Principi di Econometria

lezione 14

AA 2016-2017

Paolo Brunori

4.dati mancanti

- Y non è osservato ma ciò non dipende nè dal valore di X né dal valore di Y
- Y non osservato ciò dipende dal valore di X ma non dal valore di Y
- Y non osservato e ciò dipende dal valore di Y

Y non è osservato ma ciò non dipende nè dal valore di X né dal valore di Y

- riduzione della numerosità campionaria
- stime corrette
- meno gradi di libertà

Y non osservato ciò dipende dal valore di X ma non dal valore di Y

- riduzione della numerosità campionaria
- stime corrette solo per i valori per i quali X e Y sono osservabili
- meno gradi di libertà $\rightarrow$ ES più ampi
- meno variabilià di X $\rightarrow$ ES più ampi

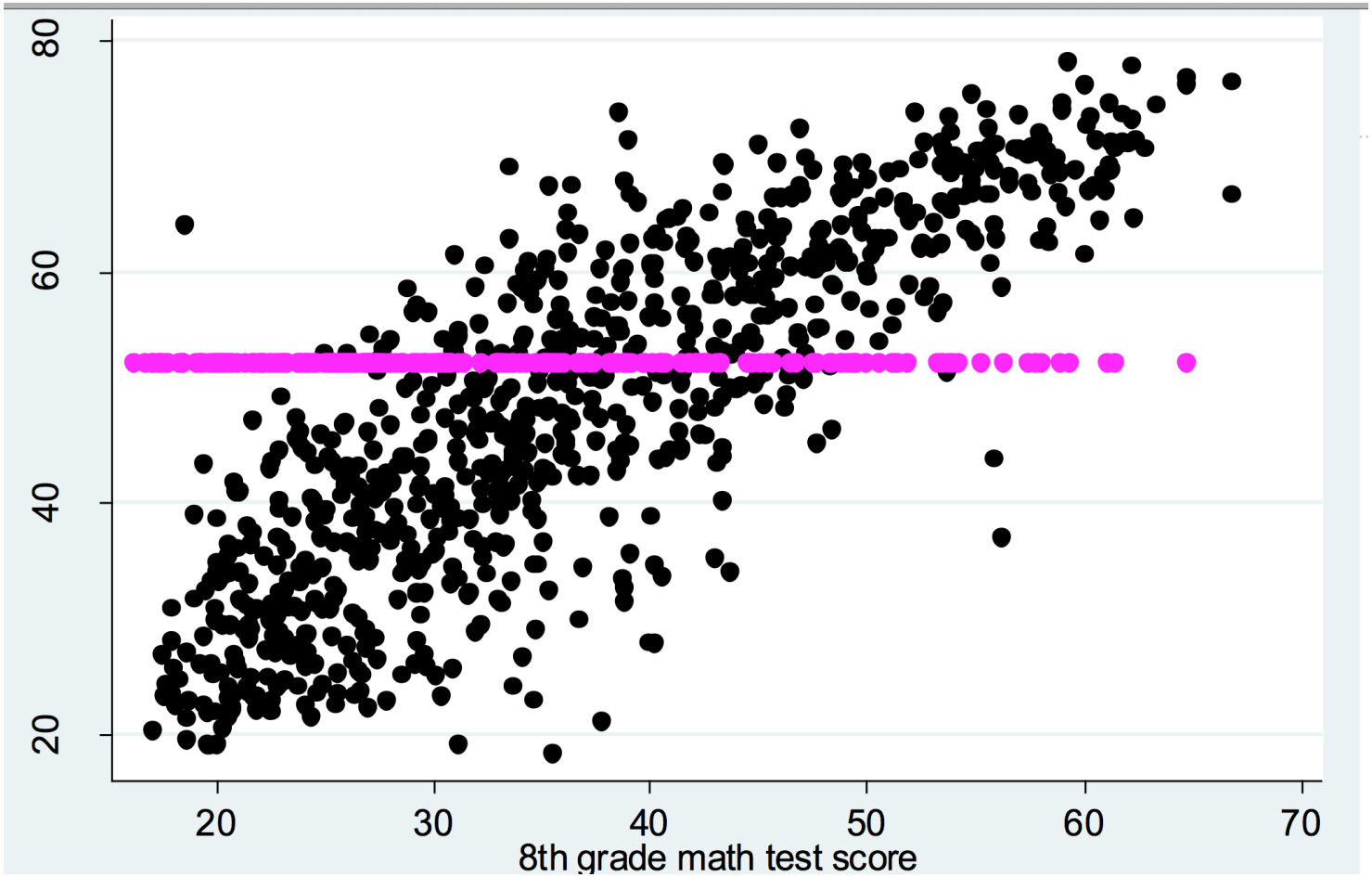
Y non osservato e ciò dipende dal valore di Y

