Econometria | 2023/2024

Lezione 10: Modelli per probabilità

Giuseppe Ragusa

https://gragusa.org

Roma, marzo 2023

Sommario

Variabili dipendenti binarie: qual è la differenza?

Variabile dipendente (Y) continua:

- punteggio nei test medio a livello di tutto il distretto
- tasso di mortalità stradale

Che cosa succede se Y è binaria?

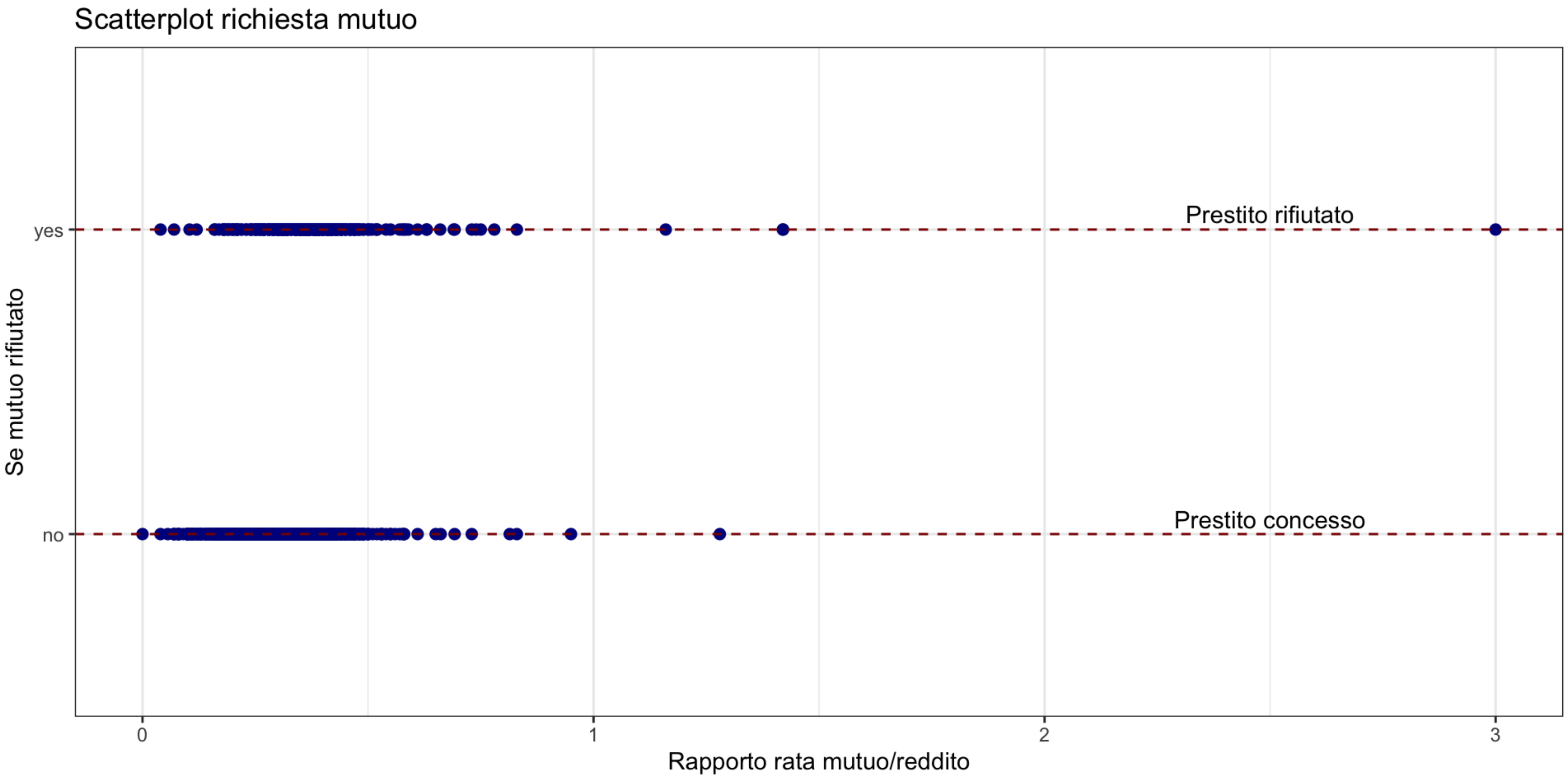
- $Y =$ frequentare università oppure no; $X =$ voti del liceo, punteggi SAT, variabili demografiche
- $Y =$ la persona fuma oppure no; $X =$ imposte sulle sigarette, reddito, variabili demografiche
- $Y =$ domanda di mutuo accettata oppure no; $X =$ etnicity, reddito, caratteristiche della casa, stato civile

Esempio: Boston Fed HMDA

- Domande individuali per mutui unifamiliari effettuate nel 1990 nell'area della città di Boston
- 2380 osservazioni, raccolte ai sensi della legge Home Mortgage Disclosure Act (HMDA)

Variabili

- Variabile dipendente:
  - il mutuo è concesso o negato?
- Variabili indipendenti:
  - reddito, ricchezza, stato occupazionale
  - altro prestito, caratteristiche della proprietà
  - etnia del richiedente



Cosa significa adattare una retta di regressione ad una variabile dipendente che può assumere solo valori zero e uno?

