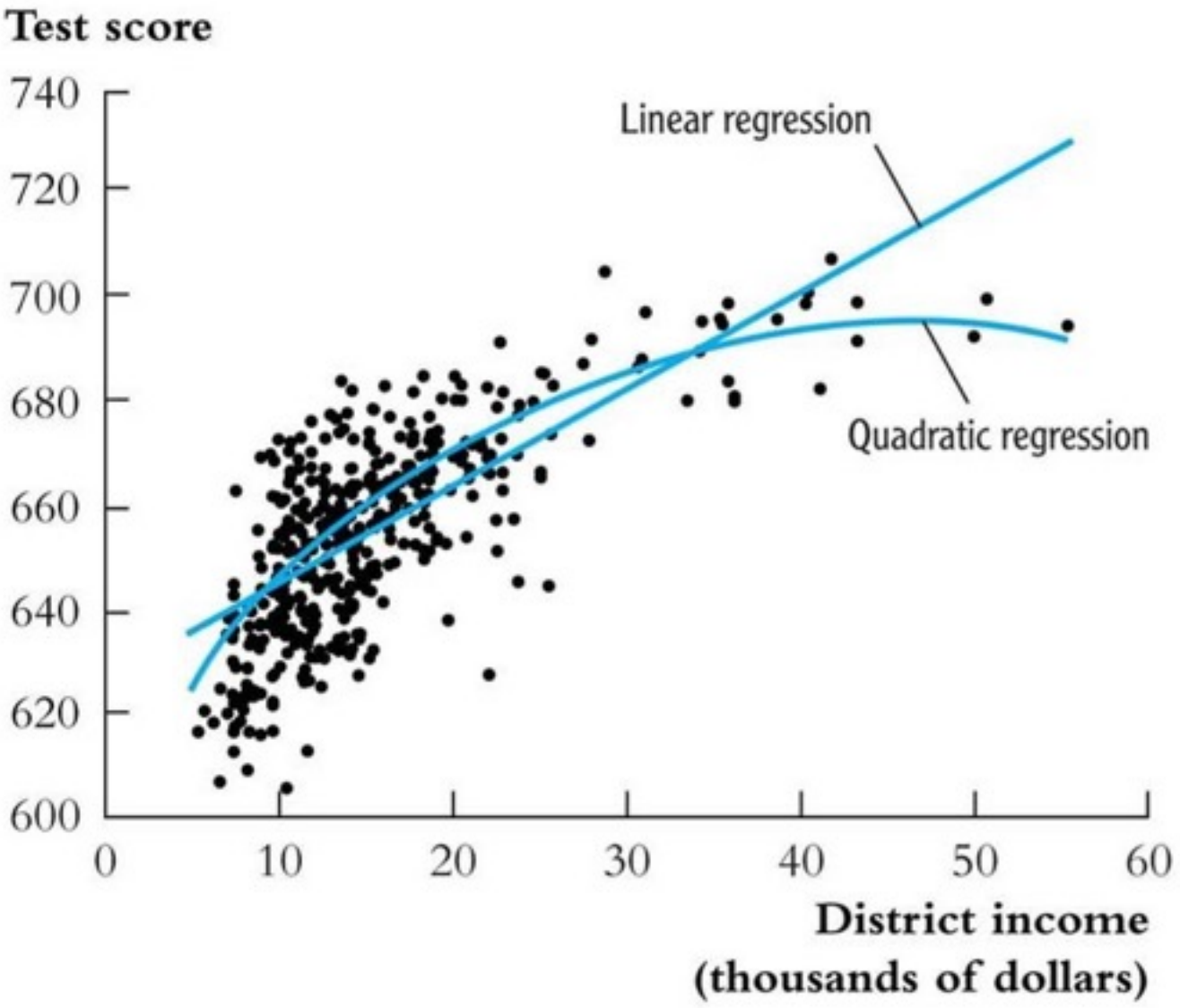
$$Testscore = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i$$

- Tale modello è una variante della regressione multipla. Pertanto possiamo usare gli OLS per stimare i coefficienti della regressione,

$$\hat{TestScore} = \underset{(2.9)}{607.3} + \underset{0.27}{3.85}\, Income - \underset{(0.0048)}{0.0423}\, Income^2, \overline{R}^2 = 0.554$$

- Possiamo verificare l'ipotesi che la relazione tra le variabili sia lineare contro l'alternativa che sia non lineare. Basterebbe testare l'ipotesi nulla H_0: β_2=0 contro l'alternativa bilaterale.

- Nel nostro caso t=(-0.0423/0.0048)=-8.81. Quindi rifiutiamo l'ipotesi nulla


Test score
740
720
700
680
660
640
620
600
0
10
20
30
40
50
60
Linear regression
Quadratic regression
District income
(thousands of dollars)

L’effetto su Y di una variazione di X nelle funzioni non lineari

- La variazione attesa in Y, ΔY, associata alla variazione ΔX_1 in X_1, tenendo costanti gli altri regressori, è la differenza tra il valore della funzione di regressione della popolazione prima e dopo la variazione di X_1, tenendo costanti i regressori.

