Econometria | 2022/2023

Lezione 9: Modelli non-lineari

Giuseppe Ragusa
https://gragusa.org

Roma, marzo 2023

Sommario

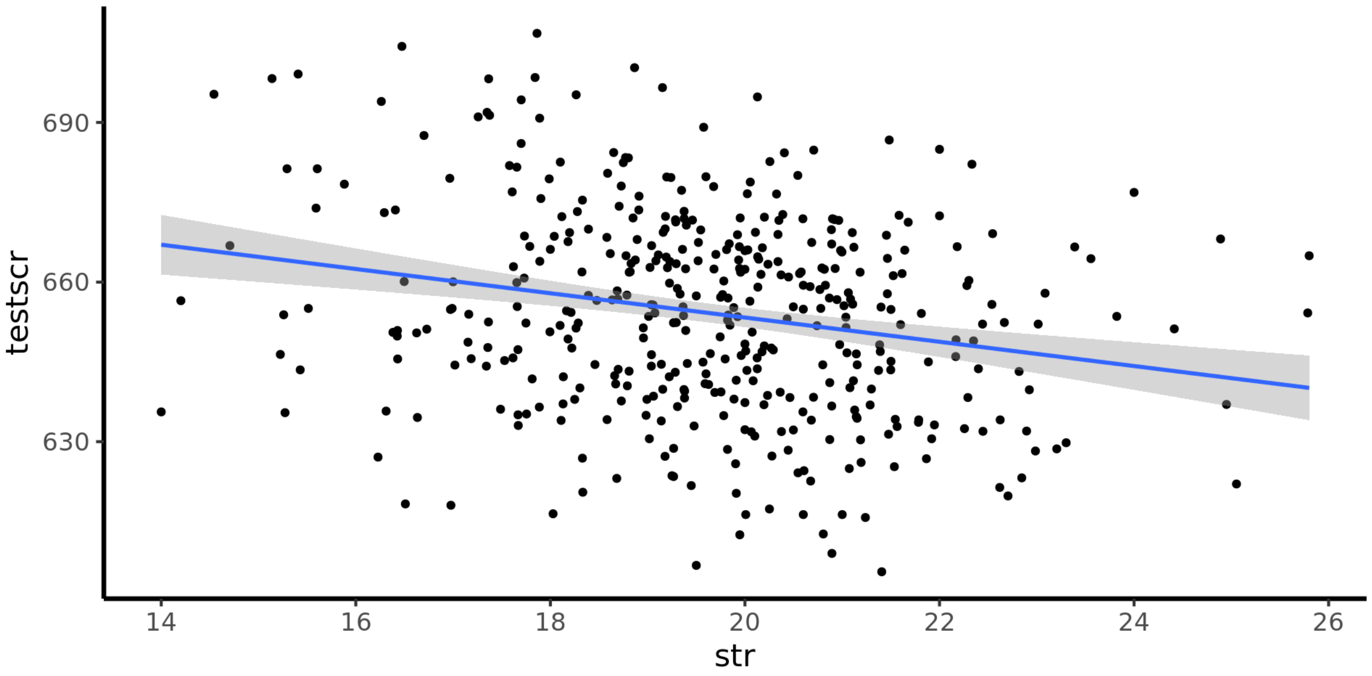
- Funzioni di regressione non lineari – note generali
- Funzioni non lineari a una variabile
  - Polinomiali
  - Trasformazioni logaritmiche
- Funzioni non lineari a due variabili: interazioni
- Applicazione al dataset dei punteggi nei test della California

Funzioni di regressione non lineari

- Le funzioni di regressione viste finora erano lineari rispetto alla variabile X.
- Ma l'approssimazione lineare non è necessariamente la migliore
- Il modello di regressione multipla può gestire funzioni di regressione non lineari in una o più X.