- cercando di stimare la relazione ORE studiate - VOTO
- supponiamo di non osservare persone che studiano prendono voto inferiore a 18
- in questo caso per valori bassi di ORE troveremo solo voti superiori a 18 e il nostro coefficiente sarà distorto

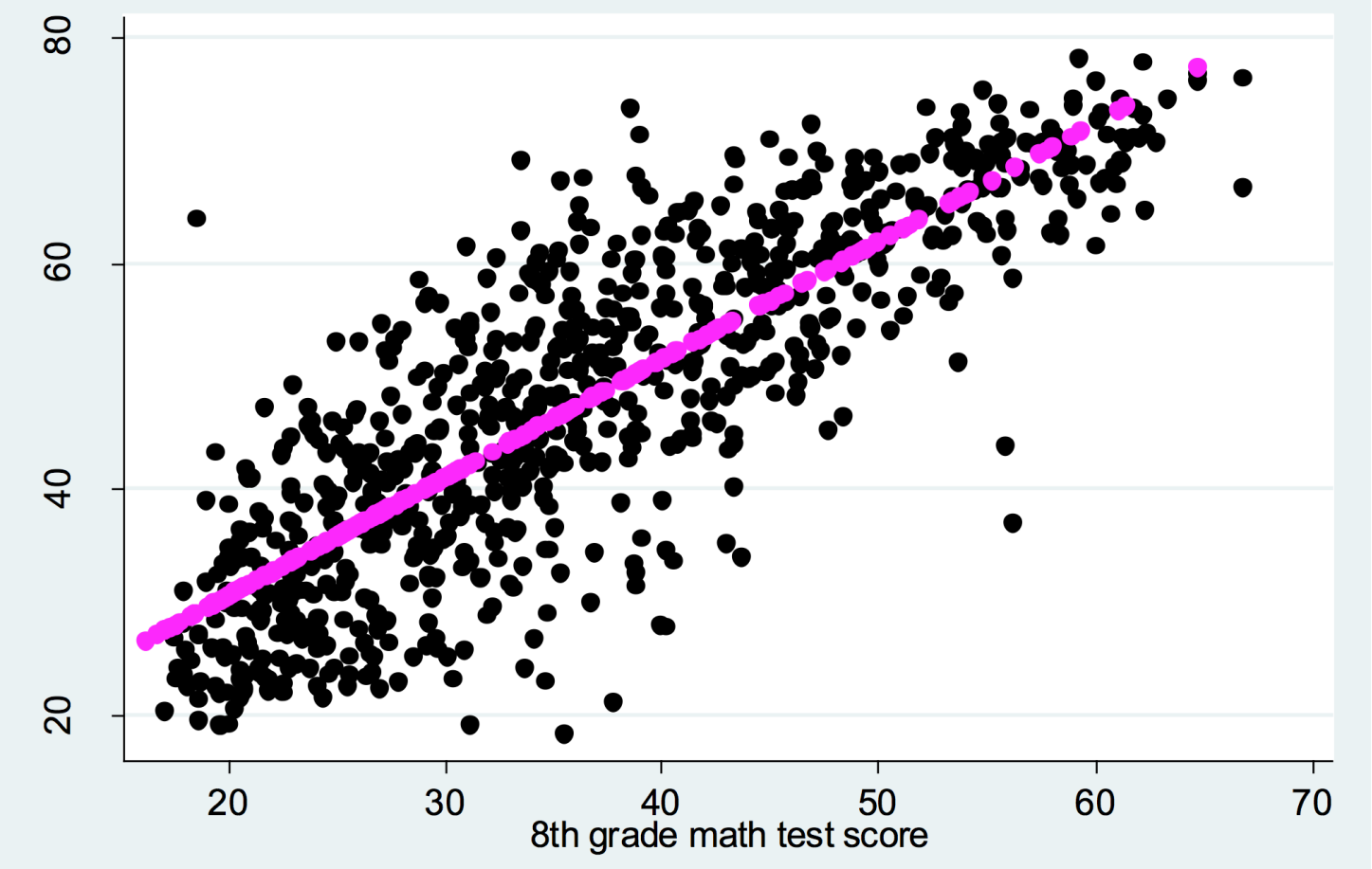
Come ci si comporta quando vi sono valori mancanti?