- $\Delta Y = f(X_1+\Delta X_1, X_2,\ldots,X_n) - f(X_1,X_2,\ldots,X_n)$

- Lo stimatore di tale differenza ignota è la differenza tra i valori predetti nei due casi.
- Sia $\hat{f}(X_1, X_2, ..., X_n)$ il valore predetto di Y basato sullo stimatore della funzione di regressione della popolazione. Allora la variazione predetta è:

$$\Delta\hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, ..., X_n) - \hat{f}(X_1, X_2, ..., X_n)$$

- Tornando al nostro esempio consideriamo una variazione del reddito del distretto e la previsione della regressione relativamente al punteggio.
- Ipotizziamo un aumento del reddito da 10000$ ad 11000$. La variazione nel punteggio sarà

$$\Delta\hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2) =$$
$$= (607.3 + 3.85 \times 11 - 0.0423 \times 121) - (607.3 + 3.85 \times 10 - 0.0423 \times 100) =$$
$$= 644.53 - 641.57 = 2.96$$

Ricapitolando:

1. Identificare una possibile relazione non lineare (teoria economica, esperienza)
2. Specificare una funzione non lineare e stimarne i parametri tramite gli OLS
3. Capire se la forma funzionale scelta è un miglioramento rispetto quella lineare
4. Disegnare la funzione stimata
5. Stimare l'effetto della variazione di Y

- Attenzione all'interpretazione dei coefficienti nelle specificazioni non lineari.

Funzioni non lineari di una singola variabile indipendente

- Abbiamo due metodi
  - Polinomiale
  - Logaritmico

- Il metodo polinomiale suggerisce il seguente modello: $Y_i=\beta_0+\beta_1X_i+\beta_2X^2_i+\ldots+\beta_rX^r_i+u_i$

- Quando ad esempio r=3 abbiamo un modello di regressione cubica.

- Il modello di regressione polinomiale è simile al modello di regressione multipla con la sola differenza che in questo caso sono potenze della stessa variabile indipendente.
- Quindi le tecniche di stima e verifica di ipotesi sono le stesse già viste per la regressione multipla.
- Come scegliere il grado del polinomio?
  - Aumentare il grado garantisce una maggiore flessibilità nella funzione per catturare forme più varie.
  - Tuttavia, aumentare i regressori può ridurre la precisione delle stime.

Si utilizza il metodo di verifica d'ipotesi sequenziale

1. Scegliere un valore massimo per il grado r del polinomio e stimare la regressione
2. Utilizzare la statistica t per verificare l'ipotesi che il coefficiente di X^r sia nullo. Se rifiutiamo tale variabile entra nella regressione e ci fermiamo.
3. Se non si rifiuta, si elimina tale regressore e si stima una regressione polinomiale di grado r-1. Poi si testa la stessa ipotesi nulla sul coefficiente di X^{r-1} etc…

Problema non sappiamo il grado del polinomio con cui iniziare. Meglio iniziare con un polinomio di piccolo ordine (non superiore a 4).

$$\underset{}{\widehat{TestScore}} = \underset{(5.1)}{600.1} + \underset{0.71}{5.02}\,Income - \underset{(0.029)}{0.096}\,Income^2 + \underset{(0.00035)}{0.00069}\,Income^3$$

Logaritmi

Ci sono tre casi diversi in cui si possono impiegare i logaritmi

1. $Y_i=\beta_0+\beta_1 \ln(X_i)+u_i$ in cui una variazione percentuale dell'1% in X determina una variazione pari a $0.01\beta_1$ in Y
2. $\ln(Y_i)=\beta_0+\beta_1 X_i+u_i$ in cui una variazione di un'unità in X determina una variazione pari al $100\beta_1$% in Y
3. $\ln(Y_i)=\beta_0+ \beta_1 \ln(X_i)+u_i$ in cui una variazione pari all'1% in X determina una variazione pari a β_1% in Y, quindi β_1 è l'elasticità di Y rispetto a X

Il modello lineare-logaritmico

- $Y_i=\beta_0+\beta_1 \ln(X_i)+u_i$. Ad esempio abbiamo

$$\underset{}{\widehat{TestScore}} = \underset{(3.8)}{557.8} + \underset{(1.40)}{36.42}\ln(Income), \overline{R}^2 = 0.561$$

- Un incremento di reddito dell'1% è associato ad un aumento medio del punteggio pari a 0.01x36.42=0.36 punti.
- Per stimare l'effetto atteso su Y di una variazione di X nell'unità di misura originaria (migliaia di dollari) si usa il metodo precedentemente visto.

Il modello log-lineare

- $Ln(Y_i)=\beta_0+\beta_1 X_i+u_i$. Come esempio riconsideriamo la relazione tra età e retribuzione dei laureati.

$$\ln(\hat{Earnings}) = \underset{(0.024)}{2.453} + \underset{(0.0006)}{0.0128}\, Age, \bar{R}^2 = 0.0387$$