Modello lineare di probabilità (Paragrafo 11.1)

Consideriamo modello di regressione lineare con un singolo regressore e Y binaria:

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

Interpretazione di:

- β_1
- $\beta_0 + \beta_1 X_i$
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Il modello lineare di probabilità (continua)

Modello lineare di probabilità:

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

$$E(Y|X) = 1 \times \Pr(Y = 1|X) + 0 \times \Pr(Y = 0|X) = \Pr(Y = 1|X)$$

Sotto l'assunzione $E(u_i|X_i) = 0$:

$$E(Y_i|X_i) = E(\beta_0 + \beta_1 X_i + u_i|X_i) = \beta_0 + \beta_1 X_i$$

$$\implies \Pr(Y = 1|X) = \beta_0 + \beta_1 X_i$$

Il modello lineare di probabilità (continua)

Il modello lineare di regressione

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

è chiamato modello lineare di probabilità quando Y è binaria perchè

$$Pr(Y = 1|X) = \beta_0 + \beta_1 X_i$$

- Implicazioni:
  - $E(Y|X = x) = Pr(Y = 1|X = x)$: prob. che $Y = 1$ data X
  - $\beta_1 =$ variazione della probabilità che $Y = 1$ per una variazione unitaria in X, o più generalmente

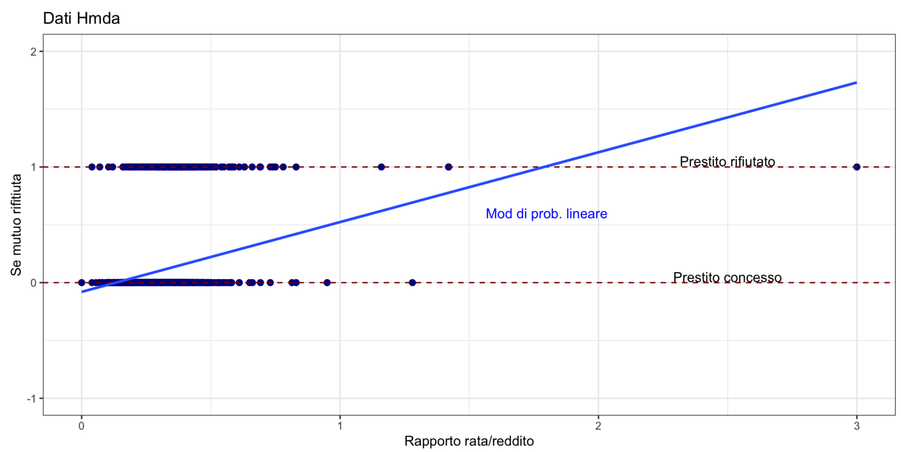
$$\beta_1 = \frac{Pr(Y = 1|X = x + \Delta x) - Pr(Y = 1|X = x)}{\Delta x}$$

  - $\hat{Y} =$ probabilità attesa stimata che $Y_i = 1$, data X

Esempio: modello lineare di probabilità, dati HMDA

```r
Hmda <- Hmda %>%
  mutate(denied = ifelse(deny=="yes", 1, 0))

ggplot(Hmda, aes(x = dir, y = denied)) +
  geom_point(size=2, col = 'darkblue') +
  theme_bw() +
  labs(x = "Rapporto rata/reddito",
       y = "Se mutuo rifitiuta",
       title = "Dati Hmda") +
  ylim(c(-1,2)) +
  geom_hline(yintercept=1, lty = 2,
             col = "darkred") +
  geom_hline(yintercept=0, lty = 2,
             col = "darkred")  +
  annotate("text", x = 2.5, y = 1.05,
           label = "Prestito rifiutato") +
  annotate("text", x = 2.5, y = 0.05,
           label = "Prestito concesso") +
  geom_smooth(method='lm',
              formula= y~x, se=FALSE) +
  annotate("text", x = 1.8, y = 0.6,
           label = "Mod di prob. lineare",
           col="blue")
```