- ci sono alcune tecniche di imputazione (moda, media)
- queste hanno un effetto sulle stime

Imputazione con valore medio di Y



Imputazione con valore predetto di Y



si usano tutte le informazioni ma si sopravvaluta l'accuratezza della stima

imputazione di dati mancanti

- nel primo caso la stima della relazione viene attenuata
- ciò potrebbe sia correggere che peggiorare la distorsione
- nel secondo caso se la stima è distorta rimarrà tale
- se l'imputazione si basa solo sul valore di X le informazioni usate sono le stesse
- si sopravvaluta l'accuratezza del coefficiente

5. causalità simultanea

- perché le classi sono formate da circa 30 studenti?
- cosa si valuta per determinare questo numero?
- immaginate di voler valutare se un numero minore di studenti per insegnante favorisca l'apprendimento
- come procedete?

Causalità simultanea

- avendo dati riguardo alla numerosità delle classi e risultati scolastici si può cercare di isolare l'effetto della numerosità della classe
- ma è possibile che la dimensione delle classi sia in qualche modo decisa sulla base dei risultati?
- se così fosse saremmo in presenza di una situazione di causalità simultanea
- situazione molto complessa per riuscire ad ottenere stime affidabili

Causalità simultanea

- abbiamo sempre ipotizzato $X \rightarrow Y$
- in molti casi è vero anche che $Y \rightarrow X$
- in questo caso si parla di causalità simultanea
- la conseguenza è correlazione fra regressore ed errore:

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

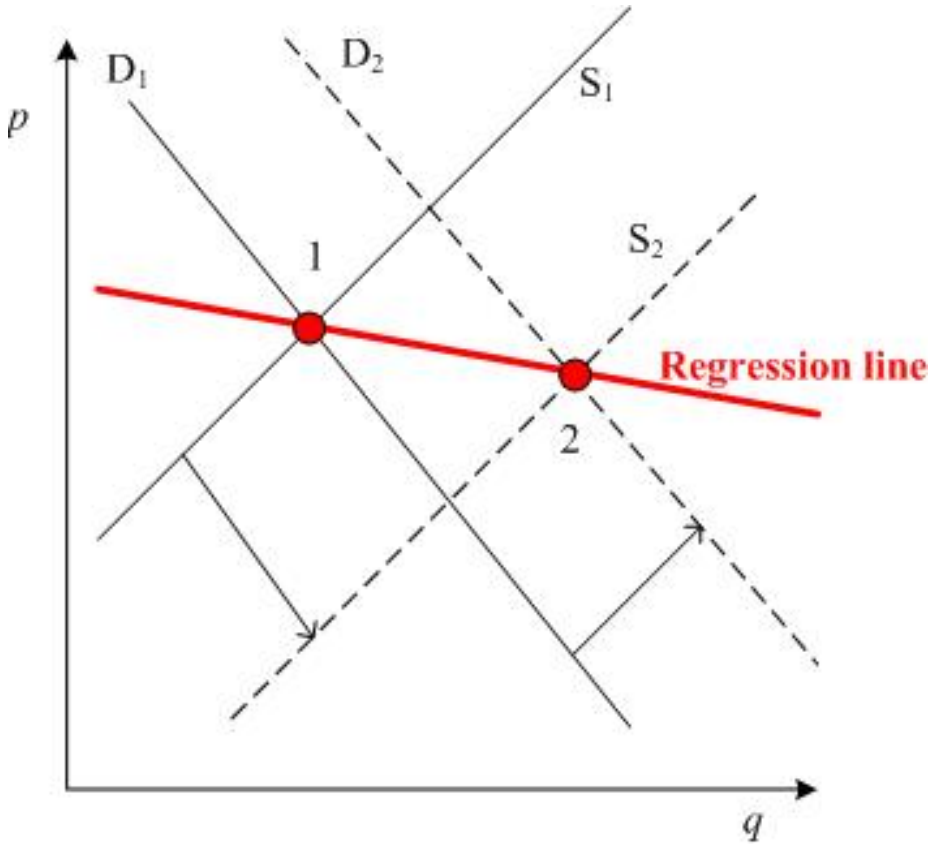
$$X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

