Principi di Econometria

lezione 13

AA 2015-2016

Paolo Brunori

Variabile indipendente qualitativa

- Spesso capita di avere variabili fra i regressori che descrivono un fenomeno qualitativo
- abitare nel nord Italia, essere donna, essere laureato
- questi fenomeni possono essere inseriti come regressori soltanto utilizzando variabili dicotomiche
- un modello che spiega un risultato scolastico:

$$Y = \beta_0 + \beta_1 Nord + \beta_2 Donna$$

- in questo caso tutti i β colgono l'effetto di variabili dicotomiche

Attenzione alle variabili qualitative

- Ma nelle indagini statistiche alcune variabili qualitative possono indurci in errore
- L'Istat ad esempio tipicamente divide i residenti in macreoaree:
  1=Nord-ovest; 2=Nord-est; 3=Centro; 4=Sud; 5=Isole
- Se regrediamo il voto finale alla scuola secondaria superiore così:

$$Y = \beta_0 + \beta_1 Donna + \beta_2 Area\ Istat$$

- commettiamo un serio errore

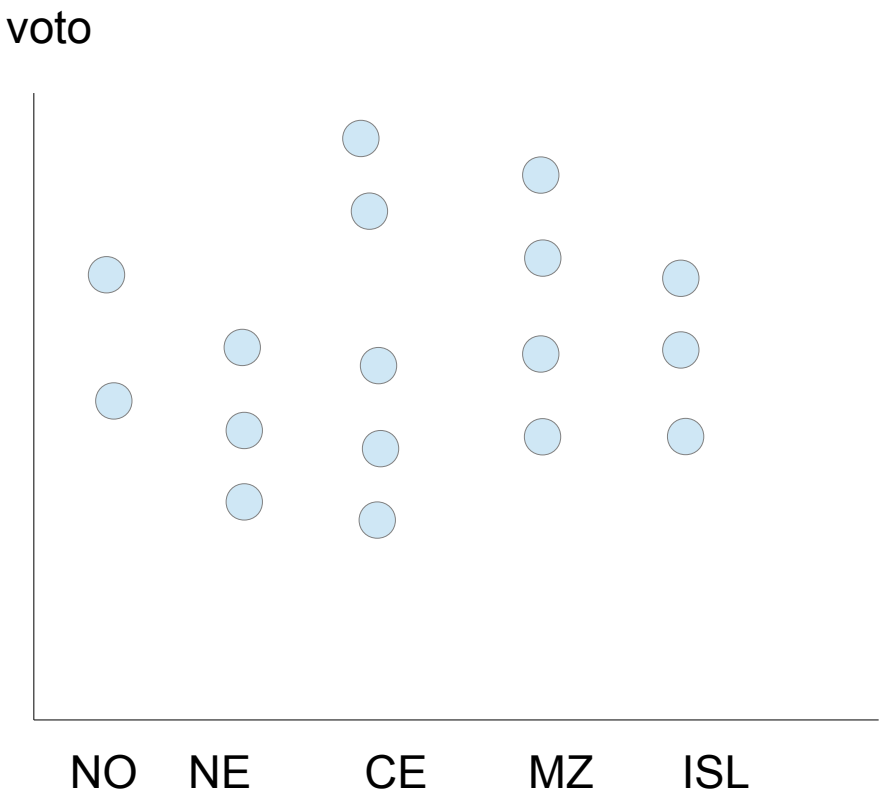
Attenzione alle variabili qualitative

- Stiamo dando un valore cardinale a una variabile qualitativa
- Immaginate che il risultato sia il seguente

	coefficiente	errore standard	t	$p - value$
β_0	72.488	11.0437	69.454	0.0000
β_{donna}	2.9786	0.8116	3.670	0.0002
$\beta_{arealstat}$	0.3409	0.3051	1.117	0.2642

- come potete interpretare i coefficienti?