Modello lineare di probabilità: Hmda

```r
lpm1 <- feols(denied ~ dir, data=Hmda, vcov = "hetero")
lpm1
```

```
OLS estimation, Dep. Var.: denied
Observations: 2,381
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)    -0.080     0.0320   -2.50 1.2436e-02 *
dir             0.604     0.0985    6.13 1.0353e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.318093   Adj. R2: 0.039332
```

- Probabilità stimata quando $dir = 0.3$ (rata pari al 30% del reddito)?

$$\Pr(deny = 1|dir = 0.3) = -0.08 + 0.604 \times 0.3 = 0.101 = 10.1\%$$

- Probabilità stimata quando $dir = 0.4$ (rata pari al 40% del reddito)?

$$\Pr(deny = 1|dir = 0.4) = -0.08 + 0.604 \times 0.4 = 0.162 = 16.2\%$$

- L'effetto di un aumento del rapporto mutuo/reddito, dir, di 0.1 è un aumento della probabilità che il muuo venga rigettato di 0.061 (o 6.1 punti percentuali)

Modello lineare di probabilità: Hmda, ctd.

```r
lpm2 <- feols(denied ~ dir+black, data=Hmda, vcov = "hetero")
lpm2
```

```
OLS estimation, Dep. Var.: denied
Observations: 2,381
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)  -0.0906     0.0286   -3.17 1.5626e-03 **
dir           0.5592     0.0887    6.31 3.3839e-10 ***
blackyes      0.1775     0.0249    7.11 1.4801e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.312025   Adj. R2: 0.075244
```

- Probabilità stimata quando $dir = 0.3$ (rata pari al 30% del reddito) e $black = 1$

$$Pr(denied = 1|dir = 0.3, black = 1) = -0.091 + 0.559 \times 0.3 + 0.177 \times 1 = 0.254$$

- Probabilità stimata quando $dir = 0.4$ (rata pari al 40% del reddito) e $black = 1$

$$Pr(denied = 1|dir = 0.3, black = 1) = -0.091 + 0.559 \times 0.3 + 0.177 \times 0 = 0.077$$

- differenza $0.254 - 0.077 = 0.177 = 17.7$ punti percentuali

Modello lineare di probabilità: riepilogo

Nel modello lineare di probabilità, $\Pr(Y = 1|X)$ è lineare in X

- Vantaggi:
  - semplice da stimare e interpretare
  - l'inferenza è la stessa della regressione multipla (occorrono errori standard robusti all'eteroschedasticità)
- Svantaggi:
  - Le probabilità previste possono essere < 0 o > 1!
  - Implicazioni del modello lineare per la probabilità che $Y = 1$ poco sensate
- Alternativa: modello non lineare di probabilità: probit e logit

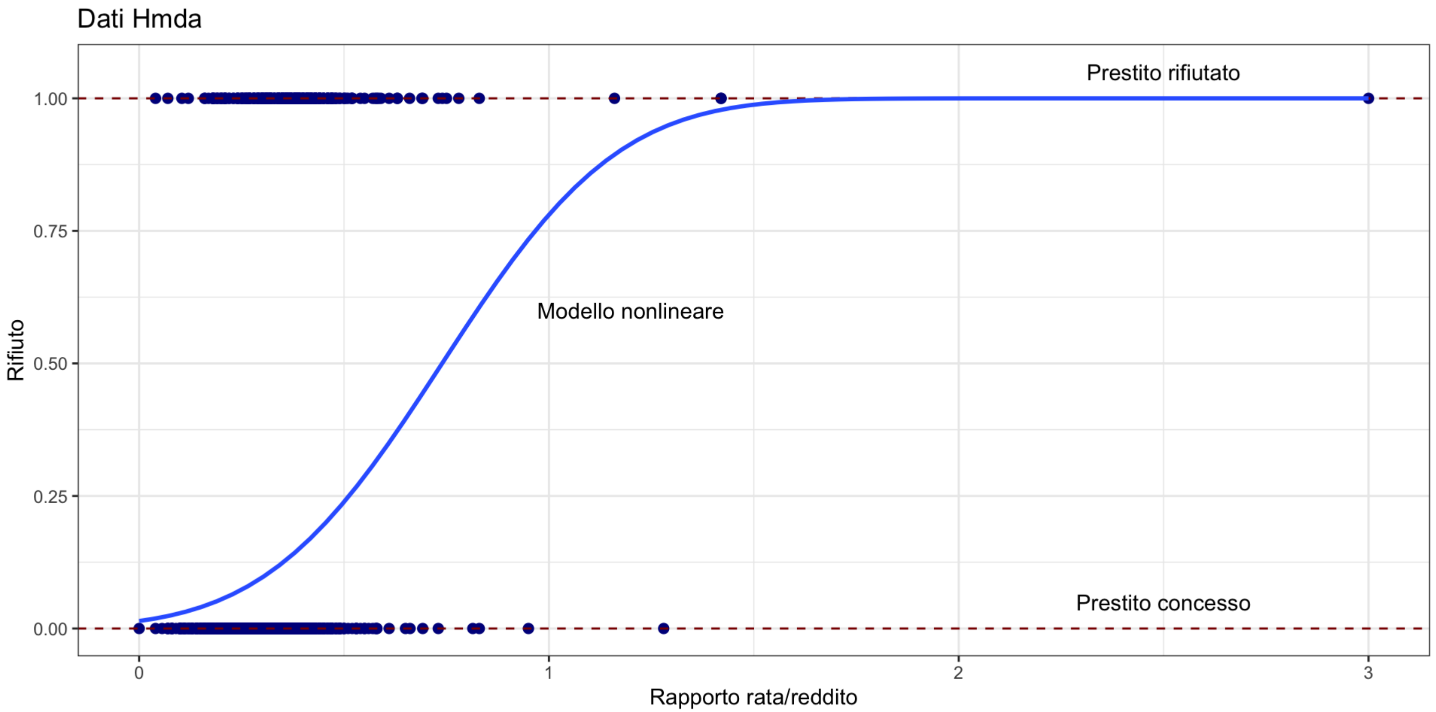
Regressioni probit e logit (Paragrafo 11.2)

