Principi di Econometria

lezione 9

AA 2015-2016

Paolo Brunori

Precisazione 1

- le assunzioni sono le stesse già viste $+ 1$
- i regressori mostrano collinearità perfetta se uno dei regressori è funzione lineare degli altri
- $\Delta X_1 = \Delta X_2 + \Delta X_3$, caso tipico quando queste sono variabili dicotomiche indicanti per esempio le tre aree di residenza (Nort, Centro, Sud)

Precisazione 2

- se tutte le assunzioni del modello sono vere possiamo interpretare i coefficienti
- può inoltre essere utile valutare quanto precisamente riusciamo a spiegare la variabilità del fenomeno
- $R^2 = 1 - \frac{\sum_i^n \hat{u}_i^2}{\sum_i^n (y_i - \bar{y})^2}$

modello di regressione quadratico

- se il DGP è non lineare il modello che stima Y come funzione lineare dei regressori non è soddisfacente

$$Y = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{1,i}^2 + u_i$$

- che segno mi aspetto per β_1 e β_2?
- come possiamo verificare che la nostra ipotesi sia corretta?

Regressione multipla redditi svedesi

	coefficiente	errore standard	t	$p - value$
β_0	-1166.809	4030.44	-0.289	0.7720
β_{age}	1251.04	178.30	7.016	0.0000
β_{age^2}	-12.509	1.843	-6.786	0.0000

- $R^2 = 0.05652$, R^2 corretto $= 0.05426$
- errore standard di regressione $= \sqrt{\frac{1}{n-k}\sum_i^n u_i^2}$ =12920

qual'è l'effetto di ΔX su Y?

- qui ΔX ha un doppio effetto su Y:

$$Y = -1166.81 + 1251.04 \times age - 12.51 age^2$$

- ΔY dipende dal valore iniziale di X
- fra 18 e 19 anni:

$$\Delta \hat{Y} = \hat{Y}|_{X=19} - \hat{Y}|_{X=18}$$

$$\Delta \hat{Y}|_{X=18 \to 19} = \left(1251.04 \times 19 - 12.51 \times 19^2\right) - \left(1251.04 \times 18 - 12.51 \times 18^2\right)$$

$$\Delta \hat{Y}|_{X=18 \to 19} = 1251.04 - 12.51(19^2 - 18^2) = \beta_1 - 37\beta_2$$

$$\Delta \hat{Y}|_{X=18 \to 19} = 788.17$$

qual'è l'effetto di ΔX su Y?

- qui ΔX ha un doppio effetto su Y:

$$Y = -1166.81 + 1251.04 \times age - 12.51 age^2$$

- ΔY dipende dal valore iniziale di X
- fra 18 e 19 anni:

$$\Delta \hat{Y} = \hat{Y}|_{X=19} - \hat{Y}|_{X=18}$$

$$\Delta \hat{Y}|_{X=48\rightarrow 49} = \left(1251.04 \times 49 - 12.51 \times 49^2\right) - \left(1251.04 \times 48 - 12.51 \times 48^2\right)$$

$$\Delta \hat{Y}|_{X=18\rightarrow 19} = \beta_1 - 97\beta_2$$

$$\Delta \hat{Y}|_{X=48\rightarrow 49} = 37.57$$

come procedere?

1. identificare una possibile relazione non lineare
2. stimare un OLS con funzione non lineare
3. testare se costituisce un miglioramento rispetto a quella lineare
4. rappresentare quello che stiamo facendo graficamente
5. stimare l'effetto di una variazione di X su Y

Regressione polinomiale

- se la relazione è non lineare potrebbe essere meglio approssimata da una funzione più complessa di quella quadratica
- il modello polinomiale di grado r:

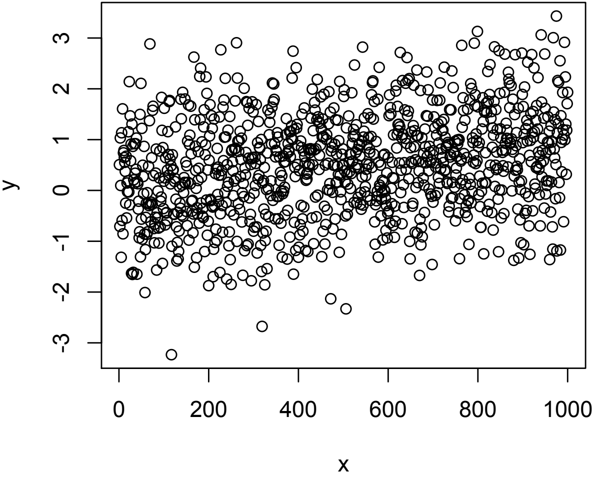
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