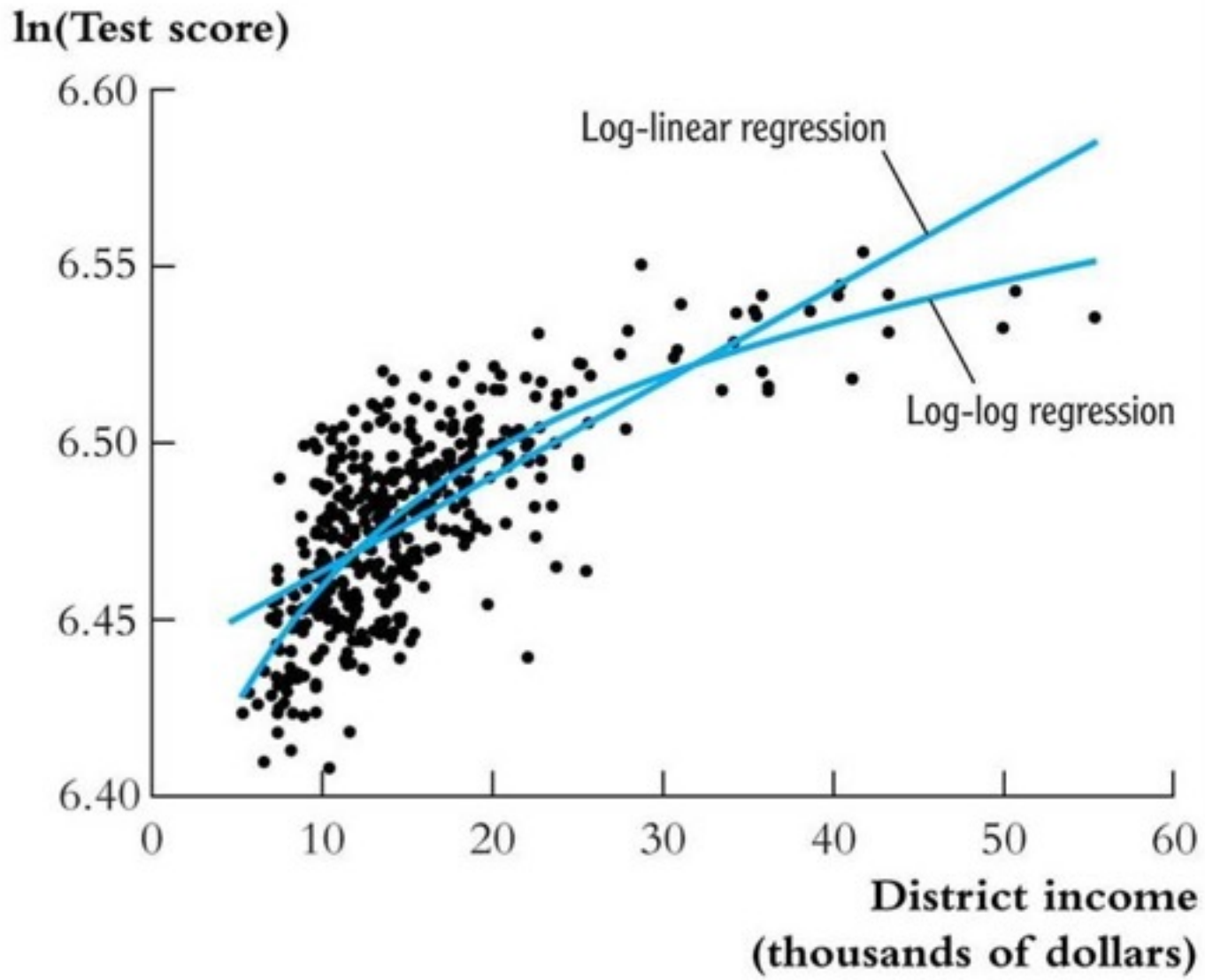
- Secondo questa regressione ci si attende che le retribuzioni crescano dell'1.28% per ogni anno d'età in più.

Il modello log-log

- $Ln(Y_i)=\beta_0+ \beta_1 ln(X_i)+u_i$. In questo modello β_1 è l'elasticità di Y rispetto a X, cioè è la variazione percentuale di Y associata ad una variazione di X dell'1%.
- Ad esempio

$$\ln(\hat{TestScore}) = \underset{(0.006)}{6.336} + \underset{(0.0021)}{0.0554}\ln(Income), \overline{R}^2 = 0.557$$

- Un incremento del reddito dell'1% corrisponde ad un incremento medio dello 0.0554% nei punteggi


ln(Test score)
6.60
6.55
6.50
6.45
6.40
Log-linear regression
Log-log regression
0
10
20
30
40
50
60
District income
(thousands of dollars)

- Come mostrato in figura la specificazione log-log approssima meglio i dati rispetto la specificazione log-lineare. Ciò è anche confermato da un maggiore R corretto, anche se la log-log presenta pure delle imprecisione nell'adattarsi ai dati.
- Come facciamo a scegliere tra i modelli di regressione logaritmica? Si deve usare l'R^2corretto.
- Tuttavia bisogna ricordare che non è possibile fare tale confronto tra modello lineare-logaritmico e quello log-log in quanto le variabili dipendenti sono diverse (Y_i e $\ln(Y_i)$).