Causalità simultanea

- esempio: $u_i < 0 \rightarrow Y_i < \hat{Y}_i$
- Y_i piccolo influenza X_i attraverso γ_1
- quindi u e X sono correlati
- le stime sono distorte
- iproblema non facilmente risolvibile

Endogeneità: il caso di prezzo e quantità di un bene

- $Q_D = a - bP$
- $Q_S = c + dP$



Soluzioni per la causalità simultanea

- le soluzioni non sono semplici
- è necessario osservare una variazione di X che avvenga indipendentemente da Y
- come avviene ad esempio negli esperimenti controllati in laboratorio per altre scienze
- ci sono casi fortunati in cui questo avviene, si tratta di situazioni quasi-sperimentali o sperimentali
- in alternativa ci sono tecniche statistiche che sfruttano la covarianza della variabile X con altre variabili non correlate con Y

Variabili strumentali: intuizione

- supponiamo di voler stimare la funzione di domanda di ciliegie
- il prezzo determina la quantità domandata ($Q_D = a - bX_P$)
- ma la quantità domandata determina il prezzo ($Q_S = c + dX_P$)
- questa relazione implica che l'errore di regressione è correlato con il prezzo
- se trovo una variabile G che è correlata con il prezzo ma non con la quantità domandata questa è una variabile strumentale
- regredendo Q_D sulla parte della variabilià di X_P correlata con G ottengo stime consistenti di $\hat{b}$
- gli errori standard saranno maggiori

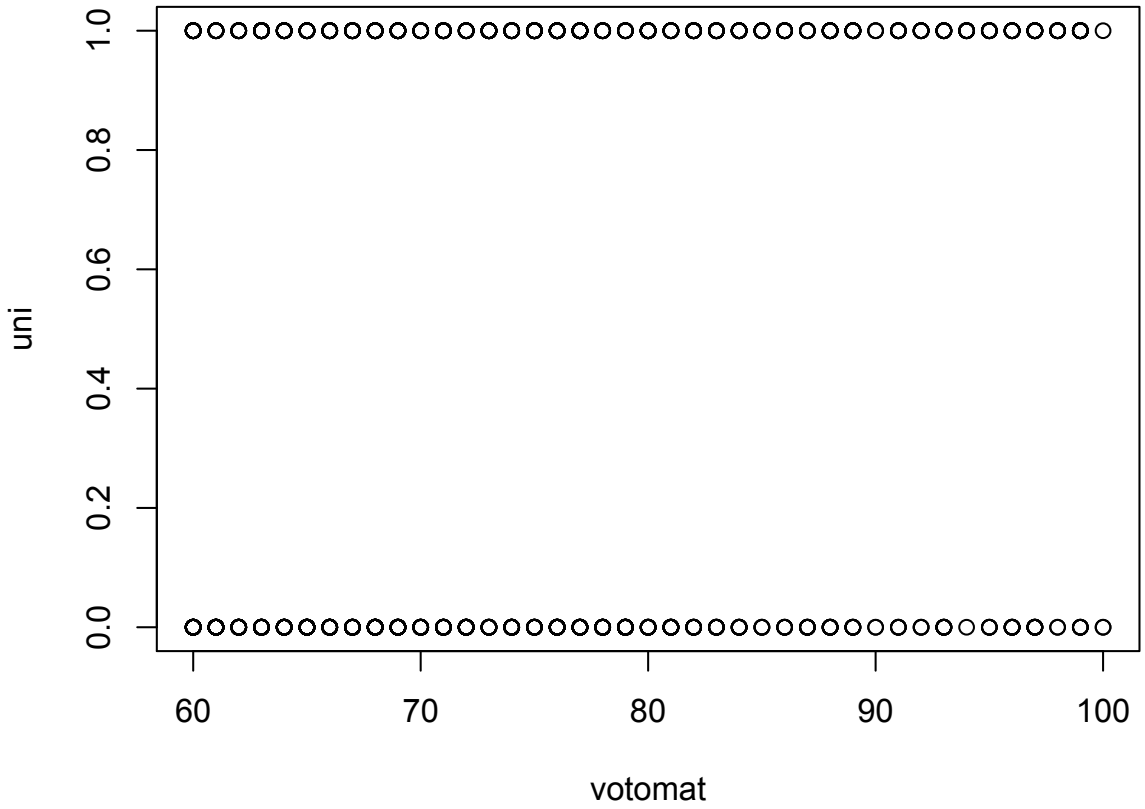
Chi si iscrive all'università?