- stesse tecniche di inferenza usate nel caso lineare

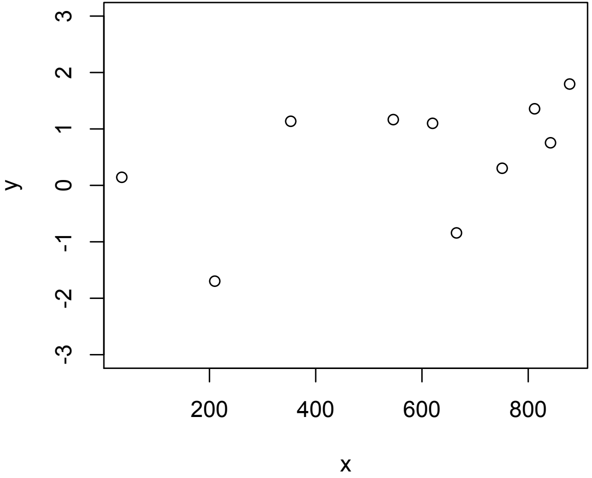
quale grado di polinomio usare?

- sappiamo che se usiamo $r = n$ avremo un'interpolazione perfetta dei dati
- d'altra parte stiamo usando tutte le informazioni a disposizione
- restiamo senza gradi di libertà: l'errore di stima sul campione è zero
- ma non abbiamo variabilità per valutare l'affidabilità delle stime
- mi devo chiedere che modello sceglierei se potessi osservare tutta la popolazione e non solo un campione