Desiderata:

a. $Pr(Y=1|X)$ crescente in X per $\beta_1 > 0$ e

b. $0 \leq Pr(Y=1|X) \leq 1$ per tutte le X

Ciò richiede l'utilizzo di una forma funzionale non lineare per la probabilità



Regressione probit

- Il modello `probit` è:

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$

dove $\Phi(\cdot)$ è la funzione di ripartizione normale

Esempio:

Se $\beta_0 = -2$, $\beta_1 = 3$, and $X = 0.4$, allora:

$$\Pr(Y = 1|X = 0.4) = \Phi(-2 + 3 \times 0.4) = \Phi(-0.8) = 0.212$$

Regressione probit (continua)

- La funzione di ripartizione normale ha una forma a S e, quindi,
  a. $Pr(Y = 1|X)$ crescente in X per $\beta_1 > 0$ e
  b. $0 \leq Pr(Y = 1|X) \leq 1$ per tutte le X
- Legame con modelli economici di scelta ottima
- Interpretazione di β_1

$$\frac{d\,\mathrm{Pr}(Y = 1|X)}{dX} = \frac{d\Phi(\beta_0 + \beta_1 X)}{dX} = \phi(\beta_0 + \beta_1 X)\beta_1$$

β_1 determina il segno dell'effetto di una variazione di X su $\mathrm{Pr}(Y = 1|X)$

Esempio in R: dati HMDA

```r
probit1 <- feglm(denied ~ dir, family = binomial("probit"), data=Hmda)
probit1
```

```
GLM estimation, family = binomial, Dep. Var.: denied
Observations: 2,381
Standard-errors: IID
            Estimate Std. Error z value   Pr(>|z|)
(Intercept)    -2.19      0.138  -15.93  < 2.2e-16 ***
dir             2.97      0.386    7.69 1.4442e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Log-Likelihood:  -831.9   Adj. Pseudo R2: 0.045059
           BIC: 1,679.4     Squared Cor.: 0.051522
```

$$\Pr(deny = 1|dir) = \Phi(-2.19 + 2.97 \times dir)$$

- Coefficiente positivo: maggiore dir maggiore è la probabilità che il mutuo sia rifiutato

Esempio in R: Hmda, ctd.

```
GLM estimation, family = binomial, Dep. Var.: denied
Observations: 2,381
Standard-errors: IID
            Estimate Std. Error z value   Pr(>|z|)
(Intercept)    -2.19      0.138  -15.93  < 2.2e-16 ***
dir             2.97      0.386    7.69 1.4442e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Log-Likelihood:  -831.9   Adj. Pseudo R2: 0.045059
           BIC: 1,679.4     Squared Cor.: 0.051522
```

- Probabilità prevista quando $dir = 0.3$:

$$\Pr(deny = 1|dir = 0.3) = \Phi(-2.19 + 2.97 \times 0.3) = \Phi(-1.30) = 0.097$$

- Probabilità stimate quando $dir = 0.4$:

$$\Pr(deny = 1|dir = 0.4) = \Phi(-2.19 + 2.97 \times 0.4) = \Phi(-1.00) = 0.159$$

La probabilità prevista di rifiuto passa da 0.097 a 0.159

Regressione probit con regressori multipli

$$Pr(Y = 1|X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

- $\Phi(\cdot)$ è la funzione di ripartizione normale
- Interpretazione coefficienti:

$$\frac{\partial \Pr(Y = 1|X_1, X_2)}{\partial X_1} = \frac{d\Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{dX} = \phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)\beta_1$$