- fra le variabili nel campione c' è la risposta alla domanda: 'si è mai iscritto all'università?'
- in questo caso stiamo cercando di spiegare cosa influenza una variabile dicotomica
- si tratta di capire cosa influenzi la probabilità di iscriversi
- non si osserva però ovviamente la probabilità ma solo se una certa persona si è iscritta o meno
- stiamo stimando un modello di probabilità, se per effettuare questa stima si utilizza un OLS si tratta di un modello lineare di probabilità (MLP)

regressione con variabile dipendente binaria

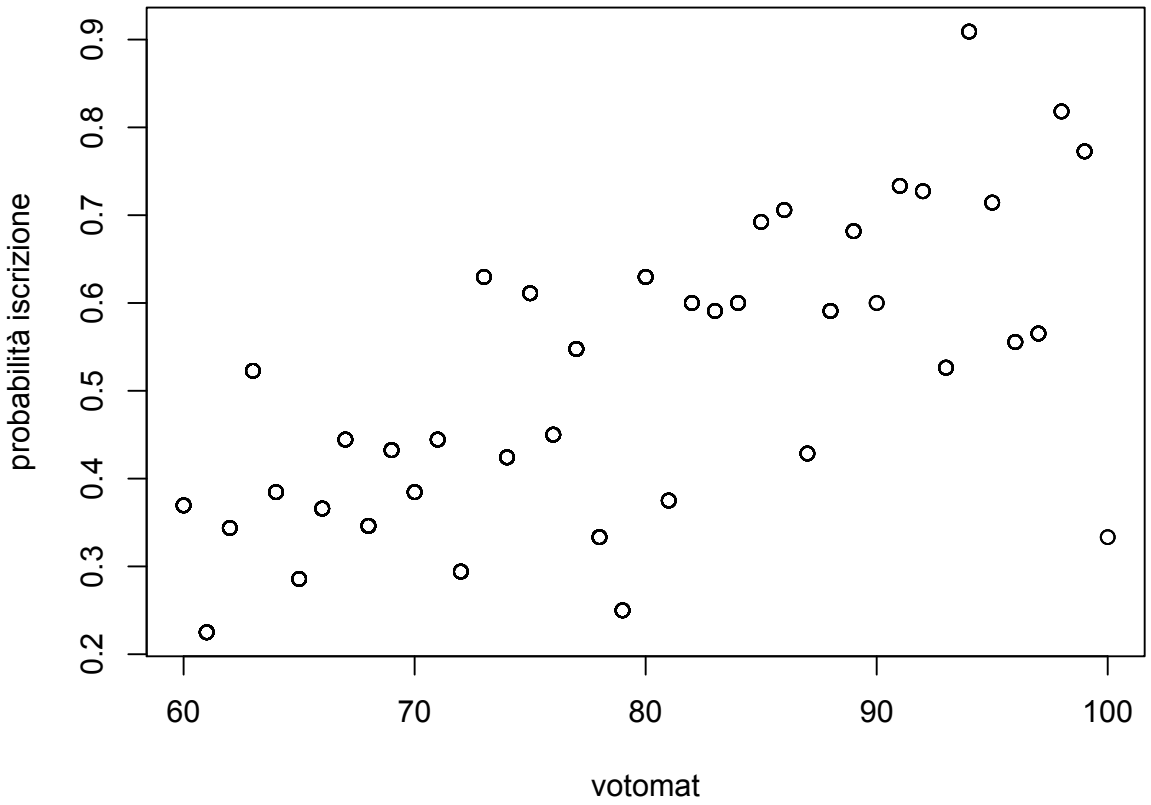
- un fattore rilevante per la probabilità di iscriversi all'università è il voto alla maturità
- la rappresentazione di questi dati chiarisce che si tratta di dati particolari

diagramma a nuvola per *VOTOMAT* e *uni*



Perché non si vede nessuna relazione fra le due variabili?