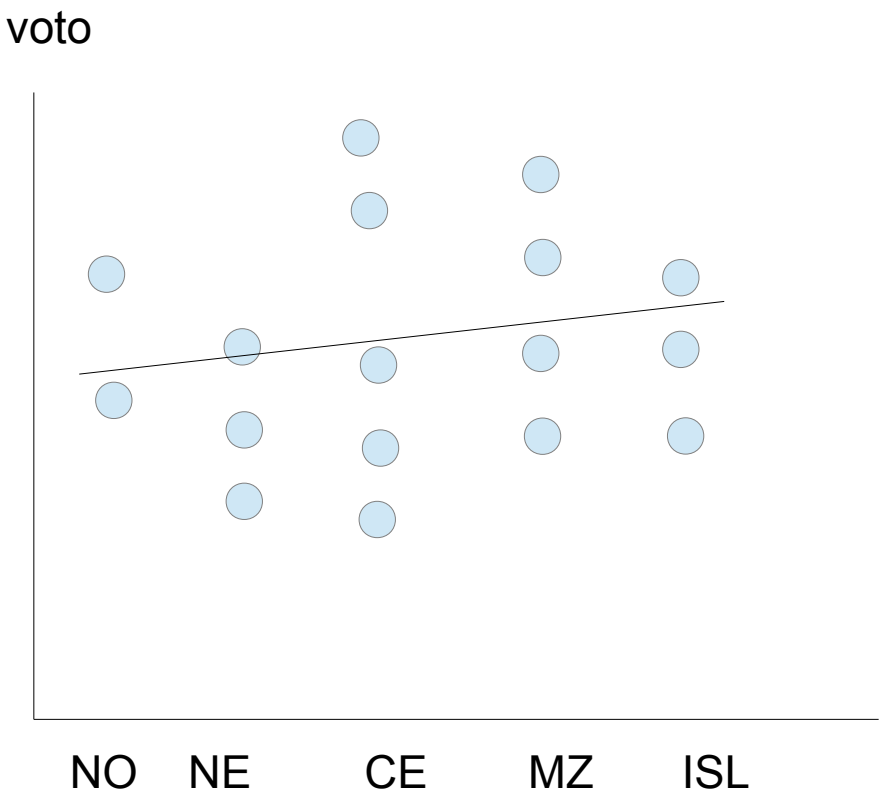
Area di residenza e voti



Fare attenzione alle variabili qualitative!

Area di residenza e voti

Utilizzando la variabile VOTOMAT come codificata dall'Istat:



Area di residenza e voti

Ma noi forse abbiamo in mente di identificare un effetto fisso di area geografica:

Ricodifica delle variabili qualitative

- per identificare l'effetto fisso dell'area geografica specifichiamo il modello:

$$Y = \beta_0 + \beta_1 Donna + \beta_2 Nord\text{-}Ovest + \beta_3 Nord\text{-}Est +$$

$$+\beta_4 Centro + \beta_5 Sud + \beta_6 Isole$$

- ognuna delle variabili è ottenuta creando una variabile dicotomica
- Sud ad esempio assume valore 1 quando Area Istat $= 4$ e assume valore 0 in ogni altro caso
- purtroppo anche questo modello è problematico. Perchè?

Variabili qualitative e multicollinearità perfetta

- non è possibile stimare il coefficiente per tutte le variabili dicotomiche

	coefficiente	errore standard	t	$p - value$
β_0	72.7407	0.9376	77.582	0.0000
β_{donna}	2.9491	0.8152	3.618	0.0003
β_{NE}	0.8470	1.3494	0.628	0.5303
β_C	0.7013	1.1625	0.603	0.5465
β_{SUD}	0.8867	1.2148	0.730	0.4656
β_{ISL}	1.7455	1.4242	1.226	0.2207

- come potete interpretare i coefficienti?
- perché il coefficiente del Nord-Ovest è assente?

Variabili qualitative e multicollinearità perfetta

- in alternativa si possono tenere tutte le variabili che descrivono le macroaree ma è necessario eliminare l'intercetta

	coefficiente	errore standard	t	$p-value$
β_{donna}	2.949	0.8152	3.618	0.0003
β_{NO}	72.7407	0.9376	77.582	0.0000
β_{NE}	73.5877	1.1128	66.128	0.0000
β_{C}	73.4420	0.857	85.668	0.0000
β_{SUD}	73.6274	0.9084	81.050	0.0000
β_{ISL}	74.4862	1.1868	62.760	0.0000

- in questo caso i coefficienti delle macroaree sono l'intercetta di ogni macroarea
- mentre prima si interpretavano come distanza dell'intercetta dalla macroarea di base (NO)