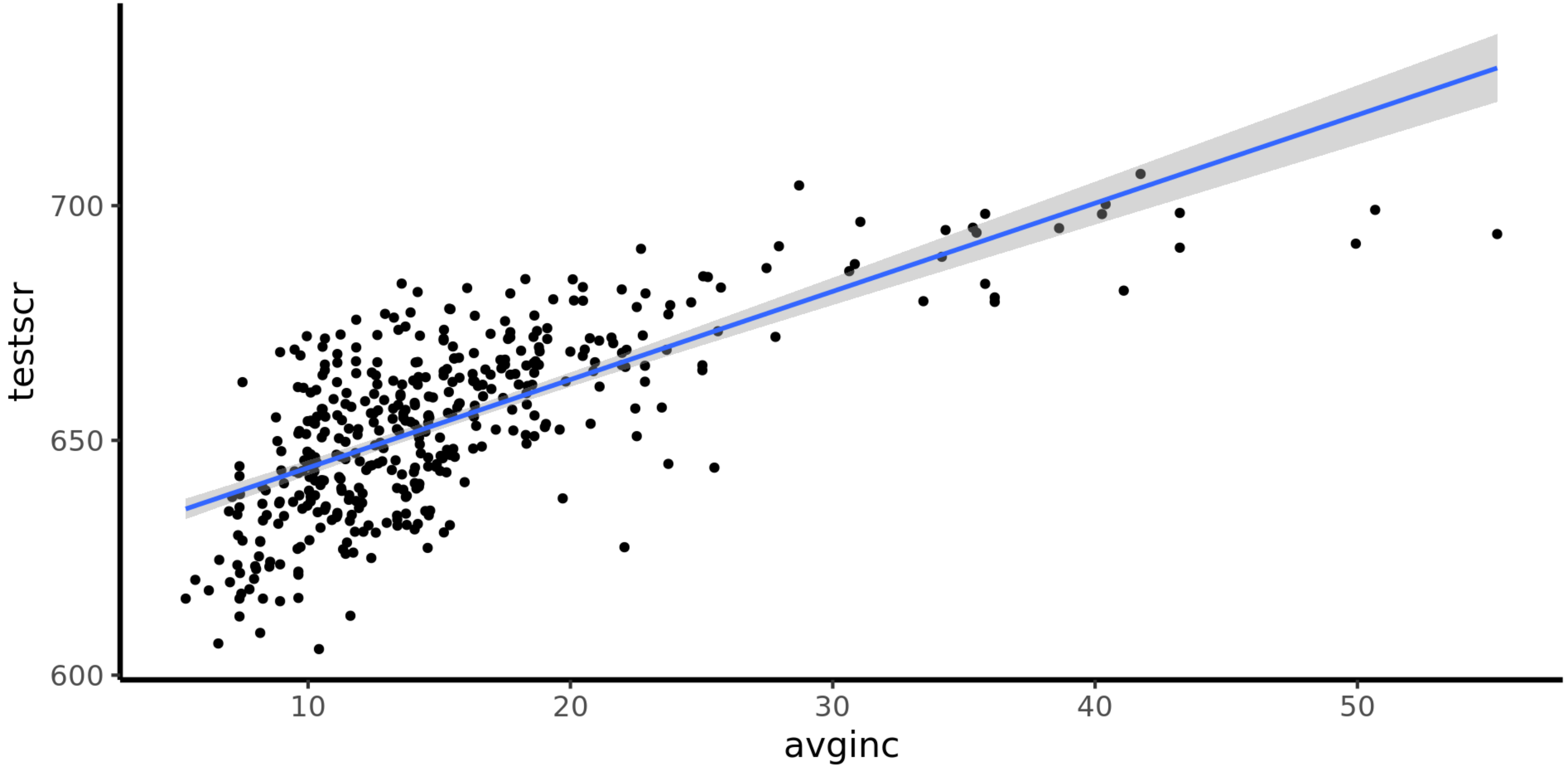
La relazione tra `testscr` e `str` sembra lineare

```r
library(ggplot2)
library(Ecdat)
data(Caschool)
ggplot(Caschool, aes(y=testscr, x=str)) +
  geom_point() +
  geom_smooth(method="lm", SE=FALSE) +
  theme_gragusa()
```