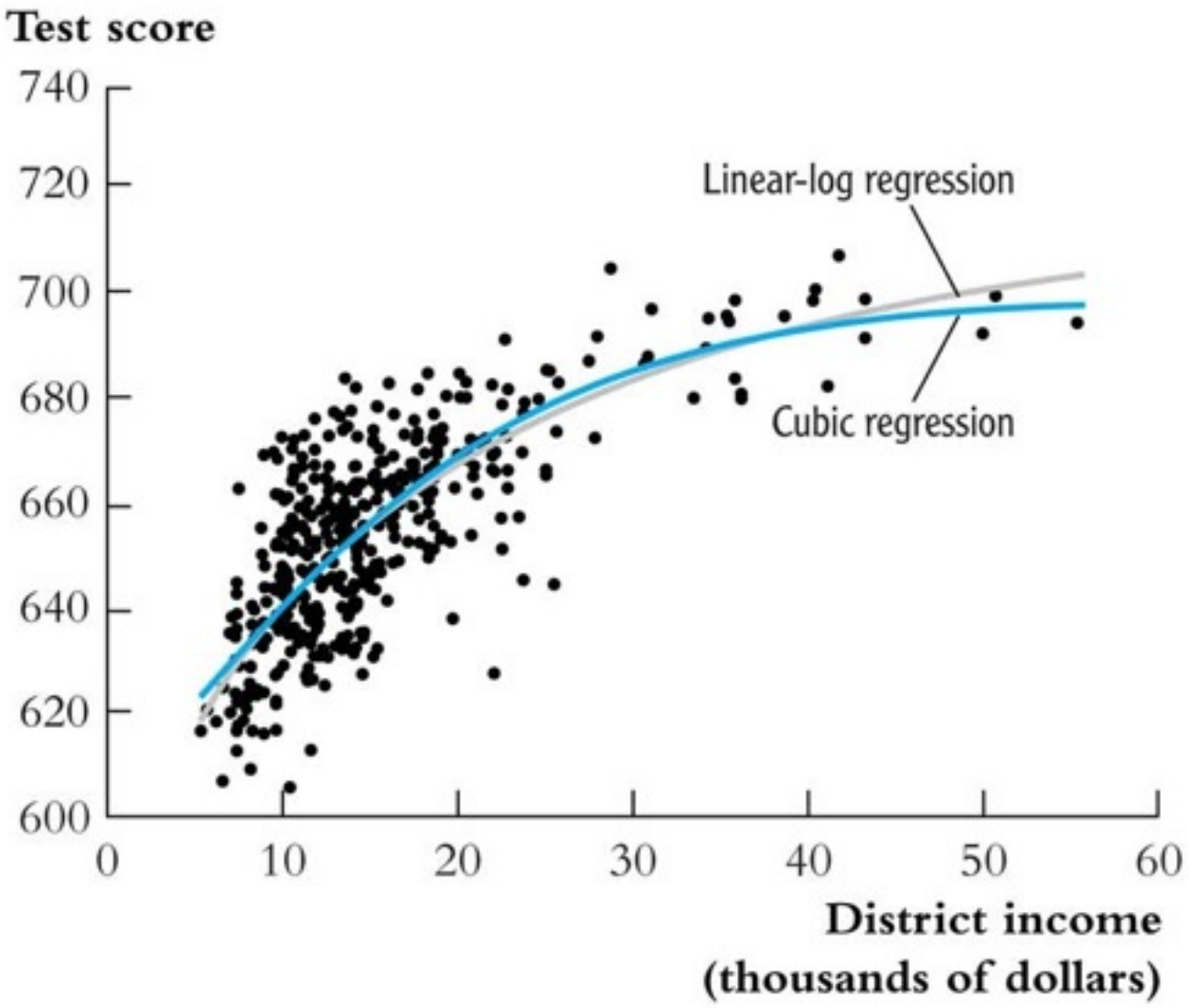
Confronto modello polinomiale e logaritmico

- Relativamente al nostro esempio abbiamo concluso che la forma polinomiale preferita è quella cubica.
- Al tempo stesso la specificazione logaritmica lineare-logaritmica sembrava fornire una buona interpolazione.
- Per verificarlo possiamo inserire in tale modello delle potenze del logaritmo del reddito. Se tali termini non sono statisticamente diversi da zero allora la specificazione iniziale è quella corretta. Quindi stimiamo,

$$\hat{TestScore} = \underset{(79.4)}{486.1} + \underset{(87.9)}{113.4}\ln(Income) - \underset{(31.7)}{26.9}\left[\ln(Income)\right]^2$$

$$+ \underset{(3.74)}{3.06}\left[\ln(Income)\right]^3, \overline{R}^2 = 0.560$$

- La statistica t del coefficiente del termine cubico è 0.818 e quindi non significativa non è rifiutata al 10%(1.64).
- Lo stesso risultato si ottiene considerando la statistica F per l'ipotesi congiunta che i veri coefficienti del termine cubico e quadratico siano nulli.
- Il modello cubico logaritmico non fornisce alcun miglioramento rispetto al modello lineare logaritmico.