Regressori ordinali

- per l'aerea di residenza non sapevamo esattamente cosa attenderci in quanto a voti
- altre variabili invece identificano variabili qualitative che identificano fenomeni intrinsecamente ordinali
- l'esempio nel dataset è l'istruzione del padre (variabile edu_padre) che assume valori:
  1=elementare, 2=media inferiore, 3=media superiore, 4=universitaria
- anche in questo caso è molto pericoloso usare la variabile così come codificata dall'Istat
- pur essendo vero che:
  elementare<*media inferiore*<*media superiore*<*universitaria*

Regressori ordinali

- se si specifica il modello

$$Y = \beta_0 + \beta_1 Donna + \beta_2 Nord\text{-}Ovest + \beta_3 Nord\text{-}Est +$$

$$+\beta_4 Centro + \beta_5 Sud + \beta_6 Isole + \beta_7 edu\_padre$$

- questo implica imporre che la differenza di voto attesa fra avere un padre con licenza media invece che con la licenza elementare è la stessa differenza di voto attesa fra avere un padre con laurea invece che con licenza media superiore
- anche in questo caso è prudente stimare una variabile dicotomica per ogni titolo di studio

Determinanti del voto di laurea

- stessa cosa vale per la variabile *VOTOMED* che assume valori da 'sufficiente' a 'ottimo'
- aggiungiamo anche il tipo di scuola frequentata (liceo, istituto tecnico,...) e se si tratta di una scuola privata o pubblica
- il nostro output di regressione diventa lungo per cui mostriamo solo i coefficienti statisticamente significativi

Determinanti del voto di laurea

	coefficiente	errore standard	t	$p-value$
β_0	73.2902	1.5768	46.481	0.0000
β_{donna}	1.4040	0.6785	2.069	0.0387
β_{ISL}	1.7942	1.0649	1.685	0.0923
$\beta_{DISTINTO}$	4.1482	0.9575	4.332	0.0000
β_{OTTIMO}	7.148	1.283	5.569	0.0000
β_{SUFF}	-2.405	0.7351	-3.272	0.0011
$\beta_{EDUP=SECINF}$	-1.8663	0.9744	-1.915	0.0557
$\beta_{PRIVATA}$	-2.1789	0.7015	-3.106	0.0019
β_{LICEO}	3.0605	1.2547	2.439	0.0148

RSE$=10.33$, $R^2=0.1251$, $Adj-R^2$=0.1132

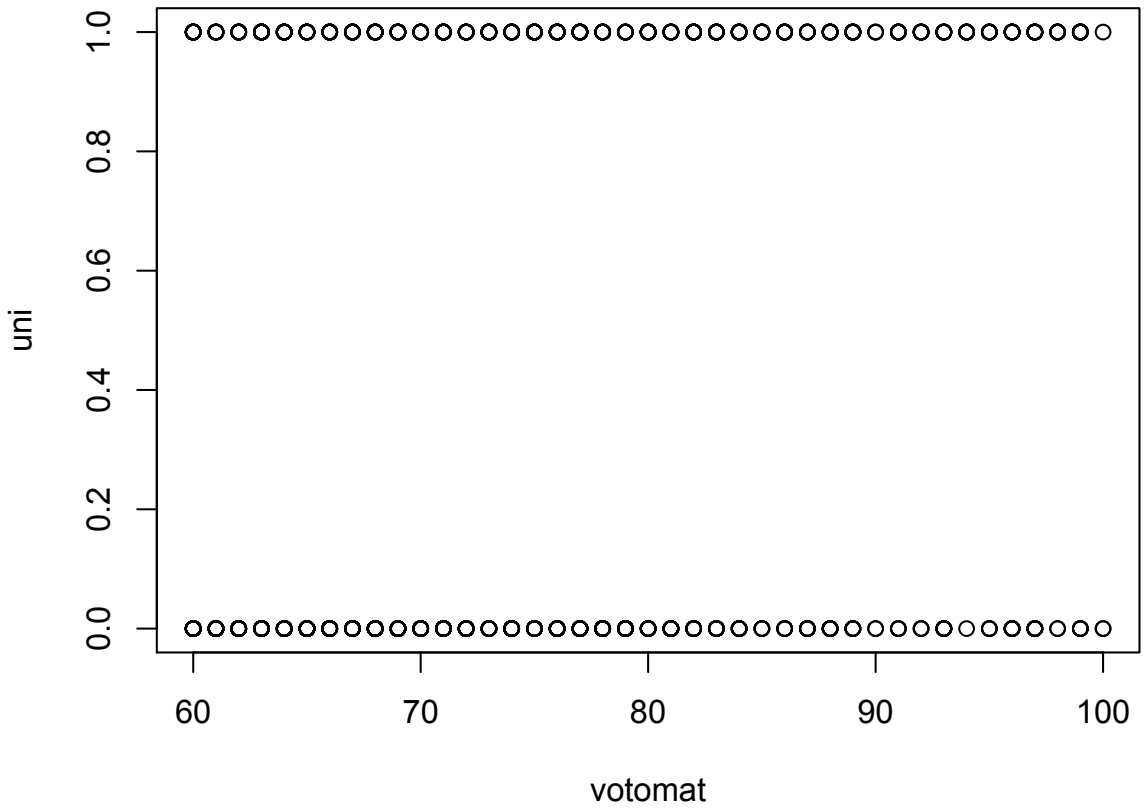
Chi si iscrive all'università?