Ma la relazione tra `testscr` e `avginc` sembra non lineare

```r
ggplot(Caschool, aes(y=testscr, x=avginc)) +
  geom_point() +
  geom_smooth(method="lm", SE=FALSE) +
  theme_gragusa()
```



Funzioni di regressione non lineari

Se una relazione tra Y e X è non lineare:

- L'effetto su Y di una variazione in X dipende dal valore di X – ovvero , l'effetto marginale di X non è costante
- Il modello lineare è misspecificato e $E(u|X_1, \ldots, X_n) = 0$ improbabile e quindi
- lo stimatore dell'effetto su Y di X è distorto
- La soluzione: funzione di regressione che sia non lineare in X

La formula generale per una funzione di regressione non lineare

$$Y_i = f(X_{1i}, X_{2i}, \ldots, X_{Ki}) + u_i, \; i = 1, \ldots, n$$

Assunzioni:

1. $E(u_i | X_{1i}, X_{2i}, \ldots, X_{Ki}) = 0$; implica che f è il valore atteso di Y condizionato alle X
2. $(X_{1i}, X_{2i}, \ldots, X_{Ki})$ sono i.i.d.
3. Gli outlier sono rari (stessa idea; la condizione matematica precisa dipende dalla f in esame)
4. Assenza di multicollinearità perfetta (la formulazione precisa dipende dalla f in esame).

La variazione in Y associata a una variazione in X_1 , mantenendo $X_2, \ldots, X_K$ costanti è:

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

Funzioni non lineari di un’unica variabile indipendente (Paragrafo 8.2)

Due approcci complementari:

1. Polinomiali in X

   La funzione di regressione della popolazione viene approssimata da una quadratica, una cubica o una polinomiale di grado più alto

2. Trasformazioni logaritmiche

   Le Y e/o le X vengono trasformate prendendone il logaritmo, che ne dà un’approssimazione “percentuale” utile in molte applicazioni

1. Polinomiali in X

Approssimiamo la funzione di regressione della popolazione con una polinomiale:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i$$

- Modello di regressione lineare multipla, ma i regressori sono potenze di X!
- Stima, verifica delle ipotesi, ecc. procedono come nel modello di regressione multipla con OLS
- I coefficienti sono difficili da interpretare, ma la funzione risultante è interpretabile

Esempio: la relazione tra punteggio nei test e reddito distrettuale

- $avginc_i$ = reddito distrettuale medio nel distretto i (migliaia di dollari pro-capite)

Approssimazione quadratica:

$$testscr_i = \beta_0 + \beta_1 avginc_i + \beta_2 (avginc_i)^2 + u_i$$

Approssimazione cubica:

$$testscr_i = \beta_0 + \beta_1 avginc_i + \beta_2 (avginc_i)^2 + \beta_3 (avginc_i)^3 + u_i$$