Test score
740
720
700
680
660
640
620
600
0
10
20
30
40
50
60
Linear-log regression
Cubic regression
District income
(thousands of dollars)

Interazione tra variabili indipendenti

- Consideriamo il caso di due variabili binarie, relative al genere ed all'istruzione dei soggetti, come regressori del logaritmo delle retribuzioni,

$$Y_i=\beta_0+\beta_1 D_{1i}+\beta_2 D_{2i}+u_i$$

- Questa specificazione non è in grado di misurare il diverso effetto della laurea sui generi. Tale relazione può essere colta da un nuovo regressore ottenuto come prodotto delle due variabili binarie

$$Y_i=\beta_0+\beta_1 D_{1i}+\beta_2 D_{2i}+\beta_3 (D_{1i} x D_{2i})+u_i$$

- Il nuovo regressore è detto termine d'interazione

Esempio

- Consideriamo la seguente regressione multipla

$$\widehat{TestScore} = \underset{(1.4)}{664.1} - \underset{(2.3)}{18.2}\,HiEL - \underset{(1.9)}{1.9}\,HiSTR - \underset{(3.1)}{3.5}(HiSTR \times HiEL), \overline{R}^2 = 0.290$$

Dove HiSTR è una variabile binaria che è uguale ad 1 se il STR è 20; HiEL è una variabile binaria che è uguale ad 1 se la percentuale di studenti che ancora apprendono l'inglese è almeno 10%.

- L'effetto predetto del passaggio da un distretto con basso STR ad uno con alto STR tenendo costante la percentuale di studenti che ancora apprendono l'inglese è dato da

$$\beta_2 + \beta_3 d_1 = -1.9 - 3.5\,HiEL$$

- Se HiEL è bassa, l'effetto sarà una diminuzione di 1.9 punti, se HiEL è alta avremo una diminuzione di 5.4

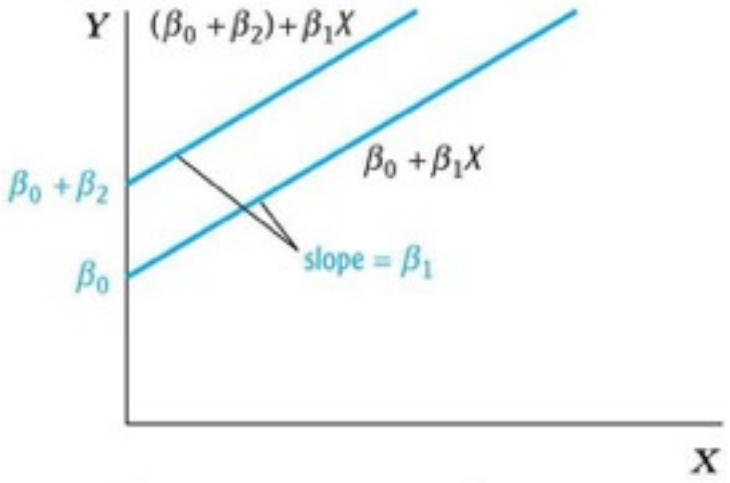
Interazione tra variabile continua e binaria

- Consideriamo la regressione del logaritmo delle retribuzioni (Y_i) come variabile continua e gli anni d'esperienza lavorativa e la presenza o meno di un lavoratore laureato come variabile binaria.
- Vi sono tre modi per mettere il relazione una variabile binaria ed una continua.

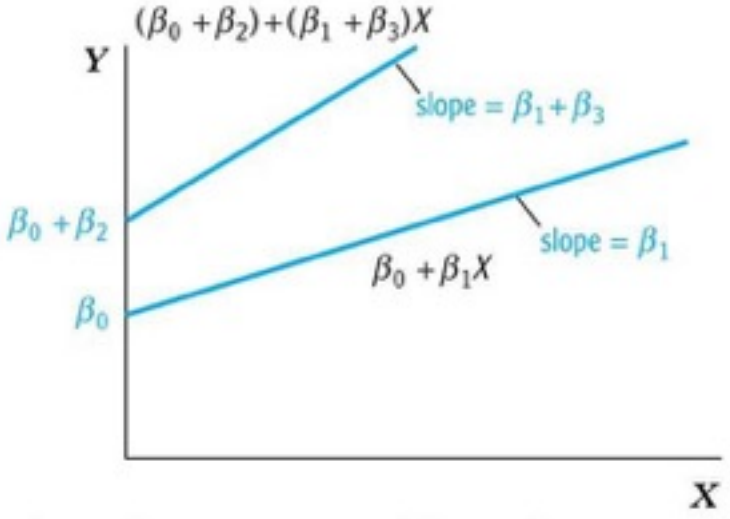
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

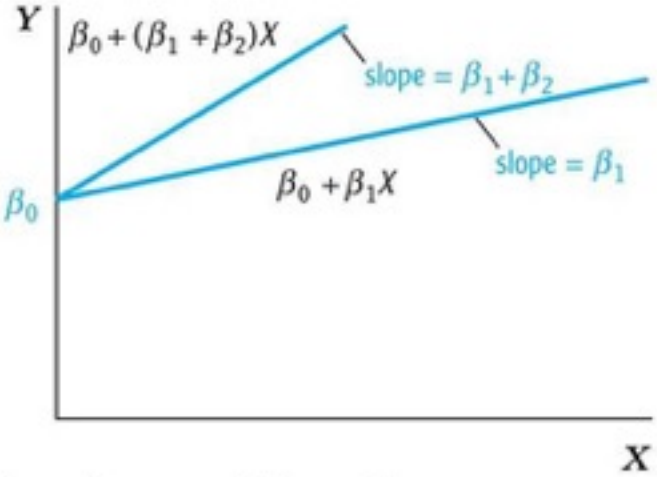
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$$



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

- Nel primo caso, le rette differiscono solo nell'intercetta.