- 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

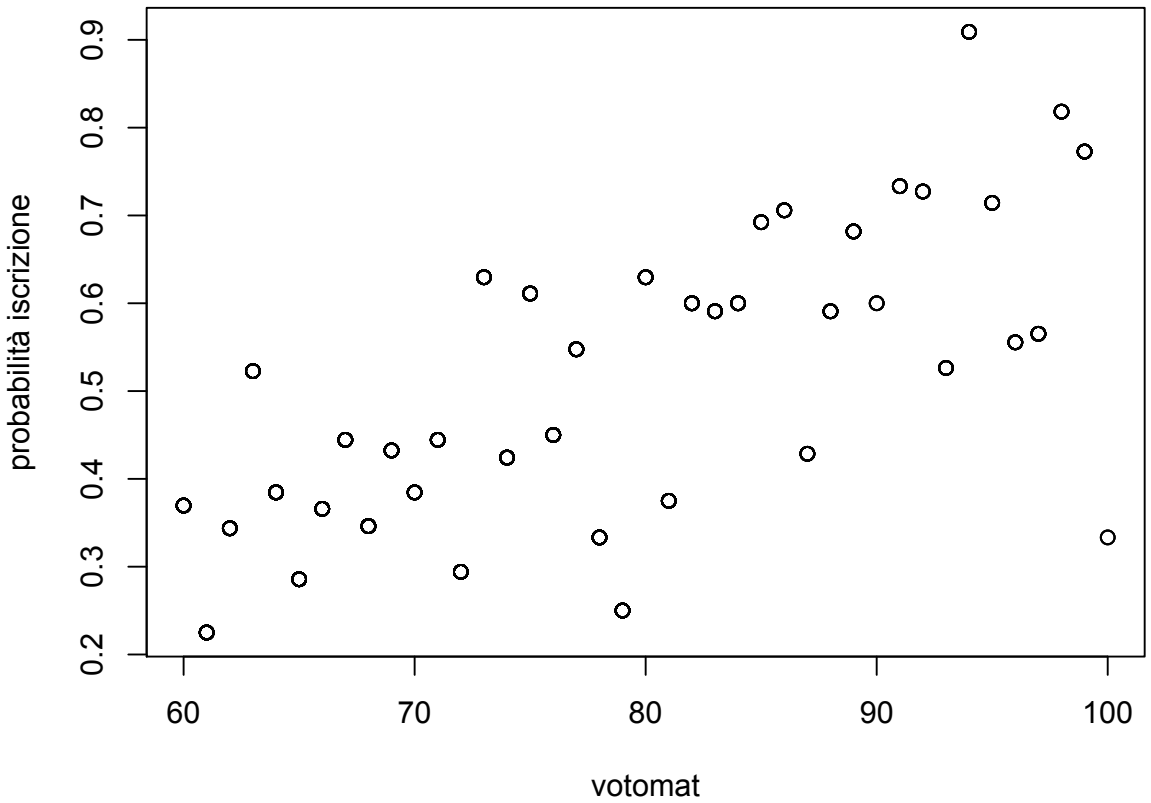
- un fattore rilevante per la probabilità di iscriversi all'università è il voto alla maturità
- la rappresentazione di questi dati chiarisce che si tratta di dati particolari