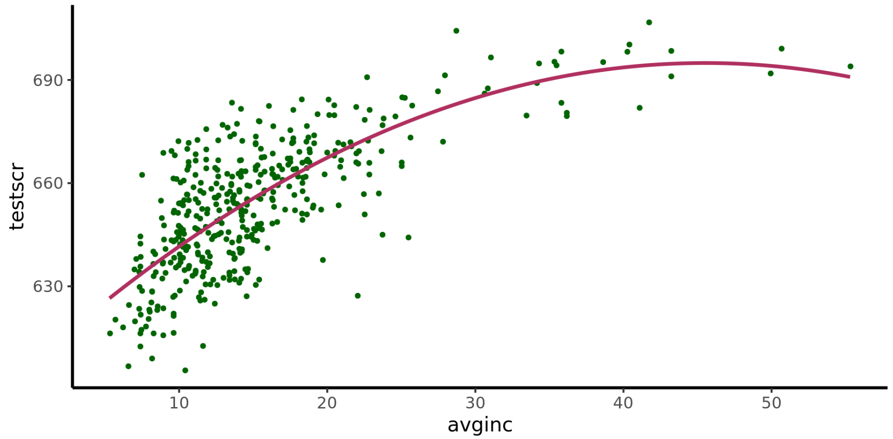
Stima dell'approssimazione quadratica in R

```r
library(estimatr)
lm1 <- lm_robust(testscr~avginc + I(avginc^2), data = Caschool)
lm1
```

```
             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
(Intercept) 607.3017    2.90706  208.91 0.00e+00 601.5874 613.0161 417
avginc        3.8510    0.26900   14.32 4.28e-38   3.3222   4.3798 417
I(avginc^2)  -0.0423    0.00482   -8.78 4.19e-17  -0.0518  -0.0328 417
```

Interpretazione della funzione di regressione

```r
## La variabile X contiene i valori
## di avginc su cui calcolare la
## predizione basata su lm1
Caschool$X <- X <- seq(5.3,55.3,
                       length.out=nrow(Caschool))
## Calacoliamo testscr_hat
Caschool$testscr_hat <-
  predict(lm1, newdata=list(avginc=X))
## Grafico
ggplot(Caschool, aes(y=testscr, x=avginc)) +
  geom_point(col="darkgreen") +
  geom_line(aes(y=testscr_hat, x=X),
    size=1.4, color = "maroon") +
  theme_gragusa()
```



Interpretazione della funzione di regressione stimata:

$$testscr = 607.3 + 3.85 \times avginc_i - 0.0423 \times avginc_i^2$$

Variazione predetta in $testscr$ per una variazione del reddito medio $\$5000 \rightarrow \$6.000$:

$$\begin{aligned}\Delta testscr =& (607.3 + 3.85 \times 6 - 0.0423 \times 6^2) \\ & - (607.3 + 3.85 \times 5 - 0.0423 \times 5^2) = 3.4\end{aligned}$$

"Effetti" attesi in base ai diversi valori di X:

Variazione del reddito ($1000 pro capite)	Var.
da 5 a 6	3.4
da 25 a 26	1.7
da 45 a 46	0.0

L'"effetto" di un cambiamento del reddito è maggiore per i redditi più bassi (forse un beneficio marginale decrescente con l'aumento dei budget delle scuole?)

Attenzione! Qual è l'effetto di una variazione da 65 a 66?

Stima dell'approssimazione cubica in R

```r
library(estimatr)
lm2 <- lm_robust(testscr~avginc + I(avginc^2)+ I(avginc^3), data = Caschool)
lm2
```

```
              Estimate Std. Error t value  Pr(>|t|)  CI Lower  CI Upper  DF
(Intercept)  6.00e+02   5.212350  115.13 1.35e-317  5.90e+02 610.32481 416
avginc       5.02e+00   0.732649    6.85  2.67e-11  3.58e+00   6.45883 416
I(avginc^2) -9.58e-02   0.030616   -3.13  1.88e-03 -1.56e-01  -0.03562 416
I(avginc^3)  6.85e-04   0.000377    1.82  6.98e-02 -5.58e-05   0.00143 416
```

Variazione del reddito ($1000 pro capite)	Var.
da 5 a 6	4.03
da 25 a 26	1.47
da 45 a 46	0.56