diagramma a nuvola per *VOTOMAT* e *probabilità di iscriversi all'università*



regressione $UNI = \beta_0 + \beta_1\ VOTOMAT$

- il nostro modello
- $E(\,Y|X_1, ..., X_k) = \Pr(\,Y = 1|X_1, ..., X_k)$
- la stima OLS restituisce
  $Y = -0.2570 + 0.0096\,VOTOMAT$
- l'interpretazione è identica a quella per l'OLS di una variabile cardinale

inferenza

- intervalli di confidenza e test delle ipotesi come nel caso dell'OLS
- variabili omesse, eteroschedasticità rimangono una minaccia
- R^2 invece non può essere utilizzato
- perché?
- riuscite ad immaginare un caso in cui $R^2 = 1$?
- il software vi resituisce l' R^2 ma voi non dovete interpretarlo

Probabilià di iscriversi all'università

- Nel modello completo inseriamo le variabili: *donna*, *VOTOMAT*, *Area_Istat*, *edu_padre*, *tipo_scuola*

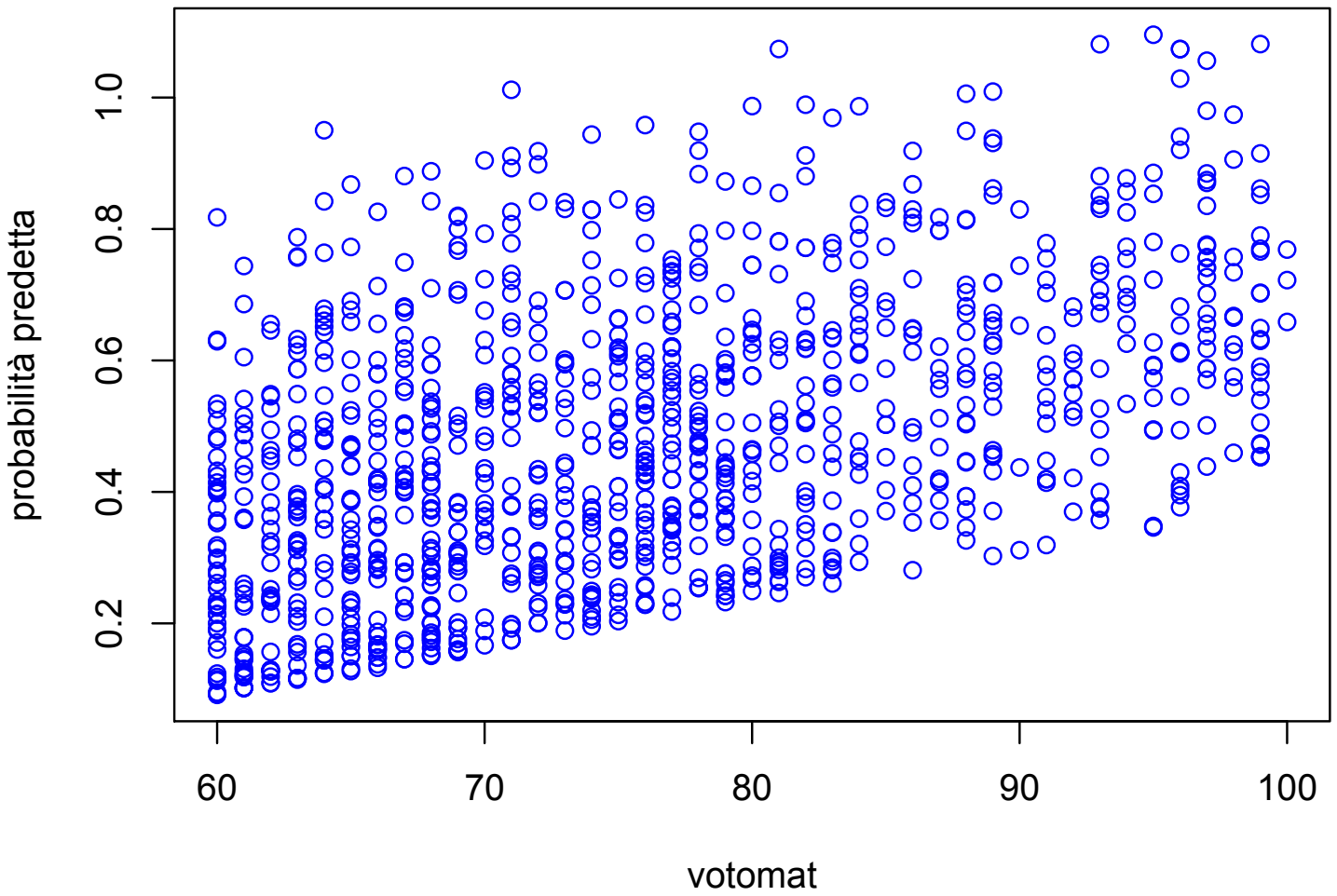
```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.2433064  0.1196088  -2.034 0.042174 *
votomat                     0.0072334  0.0013297   5.440 6.58e-08 ***
donna                       0.0184420  0.0298975   0.617 0.537468
as.factor(area_istat)2     -0.0015889  0.0446642  -0.036 0.971629
as.factor(area_istat)3      0.0292962  0.0395687   0.740 0.459223
as.factor(area_istat)4      0.0006123  0.0418955   0.015 0.988341
as.factor(area_istat)5      0.0101911  0.0468964   0.217 0.828007
as.factor(votmed)distinto -0.0288482  0.0424698  -0.679 0.497115
as.factor(votmed)ottimo     0.0897947  0.0572562   1.568 0.117101
as.factor(votmed)suff      -0.0765869  0.0324861  -2.358 0.018573 *
as.factor(edu_padre)2      -0.0234567  0.0429250  -0.546 0.584863
as.factor(edu_padre)3       0.1055730  0.0436585   2.418 0.015762 *
as.factor(edu_padre)4       0.3063408  0.0590179   5.191 2.50e-07 ***
privata                     0.1329299  0.0309898   4.289 1.95e-05 ***
as.factor(tscuola)2         0.0037162  0.0539890   0.069 0.945136
as.factor(tscuola)3         0.1831178  0.0553334   3.309 0.000966 ***
as.factor(tscuola)4         0.0052405  0.0655742   0.080 0.936318
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4541 on 1093 degrees of freedom
Multiple R-squared:  0.1854,  Adjusted R-squared:  0.1735
```