- Nel secondo caso, le rette hanno pendenze ed intercette diverse. Le pendenze diverse fanno sì che l'effetto di un anno addizionale di lavoro differisca tra laureati e non laureati.

- Nel terzo caso le due rette hanno stessa intercetta ma diversa pendenza.

Esempio

- Tornando all'esempio precedente, avremo che

$$\ln(\widehat{TestScore}) = \underset{(11.9)}{682.2} - \underset{(0.59)}{0.97}\,STR + \underset{(19.5)}{5.6}\,HiEL - \underset{(0.97)}{1.28}(STR \times HiEL), \overline{R}^2 = 0.305$$

- A seconda del livello di HiEL ridurre STR farebbe aumentare i punteggi del test di 0.97 punti nei distretti con bassa HiEL e di 2.25 nei distretti con alta HiEL.
- La loro differenza è 1.28 che è anche il coefficiente del termine d'interazione.

Interazioni tra due variabili continue

- Un esempio è il caso in cui Y_i è il logaritmo della retribuzione, X_{1i} è il numero di anni d'esperienza lavorativa e X_{2i} è il numero di anni di frequenza scolastica. Il modello diventa,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

- Il termine d'interazione permette all'effetto di una variazione unitaria di X_1 di dipendere da X_2. In questo caso, l'effetto su Y di una variazione di X_1 tenendo costante X_2 è

$$\frac{\Delta Y}{\Delta X} = \beta_1 + \beta_3 X_2$$

- Lo stesso si può dire partendo da una variazione di X_2 tenendo costante X_1.
- Mettendo insieme i due effetti si può notare che il coefficiente β_3 è l'effetto aggiuntivo di un incremento unitario in X_1 e X_2, che si somma all'effetto individuale di un incremento in X_1 da sola e in X_2 da sola.
- Nel solito esempio abbiamo,

$$\ln(\widehat{TestScore}) = \underset{(11.8)}{686.3} - \underset{(0.59)}{1.12}\,STR - \underset{(0.37)}{0.67}\,PctEL + \underset{(0.019)}{0.0012}(STR \times PctEL),$$

$$\overline{R}^2 = 0.422$$

Ricapitolazione dell'esempio

- Domande:
  - L'effetto sui punteggi del test della riduzione nel rapporto studenti-insegnanti dipende dalla frazione di studenti che ancora apprendono l'inglese?
  - Tale effetto dipende dal valore del rapporto studenti-insegnanti?
  - Prendendo in considerazione i fattori economici e la non linearità, qual è l'effetto stimato sui punteggi del test di una riduzione del rapporto studenti-insegnanti di due studenti per insegnante così come proposto dal provveditore?

TABLE 8.3 Nonlinear Regression Models of Test Scores

Dependent variable: average test score in district; 420 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (STR)	−1.00** (0.27)	−0.73** (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33** (24.86)	83.70** (28.50)	65.29** (25.26)
STR^2					−3.42** (1.25)	−4.38** (1.44)	−3.47** (1.27)
STR^3					0.059** (0.021)	0.075** (0.024)	0.060** (0.021)
% English learners	−0.122** (0.033)	−0.176** (0.034)					−0.166** (0.034)
% English learners ≥ 10%? (Binary, $HiEL$)			5.64 (19.51)	5.50 (9.80)	−5.47** (1.03)	816.1* (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3* (50.2)	
$HiEL \times STR^2$						6.12* (2.54)	
$HiEL \times STR^3$						−0.101* (0.043)	
% Eligible for subsidized lunch	−0.547** (0.024)	−0.398** (0.033)		−0.411** (0.029)	−0.420** (0.029)	−0.418** (0.029)	−0.402** (0.033)
Average district income (logarithm)		11.57** (1.81)		12.12** (1.80)	11.75** (1.78)	11.80** (1.78)	11.51** (1.81)
Intercept	700.2** (5.6)	658.6** (8.6)	682.2** (11.9)	653.6** (9.9)	252.0 (163.6)	122.3 (185.5)	244.8 (165.7)

F-Statistics and *p*-Values on Joint Hypotheses

(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR$, $HiEL \times STR^2$, $HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\overline{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