Verifica dell'ipotesi nulla di linearità

H_0 : coefficienti di popolazione per $avginc^2$ e $avginc^3$=0

H_1 : almeno uno di questi coefficienti è diverso da zero.

```r
linearHypothesis(lm2, c("I(avginc^2)=0", "I(avginc^3)=0"))
```

```
Linear hypothesis test

Hypothesis:
I(avginc^2) = 0
I(avginc^3) = 0

Model 1: restricted model
Model 2: testscr ~ avginc + I(avginc^2) + I(avginc^3)

  Res.Df Df Chisq Pr(>Chisq)
1    418
2    416  2  67.5    2.2e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'ipotesi che la funzione di regressione della popolazione sia lineare viene rifiutata al livello di significatività dell'1% (vs H_1 polinomiale di grado fino a 3).

Riepilogo: funzioni di regressione polinomiali

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i$$

- Stima: via OLS dopo aver definito nuovi regressori
- I coefficienti hanno interpretazioni complicate
- Per interpretare la funzione di regressione stimata:
  - rappresentare graficamente i valori predetti come funzione di x
  - calcolare gli scarti predetti $\Delta Y / \Delta X$ per i diversi valori di x
- Le ipotesi r (il grado del polinomio) possono essere verificate tramite test (t e $Wald$)
- Scelta del grado r
  - rappresentare i dati graficamente, effettuare i test t e $Wald$, verificare la sensibilità e gli effetti stimati, giudicare.
  - In alternativa usare il criterio di scelta del modello (più avanti).

2. Funzioni logaritmiche di Y e/o X

- $\log(X) =$ è il logaritmo naturale di X
- Le trasformazioni logaritmiche permettono di modellare le relazioni in termini "percentuali" (come l'elasticità) invece che linearmente.

  Ecco perché:

$$\Delta \log(x) = \log(x + \Delta x) – \log(x) = log\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}$$

  (calcolo: $\frac{d \log(x)}{dx} = \frac{1}{x}$)

Numericamente:

- $x = 100$ e $\Delta x = 1$:

  $\frac{\Delta x}{x} = 0.01 \approx \log(101) - \log(100) = 0.00995$
- $x = 100$ e $\Delta x = 5$:

  $\frac{\Delta x}{x} = 0.05$ mentre $\log(105) - \log(100) = 0.04879$

Le tre specificazioni di regressione logaritmica:

- L'interpretazione del coefficiente pendenza è diversa in ciascun caso
- L'interpretazione si trova applicando la regola generale "prima e dopo": predire la variazione in Y per una data variazione in X
- Ogni caso ha una diversa interpretazione naturale (per piccole variazioni in X)

I. Funzione di regressione della popolazione lineare-logaritmica

- Y "prima" e "dopo" aver modificato la X:
  1. "Prima": $Y = \beta_0 + \beta_1 \log(X)$
  2. "Dopo": $Y + \Delta Y = \beta_0 + \beta_1 \log(X + \Delta X)$
- (2)-(1):

$$\begin{aligned}\Delta Y &= \beta_1 \underbrace{[\log(X + \Delta X) - \log(X)]}_{\text{for } \Delta X \to 0\ \approx \Delta X/X} \\ &= \beta_1 \frac{\Delta X}{X} \\ &\implies \beta_1 = \frac{\Delta Y}{\Delta X/X}\end{aligned}$$

Nel modello lineare-logaritmico un incremento dell'1% in X ($\Delta X/X = 0.01$), implica una variazione di $0.01\beta_1$ di Y.

Esempio: testscr su $\log(avginc)$

```r
lm3 <- lm_robust(testscr~log(avginc), data = Caschool)
lm3
```

```
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
(Intercept)    557.8       3.85     145 0.00e+00    550.3    565.4 418
log(avginc)     36.4       1.40      26 1.91e-89     33.7     39.2 418
```

quindi un incremento dell'1% in *income* è associato a un aumento di 0.36 nel punteggio nei test

- Si applicano tutti i soliti meccanismi di regressione: errori standard, intervalli di confidenza, R^2
- Esercizio: confrontare del modello lin-log con il modello cubico

II. Funzione di regressione della popolazione log-lineare

$$\log(Y) = \beta_0 + \beta_1 X + u$$

- Y “prima” e “dopo” aver modificato la X:
  1. “Prima”: $\log(Y) = \beta_0 + \beta_1 X$
  2. “Dopo”: $\log(Y + \Delta Y) = \beta_0 + \beta_1 (X + \Delta X)$
- (2)-(1):

$$\begin{aligned}
\underbrace{\log(Y + \Delta Y) - \log(Y)}_{\text{for } \Delta Y \to 0\ \approx \Delta Y/Y} &= \beta_1 \Delta X \\
\implies \Delta Y/Y &= \beta_1 \Delta X \\
\implies \beta_1 &= \frac{\Delta Y}{Y} / \Delta X
\end{aligned}$$

Quindi nel modello logaritmico-lineare un incremento unitario in X ($\Delta X = 1$) implica una

Esempio: $\log(wage)$ su $educ$

Consideriamo di stimare il rendimento dell'istruzione con il seguente modello di regressione log-lineare:

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