β_1 cattura effetto ceteris paribus sul segno

Esempio in R: Dati HMDA

```
GLM estimation, family = binomial, Dep. Var.: denied
Observations: 2,381
Standard-errors: IID
             Estimate Std. Error z value   Pr(>|z|)
(Intercept)   -2.259     0.1368   -16.5  < 2.2e-16 ***
dir            2.742     0.3807     7.2 5.8682e-13 ***
blackyes       0.708     0.0834     8.5  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Log-Likelihood:  -797.2   Adj. Pseudo R2: 0.083677
           BIC: 1,617.8     Squared Cor.: 0.087201
```

- Il coefficiente di $black$ è statisticamente significativo?

$$\begin{aligned}\Pr(deny = 1|dir = 0.3, black = 1) &= \Phi(-2.26 + 2.74 \times 0.3 + 0.71 \times 1)\\ &= \Phi(-0.728) = 0.233\end{aligned}$$

$$\begin{aligned}\Pr(deny = 1|dir = 0.3, black = 0) &= \Phi(-2.26 + 2.74 \times 0.3 + 0.71 \times 0)\\ &= \Phi(-1.438) = 0.075\end{aligned}$$

- Differenza nelle probabilità di rifiuto= 0.158 (15.8 punti percentuali)
- Ancora molto spazio per distorsione da variabili omesse!

Esempio in R (continua): probabilità probit previste

```r
## Differenza probability black/nonblack per dir=0.3
predict(probit2, newdata=data.frame(dir=0.3, black="yes"), type = "response") -
  predict(probit2, newdata=data.frame(dir=0.3, black="no"), type = "response")
```

```
[1] 0.158
```

Effetti marginali

Valori diversi di dir danno valori diversi della variazione di $\Pr(Y = 1|X)$...

Two approaches:

- Calcolare l'effetto per un valore rappresentativo di dir, e.g., $\bar{dir}$

```r
## Differenza probability black/nonblack per dir=media(dir)
dirbar <- mean(Hmda$dir)
predict(probit2, newdata=data.frame(dir=dirbar, black="yes"), type = "response") -
  predict(probit2, newdata=data.frame(dir=dirbar, black="no"), type = "response")
```

```
[1] 0.172
```

- Calcolare gli effetto per i valori $dir_i, i = 1, \ldots, n$ e calcolare la media

```r
## Differenza probability black/nonblack per dir=dir_i
effetti <- predict(probit2, newdata=data.frame(dir=Hmda$dir, black="yes"), type = "response") -
  predict(probit2, newdata=data.frame(dir=Hmda$dir, black="no"), type = "response")
mean(effetti)
```

```
[1] 0.17
```

Effetti marginali

```r
library(marginaleffects)
## Valuta effetti alla media (avg) per e calcola standard error
avg_slopes(probit2, newdata = "mean")
```

```
  Term Contrast Estimate Std. Error    z Pr(>|z|)    S 2.5 % 97.5 %
 black yes - no    0.172     0.0247 6.95   <0.001 38.0 0.123  0.220
 dir   dY/dX       0.439     0.0625 7.02   <0.001 38.7 0.316  0.561

Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high 
Type:  response 
```

```r
## Valuta la media degli effetti (AME) e calcola standard error
avg_slopes(probit2)
```

```
  Term Contrast Estimate Std. Error    z Pr(>|z|)    S 2.5 % 97.5 %
 black yes - no    0.170     0.0243 6.99   <0.001 38.5 0.122  0.217
 dir   dY/dX       0.501     0.0690 7.26   <0.001 41.3 0.366  0.637

Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high 
Type:  response 
```

Regressione logit

- Modella la probabilità di $Y = 1$, data X, come funzione di ripartizione logistica standard, valutata in $z = \beta_0 + \beta_1 X$:

$$Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

dove F è la funzione di ripartizione logistica:

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Poiché le regressioni logit e probit utilizzano funzioni di probabilità diverse, i coefficienti (β) sono diversi nelle regressioni logit e probit.

Regressione logit (continua)

$$Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

dove:

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Esempio: $\beta_0 = \mathbf{3},\ \beta_1 = \mathbf{2},\ X = 0.4$

- quindi $\beta_0 + \beta_1 X = -3 + 2 \times 0.4 = -2.2$
- perciò $Pr(Y = 1|X = 0.4) = \frac{1}{1+e^{-(-2.2)}} = 0.0998$

A che scopo utilizzare logit se disponiamo di probit?