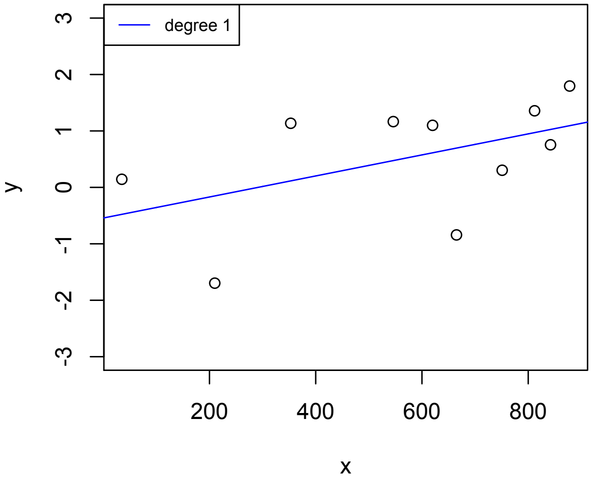
Popolazione



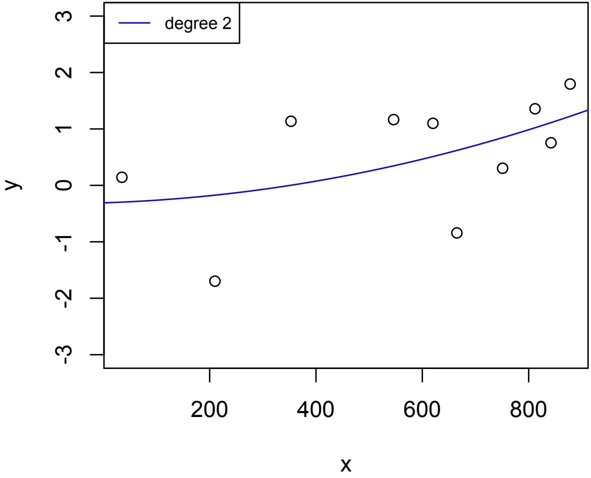
Campione $n = 10$



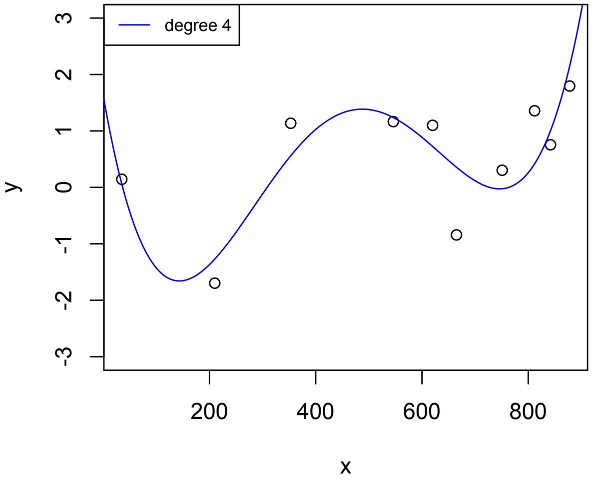
$$Y = \beta_0 + \beta_1 X$$



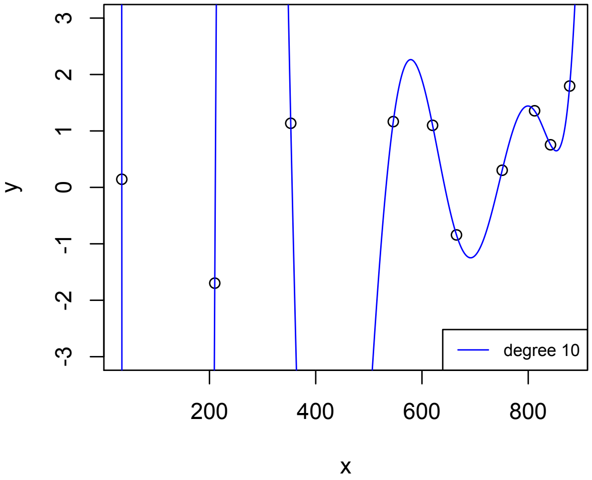
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$



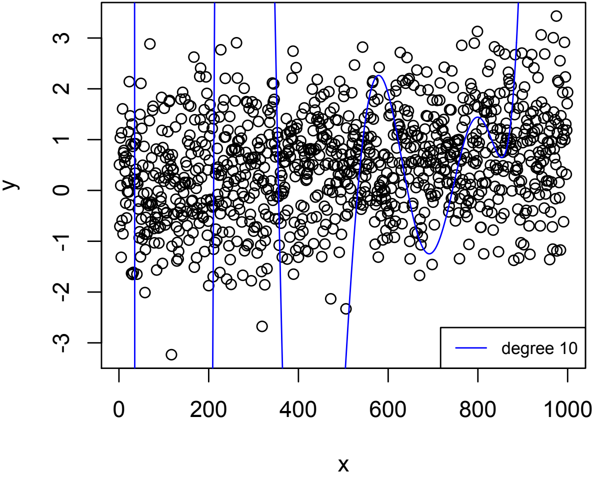
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_4 X^4$$



$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n$$



$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_n X^n$$



Quale grado di polinomio usare?

- un'interpolazione perfetta sul campione non implica una buona capacità di spiegare la relazione nella popolazione
- interpolare perfettamente un campione significa interpretare anche la componente casuale come parte della relazione
- il potere predittivo di un modello ottenuto così è inferiore al potere predittivo di un modello polinomiale di grado inferiore
- quindi non dobbiamo cercare il polinomio che meglio interpola i dati del campione!

Polinomi

	coefficiente	errore standard	t	$p-value$
β_0	1.0812	0.2509	4.309	0.00258
β_X	-0.8836	0.7934	-1.114	0.29777

- $R^2 = 0.1342$, R^2 corretto $= 0.026$
- errore standard di regressione $= \sqrt{\frac{1}{n-k}\sum_i^n u_i^2} = 0.7934$

Polinomi

	coefficiente	errore standard	t	$p - value$
β_0	1.0812	0.2509	4.096	0.00459
β_X	-0.8836	0.8347	-1.059	0.32494
β_{X^2}	0.3990	0.8347	0.478	0.64718

- $R^2 = 0.1616$, R^2 corretto $= -0.07795$
- errore standard di regressione $= \sqrt{\frac{1}{n-k}\sum_i^n u_i^2}$ =0.8347

Polinomi

	coefficiente	errore standard	t	$p - value$
β_0	1.0812	NA	NA	NA
β_X	-0.8836	NA	NA	NA
β_{X^2}	0.3990	NA	NA	NA
β_{X^3}	-0.3463	NA	NA	NA
β_{X^4}	-1.3649	NA	NA	NA
β_{X^5}	-0.1688	NA	NA	NA
β_{X^6}	0.2984	NA	NA	NA
β_{X^7}	-0.6287	NA	NA	NA
β_{X^8}	-0.8308	NA	NA	NA
β_{X^9}	1.3003	NA	NA	NA

- $R^2 = 1.0000$, R^2 corretto $= NAN$
- errore standard di regressione $= \sqrt{\frac{1}{n-k} \sum_i^n u_i^2}$ =NAN

Regressione multipla redditi svedesi: quale polinomio usare?

	coefficiente	errore standard	t	$p - value$
β_0	-14693.729	4149.33	-3.541	0.0004
β_{age}	1119.13	171.52	6.522	0.0000
β_{age^2}	-10.407	1.785	-5.831	0.0000
β_{sex}	1054.07	122.67	8.592	0.0000
β_{edu}	1776.77	856.584	2.074	0.0383

- $R^2 = 0.138$, R^2corretto $= 0.1339$
- errore standard di regressione $= \sqrt{\frac{1}{n-k}\sum_i^n u_i^2}$=12370

Regressione multipla redditi svedesi: quale polinomio usare?