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*

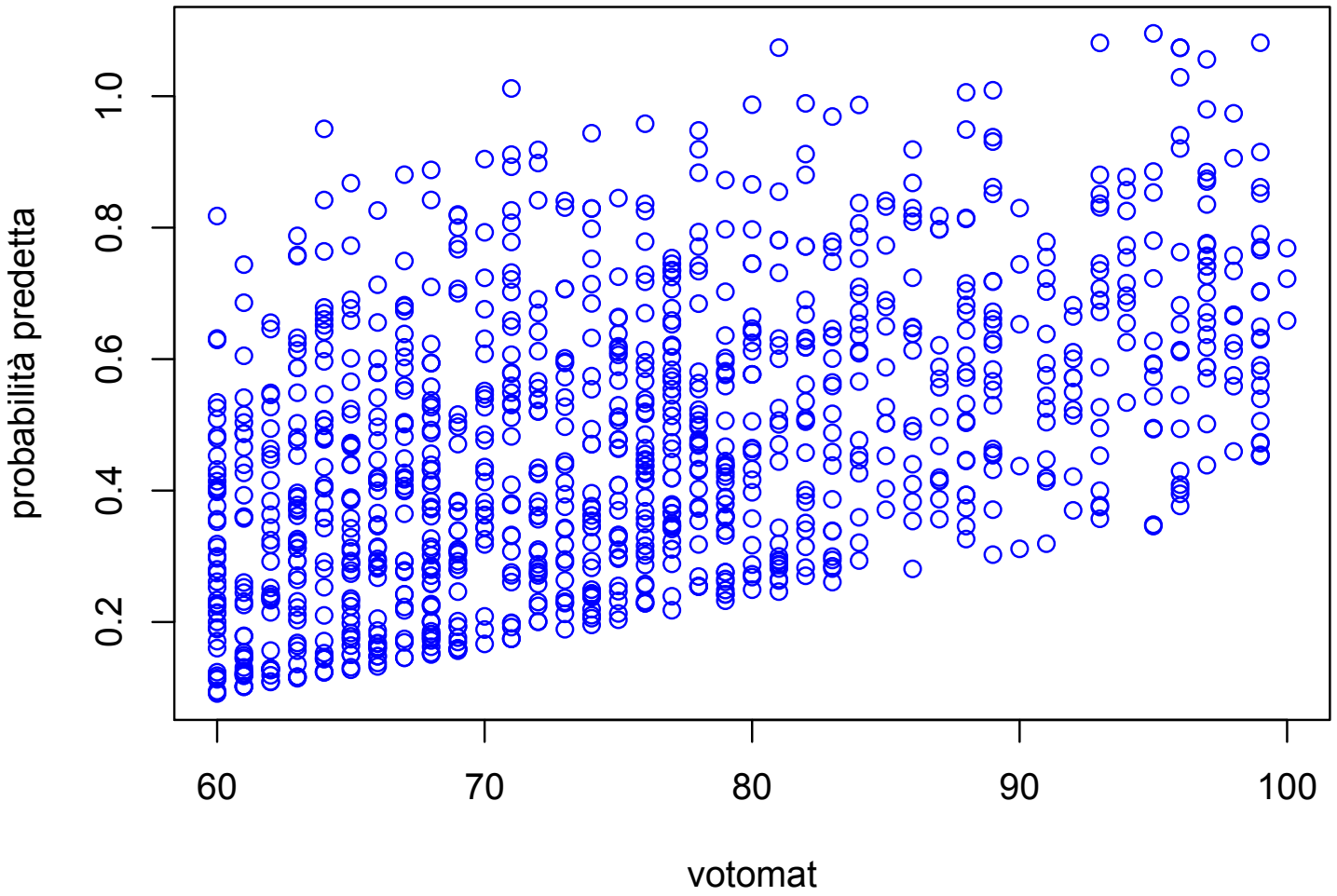
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.2433064  0.1196088  -2.034 0.042174 *
votomat                   0.0072334  0.0013297   5.440 6.58e-08 ***
donna                     0.0184420  0.0298975   0.617 0.537468
as.factor(area_istat)2   -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.