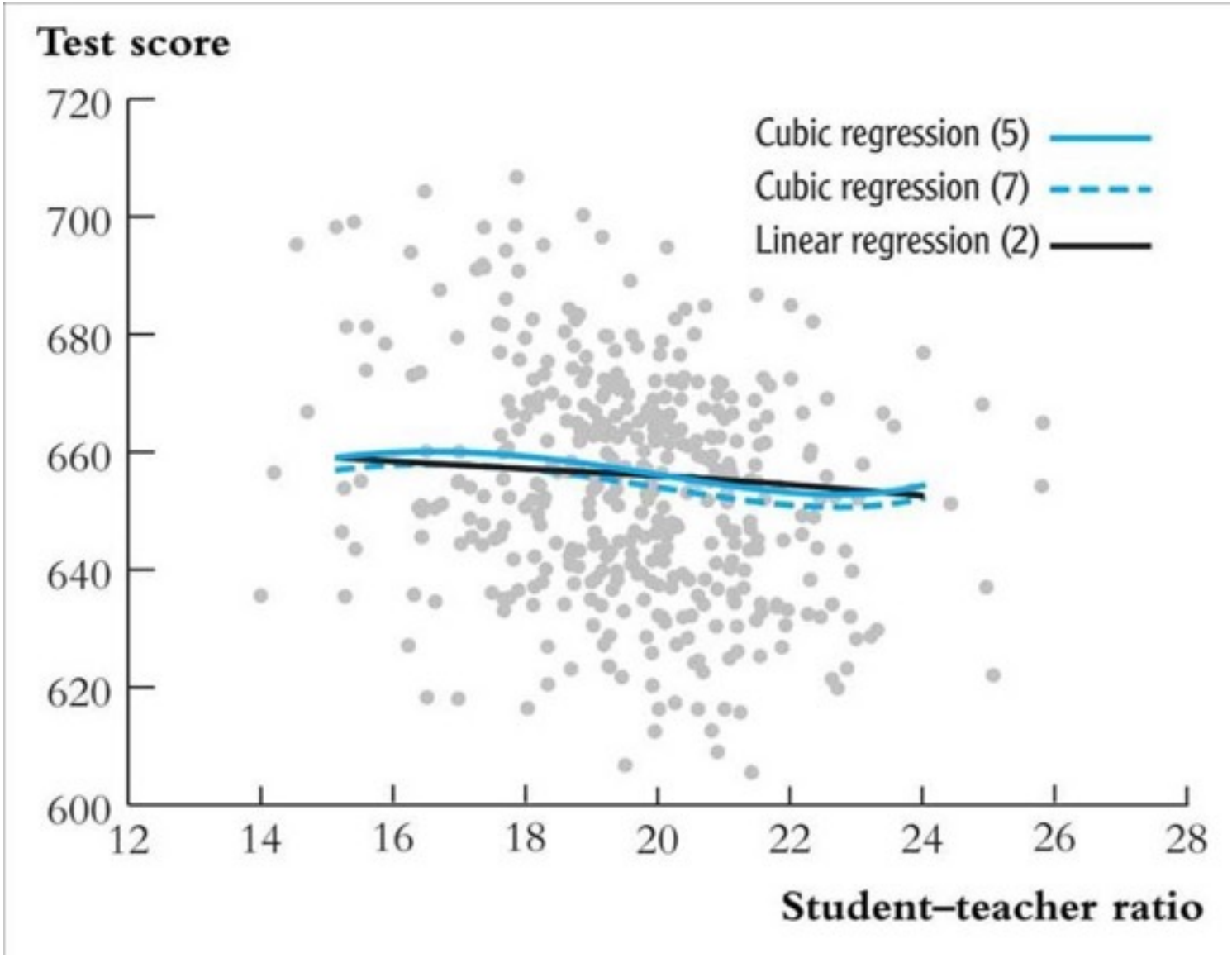
- La regressione (1) è la specificazione di base già vista precedentemente.
- La regressione (2) controlla per il reddito.
  - Tale variabile risulta significativa e la variazione del coefficiente di STR giustifica la presenza del reddito
- La regressione (3) introduce la variabile binaria che controlla per la percentuale di studenti che ancora apprendono l'inglese. Mancano ancora le variabili di controllo di tipo economico.
- La regressione (4) introduce tali variabili. I coefficienti cambiano ma i coefficienti del termine d'interazione non è mai significativo al 5%.
  - L'ipotesi che percentuali alte o basse di studenti che apprendono l'inglese non influiscano sul STR non può essere rifiutata al 5%.