- Il motivo principale è storico: dal punto di vista del calcolo, logit è più veloce
- `logit` e `probit` danno risultati spesso simili

Esempio in R: Dati HMDA

```r
logit2 <- feglm(denied ~ dir+black, family = binomial("logit"), data=Hmda)
logit2
```

```
GLM estimation, family = binomial, Dep. Var.: denied
Observations: 2,381
Standard-errors: IID
            Estimate Std. Error z value   Pr(>|z|)
(Intercept)    -4.13      0.269  -15.37  < 2.2e-16 ***
dir             5.37      0.729    7.37 1.6925e-13 ***
blackyes        1.27      0.146    8.71  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Log-Likelihood:  -795.8   Adj. Pseudo R2: 0.085331
           BIC: 1,614.9     Squared Cor.: 0.088796
```

```r
predict(logit2, newdata= data.frame(dir=0.3, black="no"), type = "response")
```

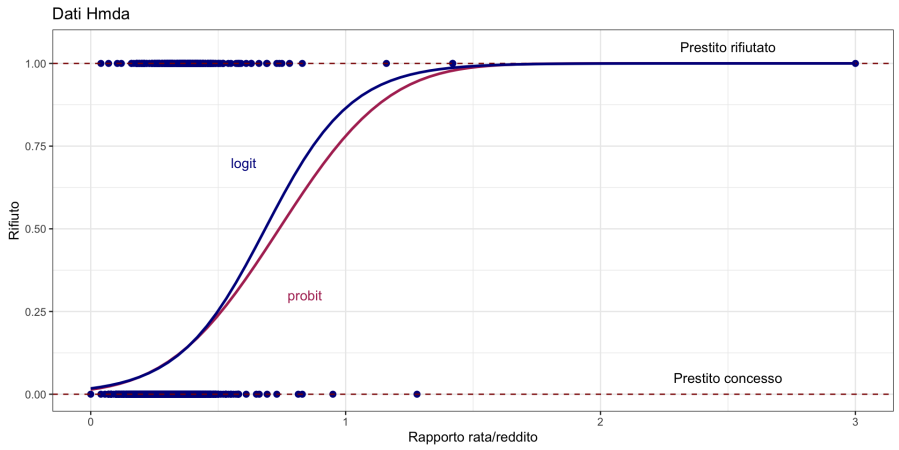
```
[1] 0.0748
```

```r
predict(probit2, newdata= data.frame(dir=0.3, black="no"), type = "response")
```

```
[1] 0.0754
```

Confronto grafico probit e logit

```r
ggplot(Hmda, aes(x = dir, y = denied)) +
  geom_point(size=2, col = 'darkblue') +
  theme_bw() +
  labs(x = "Rapporto rata/reddito") +
  labs(y = "Rifiuto", title = "Dati Hmda") +
  geom_hline(yintercept=0, lty = 2,
             col = "darkred") +
  geom_hline(yintercept=1, lty = 2,
             col = "darkred") +
  annotate("text", x = 2.5, y = 1.05,
           label = "Prestito rifiutato") +
  annotate("text", x = 2.5, y = 0.05,
           label = "Prestito concesso") +
  stat_smooth(method="glm", se=FALSE,
              col = "maroon",
              method.args = list(family=
               binomial(link = "probit"))) +
  stat_smooth(method="glm", se=FALSE,
              col = "darkblue",
              method.args = list(family=
               binomial(link = "logit"))) +
  annotate("text", x = .6, y = 0.7,
           color = "darkblue",
           label = "logit") +
  annotate("text", x = 0.84, y = 0.3,
           color = "maroon",
           label = "probit")
```



Stima e inferenza nei modelli logit e probit (Paragrafo 11.3)

Ci concentriamo sul modello probit:

$$Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$

- Stima e inferenza
  - Come possiamo stimare β_0 e β_1?
  - Qual è la distribuzione campionaria degli stimatori?
  - Perché possiamo utilizzare i metodi usuali di inferenza?
- Anzitutto motiviamo tramite i minimi quadrati non lineari
- Quindi discutiamo della stima di massima verosimiglianza (ciò che effettivamente è fatto nella pratica)

Stima probit mediante i minimi quadrati non lineari

Ricordiamo OLS: $\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1)\right]^2$

- Il risultato sono gli stimatori OLS $\hat{\beta}_0, \hat{\beta}_1$

I minimi quadrati non lineari estendono l'idea di OLS ai modelli nei quali i parametri entrano in modo non lineare:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \left[Y_i - \Phi(\hat{\beta}_0 + \hat{\beta}_1)\right]^2$$

- Come risolvere questo problema di minimizzazione?
  - Il calcolo non offre una soluzione esplicita.
  - Risolto numericamente mediante il computer (algoritmi di minimizzazione speciali)
  - In pratica, i minimi quadrati non lineari non sono utilizzati. Uno stimatore più efficiente (varianza più piccola) è

Stimatore di massima verosimiglianza

La funzione di verosimiglianza è la distribuzione di $Y_1, \ldots, Y_n$ condizionata a $X_1, \ldots, X_n$ trattata come funzione dei parametri ignoti β_0, β_1.