	coefficiente	errore standard	t	$p - value$
β_0	19625.42	10123.7622	1.939	0.05289
β_{age}	-1486.15	722.42	-2.057	0.0399
β_{age^2}	48.51	15.97	3.037	0.00247
β_{age^3}	-0.412	0.1111	-3.711	0.00022
β_{edu}	1094.39	122.23	8.953	0.0000
β_{sex}	1851.75	850.32	2.178	0.02971

- $R^2 = 0.1521$, R^2corretto $= 0.147$
- errore standard di regressione $= \sqrt{\frac{1}{n-k} \sum_i^n u_i^2} = 12270$

interazioni tra regressori

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

- dove $D_{1,i}$= laureata/o e $D_{2,i}$ = sesso
- in questo modello la laurea ha un effetto identico per i due gruppi
- è possibile considerare un possibile effetto di interazione fra le due variabili:

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

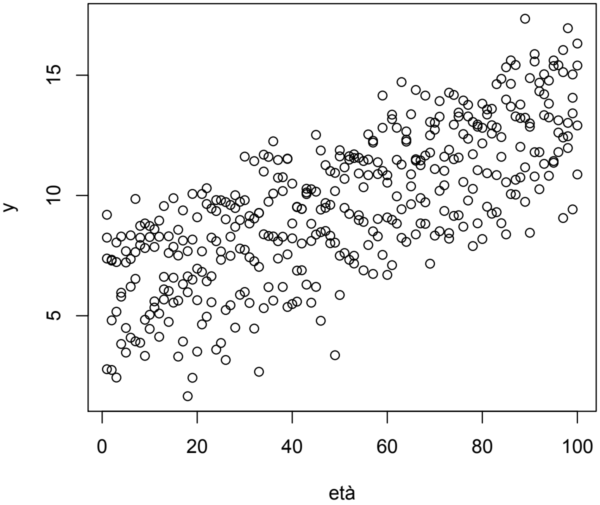
l'effetto di avere una laurea

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

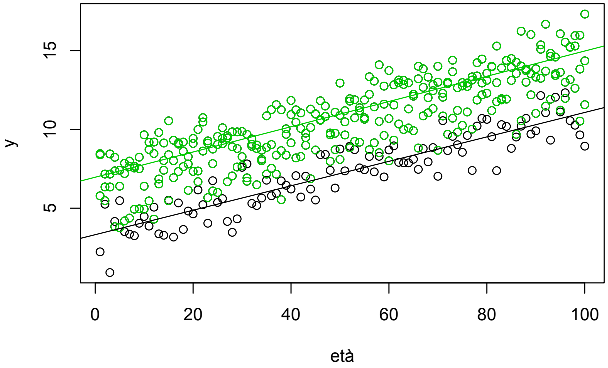
$D_{1,i}$= laureata/o e $D_{2,i}$ = sesso

- $E(Y|D_{1,i} = 0, D_{2,i} = d_2) =$
  $\beta_0 + \beta_1 \times 0 + \beta_2 d_2 + \beta_3 \times 0 \times d_2 = \beta_0 + \beta_2 d_2$
- $E(Y|D_{1,i} = 1, D_{2,i} = d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 =$
  $\beta_0 + \beta_1 + (\beta_2 + \beta_3) d_2$
- l'effetto di avere o meno una laurea D_1 dipende dal sesso (D_2): $\beta_1 + \beta_3 d_2$

premio di laurea e sesso

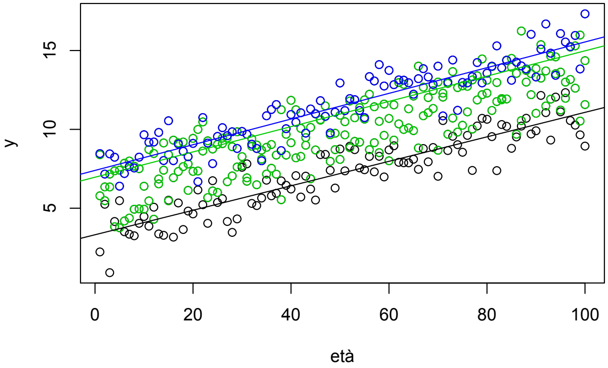


premio di laurea e sesso



Regressione con variabile dicotomica sesso

premio di laurea e sesso



Regressione con variabile dicotomica sesso e interazione fra sesso e laurea

interazione fra variabili binarie e continue

- $D =$ uomo, X=esperienza lavorativa.
- modello base:

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

- in pratica stiamo stimando due rette parallele di regressione
- ma nella realtà le prospettive di carriera potrebbero dipendere dal sesso
- anche qui si può aggiungere un termine di interazione

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

- se si tratta di un uomo la funzione di regressione è:

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

- se donna:

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

- per cui ho intercette diverse e pendenze diverse

premio di laurea e sesso