```r
library(wooldridge)
data("cps78_85")
cps85 <- cps78_85 %>% filter(year==85)
lmWage <- lm_robust(lwage~educ, data = cps85)
lmWage
```

```
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
(Intercept)   1.0599    0.10592   10.01 1.03e-21   0.8518   1.2679 532
educ          0.0768    0.00809    9.48 8.15e-20   0.0609   0.0927 532
```

Un anno addizionale di istruzione è associato con salari più elevati di circa il 7.67%.

III. Funzione di regressione log-log

$$\log(Y) = \beta_0 + \beta_1 \log(X) + u$$

- Se $\Delta u = 0$ possiamo calcolare Y “prima” e “dopo” aver modificato la X:
  1. “Prima”: $\log(Y) = \beta_0 + \beta_1 \log(X)$
  2. “Dopo”: $\log(Y + \Delta Y) = \beta_0 + \beta_1 \log(X + \Delta X)$
- (2)-(1): $\log(Y + \Delta Y) - \log(Y) = \beta_1 [\log(X + \Delta X) - \log(X)]$
- Dato che per piccole ΔY e ΔX:

$$\log(Y + \Delta Y) - \log(Y) \approx \frac{\Delta Y}{Y} \text{ e } \log(X + \Delta X) - \log(X) \approx \frac{\Delta X}{X}$$

$$\Longrightarrow \beta_1 = \frac{\Delta Y / Y}{\Delta X / X}$$

Nel modello log-log un incremento di 1% in X implica una variazione di β_1 % in Y

Esempio: log(testscr) su log(avinc)

$$\log(testscr) = \beta_0 + \beta_1 \log(avinc) + u$$

```r
lm4 <- lm_robust(log(testscr)~log(avginc), data = Caschool)
lm4
```

```
              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
(Intercept)     6.3363    0.00593  1067.7 0.00e+00   6.3247   6.3480 418
log(avginc)     0.0554    0.00215    25.8 1.86e-88   0.0512   0.0596 418
```

Un incremento dell'1% in `avginc` è associato a un aumento di 0.0554% nel punteggio nei test.

- e.g.: supponiamo che il reddito salga da 10000 dollari a 11000 dollari, o del 10%.
  - Quindi $testscr$ cresce approssimativamente di $0.0554 \times 10\% = 0.554\%$.
  - Se $testscr = 650$, questo corrisponde a un aumento di $0.00554 \times 650 = 3.6$ punti.
- Come si confronta rispetto al modello log-lineare?

Riepilogo: trasformazioni logaritmiche

- Tre casi , differiscono in base alla o alle variabili Y e/o X trasformate in logaritmi .
- La regressione diventa lineare sulla(e) nuova(e) variabile(i) $\log(Y)$ e/o $\log(X)$, mentre i coefficienti possono essere stimati attraverso l'OLS.
- I test di ipotesi e gli intervalli di affidabilità possono essere implementati e interpretati "nel solito modo".
- L'interpretazione di β_1 differisce caso per caso. La scelta della specificazione (forma funzionale) dev'essere guidata dal ragionamento – quale interpretazione ha più senso nella vostra applicazione? – da test e dall'analisi grafica dei valori predetti.