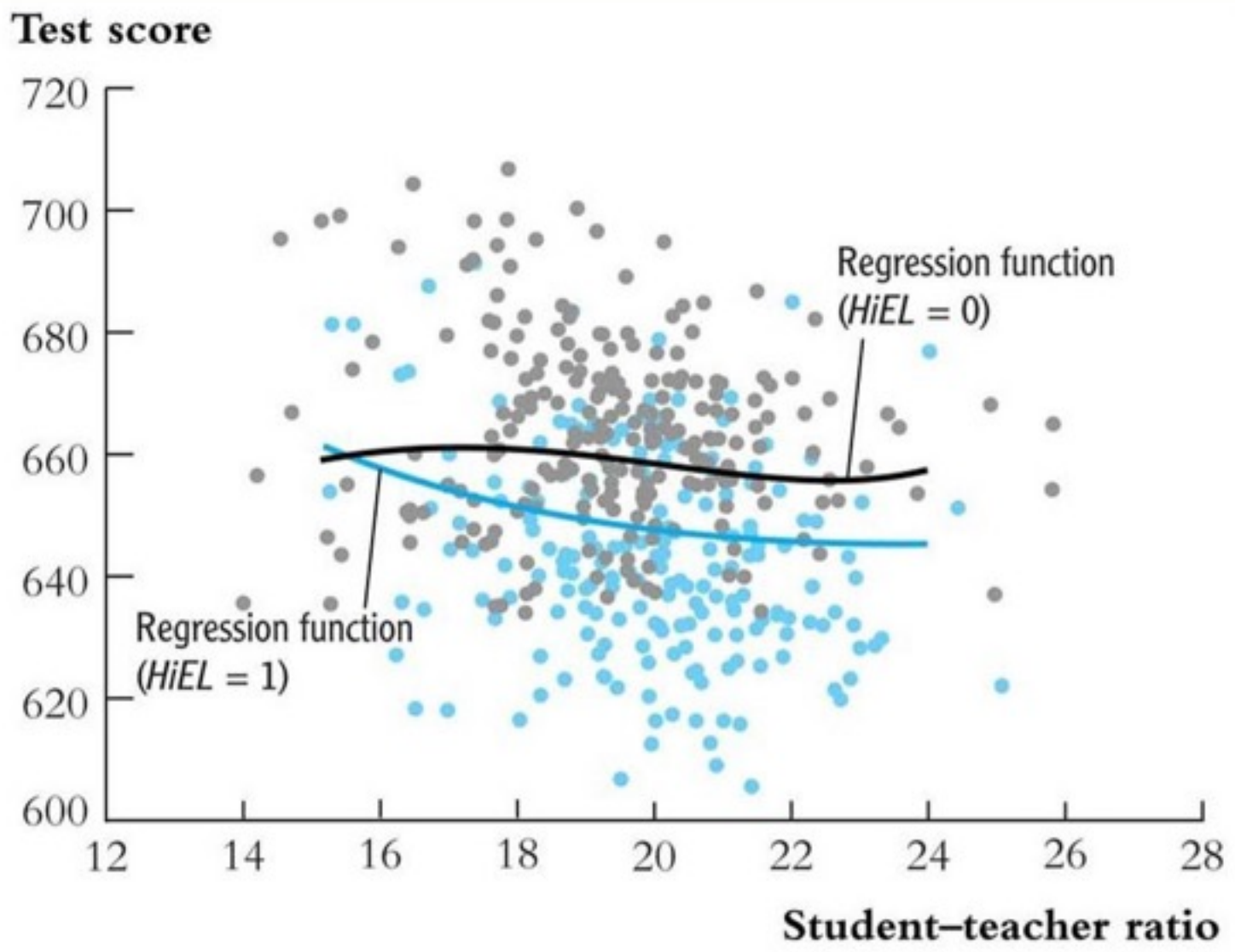
- La regressione (5) controlla per l'effetto di una specificazione cubica in STR oltre alle variabili di controllo presenti in (4). La (5) mostra che la relazione punteggio/studenti-insegnanti non è lineare.
- La regressione (6) controlla anche per l'ulteriore effetto della frazione di studenti che ancora apprendono l'inglese. Includendo anche i due termini d'interazione verifichiamo se le funzioni di regressione della popolazione che mettono in relazione i punteggi del test e STR sono diverse per percentuali alte o basse di studenti che ancora apprendono l'inglese.

- La statistica F (2.69, valore-p=0.046) fornisce evidenza che le funzioni di regressione sono diverse per distretti con percentuali diverse di studenti che devono ancora apprendere l'inglese.
- Le differenze sono dovute ai termini quadratici e cubici.
- La regressione (7) è una variante della (5) dove al posto di HiEL abbiamo la variabile continua PctEL. Non vi sono variazioni importanti dei coefficienti e quindi la (5) non sembra essere sensibile al tipo di misurazione della percentuale di studenti che devono apprendere l'inglese.

- Le regressioni cubiche (5, 7) sembrano appiattirsi verso quei valori elevati del rapporto studenti-insegnanti.

- La nonlinearità tuttavia rimane poco evidente.


Test score
720
700
680
660
640
620
600
12
14
16
18
20
22
24
26
28
Student–teacher ratio
Cubic regression (5)
Cubic regression (7)
Linear regression (2)


Test score
720
700
680
660
640
620
600
12
14
16
18
20
22
24
26
28
Student–teacher ratio
Regression function
(HiEL = 0)
Regression function
(HiEL = 1)

- La regressione (6) mostra una chiara differenza tra le regressioni cubiche che mettono in relazione punteggi del test e STR , a seconda che la percentuale di studenti non di madrelingua inglese sia grande o piccola.
- Nell'intervallo STR 17-23 (88% delle osservazioni) le due funzioni sono notevolmente distanziate.
- I distretti con una più bassa percentuale di studenti non di madrelingua inglese fanno meglio, tenendo costante STR, ma l'effetto di una variazione di STR è quasi lo stesso tra i due gruppi.
- Bisogna essere molto cauti nell'interpretare la differenza tra le regressioni quando STR è basso in quanto tali valori fanno riferimento a poche osservazioni.

Risposte ai 3 Quesiti

1. Dopo aver controllato per la condizione economica, il fatto che il distretto abbia molti o pochi studenti non di madrelingua inglese non ha un'influenza significativa sul variare dei punteggi al variare di STR.

2. Dopo aver controllato per la condizione economica, c'è evidenza di un effetto non lineare di STR sul punteggio del test

3. Vediamo l'effetto di una riduzione di STR di due unità sul punteggio del test
    1. (2) -0.73*-2=1.46 non dipende da STR
    2. Nelle specificazioni non lineari questo effetto dipende dal valore di STR. In (5) passando da 20 a 18 avremmo un miglioramento di 3 punti
    3. In (7) avremmo 2.93 punti

Passando invece da 22 a 20, avremmo con la (5) un miglioramento di 1.93 e con la (7) di 1.90

Le specificazioni non lineari suggeriscono che tagliare STR ha un effetto maggiore se tale rapporto è già piccolo.