- Lo stimatore di massima verosimiglianza (MLE) è il valore di (β_0, β_1) che massimizza la funzione di verosimiglianza.
- MLE sceglie i parametri per massimizzare la probabilità di estrarre i dati effettivamente osservati.
- MLE è il valore di $(\beta_0, \beta_1$, cioè i parametri che "più verosimilmente" hanno generato i dati.
- In grandi campioni, MLE è:
  - consistente
  - normalmente distribuito
  - efficiente (ha la più piccola varianza di tutti gli stimatori)

Caso speciale: MLE probit senza alcuna X

Distribuzione di Bernoulli:

$$Y = \begin{cases} 1 & \text{con probabilità } p, \\ 0 & \text{con probabilità } (1-p). \end{cases}$$

Dati: $Y_1, \ldots, Y_n$ i.i.d.

La derivazione della verosimiglianza inizia con la distribuzione di Y_1:

$$Pr(Y_1 = y_1) = p^{y_1}(1-p)^{1-y_1}$$

Distribuzione congiunta di (Y_1, Y_2): poiché Y_1 e Y_2 sono indipendenti:

$$\begin{aligned} Pr(Y_1 = y_1, Y_2 = y_2) &= Pr(Y_1 = y_1) \times Pr(Y_2 = y_2) = \\ &= \left[p^{y_1}(1-p)^{1-y_1}\right] \times \left[p^{y_2}(1-p)^{1-y_2}\right] = \\ &= p^{(y_1+y_2)}(1-p)^{2-(y_1+y_2)} \end{aligned}$$

Distribuzione congiunta di $(Y_1, \ldots, Y_n)$: poiché $Y_1, \ldots, Y_n$ sono indipendenti:

$$\begin{aligned} Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) &= Pr(Y_1 = y_1) \times Pr(Y_2 = y_2) \times \\ &\quad \cdots \times Pr(Y_n = y_n) \\ &= \left[p^{y_1}(1-p)^{1-y_1}\right] \times \left[p^{y_2}(1-p)^{1-y_2}\right] \times \\ &\quad \cdots \times \left[p^{y_n}(1-p)^{1-y_n}\right] \\ &= p^{\sum_{i=1}^n y_i}(1-p)^{\left(n-\sum_{i=1}^n y_i\right)} \end{aligned}$$

La verosimiglianza è la distribuzione congiunta, trattata come funzione dei parametri non noti, che qui è p:

$$f(p; Y_1, \ldots, Y_n) = p^{\sum_{i=1}^{n} y_i}(1-p)^{\left(n - \sum_{i=1}^{n} y_i\right)}$$

MLE massimizza la verosimiglianza. È più facile lavorare con il logaritmo della verosimiglianza, $\ln[f(p; Y_1, \ldots, Y_n)]$:

$$\ln[f(p; Y_1, \ldots, Y_n)] = \left(\sum_{i=1}^{n} y_i\right)\ln(p) + \left(n - \sum_{i=1}^{n} y_i\right)\ln(1-p)$$

Massimizzare la verosimiglianza impostando la derivata =0:

$$\frac{d\ln[f(p; Y_1, \ldots, Y_n)]}{dp} = \left(\sum_{i=1}^{n} y_i\right)\frac{1}{p} + \left(n - \sum_{i=1}^{n} y_i\right)\left(\frac{-1}{1-p}\right) = 0$$

Risolvendo per p si ottiene lo stimatore di massima verosimiglianza.

Ovvero, $\hat{p}^{MLE}$ soddisfa:

$$\left(\sum_{i=1}^{n} y_i\right) \frac{1}{\hat{p}^{MLE}} + \left(n - \sum_{i=1}^{n} y_i\right) \left(\frac{-1}{1-\hat{p}^{MLE}}\right) = 0$$

oppure:

$$\left(\sum_{i=1}^{n} y_i\right) \frac{1}{\hat{p}^{MLE}} = \left(n - \sum_{i=1}^{n} y_i\right) \left(\frac{1}{1-\hat{p}^{MLE}}\right)$$

oppure:

$$\frac{\bar{Y}}{1-\bar{Y}} = \frac{\hat{p}^{MLE}}{1-\hat{p}^{MLE}}$$

quindi:

$$\hat{p}^{MLE} = \bar{Y} = \textbf{frazione di 1}$$

MLE nel caso “No-X”, ctd,

$$\hat{p}^{MLE} = \bar{Y} = \text{frazione di 1}$$

- Per Y_i i.i.d. di Bernoulli, MLE è lo stimatore “naturale” di p, la frazione di 1, che è $\bar{Y}$
- Conosciamo già i fondamenti dell’inferenza:
  - Per n grande, la distribuzione campionaria di $\hat{p}^{MLE} = \bar{Y}$ è normale
  - Perciò l’inferenza è “come di consueto”: verifica di ipotesi tramite statistica t, intervallo di confidenza come $\pm 1.96 SE$