pregi e difetti del MLP

- interpretabilità
- facilità di stima
- se X può variare fra $-\infty$ e $+\infty$ quali valori può assumere $\hat{Y}$?
- è possibile che β sia lineare?

Le probabilità predette dal modello



pregi e difetti del MLP

- un effetto lineare dei regressori implica che, a meno che i regressori non assumano valori all'interno di un intervallo esiste un valore del regressore tale per cui la probabilità è negativa (o maggiore di zero)
- questo ovviamente è impossibile
- come si risolve il problema?
- o si accetta che i coefficienti stimati hanno un significato che riguarda soltanto alcuni valori dei regressori (quelli intermedi)
- o si cerca una specifica funzionale diversa nella quale la variabile dipendente è confinata nell'intervallo [0,1] e l'effetto dei regressori diventa infinitamente piccolo quando i regressori assumono valori molto grandi (in positivo o negativo)

regressione probit e logit

- vogliamo costringere $\hat{Y} \in [0, 1]$
- un modo per farlo è quello di utilizzare una funzione cumulativa di distribuzione
- le due varianti utilizzate sono quelle basate sulla c.d.f. normale e quelle basate sulla c.d.f. logistica
- la prima è utilizzata nel modello probit la seconda nel logit
- questi due modelli non fanno parte del programma

interpretazione di un modello probit

FIGURE 11.2 **Probit Model of the Probability of Denial, Given the P/I Ratio**

The probit model uses the cumulative normal distribution function to model the probability of denial given the payment-to-income ratio or, more generally, to model $\Pr(Y = 1|X)$. Unlike the linear probability model, the probit conditional probabilities are always between 0 and 1.