MLE nel caso “No-X” (distribuzione di Bernoulli), continua:

- La teoria della stima di massima verosimiglianza dice che è lo stimatore più efficiente di p - di tutti i possibili stimatori! - almeno per n grande (molto più forte del teorema di Gauss-Markov).
- Per questo motivo MLE è il principale stimatore utilizzato per i modelli nei quali i parametri (coefficienti) entrano in modo non lineare.

Siamo ora pronti a passarea MLE dei coefficienti probit, nel quale la probabilità è condizionata a X.

La verosimiglianza probit con una X

La derivazione inizia con la distribuzione di Y_1, data X_1 è:

$$Pr(Y_1 = 1|X_1) = \Phi(\beta_0 + \beta_1 X_1)$$
$$Pr(Y_1 = 0|X_1) = 1 - \Phi(\beta_0 + \beta_1 X_1)$$

quindi:

$$Pr(Y_1 = y_1|X_1) = \Phi(\beta_0 + \beta_1 X_1)^{y_1}[1 - \Phi(\beta_0 + \beta_1 X_1)]^{1-y_1}$$

La funzione di verosimiglianza probit è la distribuzione congiunta di $Y_1, \ldots, Y_n$ dati $X_1, \ldots, X_n$, come funzione di β_0, β_1

$$\begin{aligned} f(\beta_0, \beta_1; Y_1, \ldots, Y_n|X_1, \ldots, X_n) = \Phi(\beta_0 + \beta_1 X_1)^{y_1}[1 - \Phi(\beta_0 + \beta_1 X_1)]^{1-y_1} \times \\ \cdots \times \Phi(\beta_0 + \beta_1 X_n)^{y_n}[1 - \Phi(\beta_0 + \beta_1 X_n)]^{1-y_n} \end{aligned}$$

La funzione di verosimiglianza probit

$$f(\beta_0, \beta_1; Y_1, \ldots, Y_n | X_1, \ldots, X_n) = \Phi(\beta_0 + \beta_1 X_1)^{y_1} [1 - \Phi(\beta_0 + \beta_1 X_1)]^{1-y_1} \times \\ \cdots \times \Phi(\beta_0 + \beta_1 X_n)^{y_n} [1 - \Phi(\beta_0 + \beta_1 X_n)]^{1-y_n}$$

- $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ massimizzano questa funzione di verosimiglianza
- Ma non possiamo risolvere esplicitamente per il massimo! Così MLE deve essere massimizzato mediante metodi numerici
- Come nel caso di nessuna X, in grandi campioni:
  - $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ sono consistenti e normalmente distribuiti e asintoticamente efficienti
- Gli errori standard di $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ vengono calcolati automaticamente
- La verifica degli intervalli di confidenza procede nel modo consueto
- Tutto ciò si estende a X multiple, per i dettagli si veda l'Appendice 11.2

La verosimiglianza logit con una X

- L'unica differenza tra probit e logit è a forma condizionale utilizzata per la probabilità: Φ è sostituita dalla funzione di ripartizione logistica
- Altrimenti, la verosimiglianza è simile; per i dettagli si veda l'Appendice 11.2
- Come nel caso probit:
  - $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ sono consistenti
  - $\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}$ sono normalmente distribuiti
  - I loro errori standard possono essere calcolati
  - La verifica degli intervalli di confidenza procede nel modo consueto

Conclusione

- Se Y_i è binaria, allora $E(Y|X) = Pr(Y = 1|X)$
- Tre modelli:
  - modelli lineare di probabilità (regressione lineare multipla)
  - probit (funzione di ripartizione normale standard)
  - logit (funzione di ripartizione logistica standard)
- LPM, probit, logit tutti producono probabilità predette
- L'effetto di ΔX è la variazione nella probabilità condizionata che $Y = 1$. Per logit e probit, ciò dipende dalla X iniziale
- Probit e logit vengono stimati tramite la massima verosimiglianza
  - I coefficienti hanno distribuzione normale per n grande
  - La verifica di ipotesi e gli intervalli di confidenza per ngrande sono come di consueto