Variáveis Binárias em R

Bruno Holanda *

Resumo

O objetivo desta nota é fornecer um rotina em R para reproduzir uma análise de escolha entre duas marcas de ketchup que utiliza modelos de variáveis independentes binárias. Este estudo está melhor detalhado no quarto capítulo do livro de Franses and Paap (2001).

Palavras chave: Variável Binária · Probit · Logit · R

1 Introdução

Modelos de variável binária são necessários quando a variável independente do modelo só pode assumir um entre dois possíveis valores que são geralmente normalizados como constantes 0 ou 1. Por exemplo, um pesquisador pode estar interessado em explicar como o nível educacional e a experiência laboral afetam a condição de o indivívio estar ou não empregado.

Neste tipo de modelo, não podemos assumir a hipótese de linearieadade. De fato, ao assumirmos uma regressão do tipo

$$Y|X = X\beta + u$$

o lado esquerdo pode facilmente gerar valores fora do conjunto $\{0, 1\}$. Ao invés disso, podemos utilizar a regressão:

$$Prob(Y = 1|X) = F(X\beta),$$

onde $F(\cdot)$ representa alguma função de probabilidade acumulada. Com essa abordagem, garantimos que os dois lados da equação sejam compatíveis. Em geral, escolhemos F como sendo uma das seguintes alternativas:

*Professor Adjunto, Universidade Federal de Goiás, Faculdade de Administração, Contabilidade e Economia (FACE-UFG). e-mail: bholanda@ufg.br

(i) **Função Logística:**

$$F(x) = \Lambda(x) = \frac{\exp(x)}{1 + \exp(x)}$$

(ii) **Função Normal:**

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \phi(z)dz,$$

onde $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$ é a distribuição normal.

Na primeira opção, chamamos o modelo de regressão logística. Na segunda, de modelo probit.

Para estes modelos, não faz sentido utilizar o método de mínimos quadrados para estimar os parâmetros β. Utilizaremos, portanto, o método de máxima verossimilhança. Além disso, diferentemente do que ocorre no modelo linear, não é possível achar analiticamente o ponto que maximiza a função de verossimilhança para este modelo. Para obter os pontos críticos, devemos recorrer a algum método computacional interativo. Em geral, os pacotes estatíticos utilizam o método de Newton–Raphson.

Para medir a qualidade de ajuste (goodness of fit) dos modelos estimados, utilizaremos as medidas de Pseudo-R^2_{MF} de McFadden e R^2-contado.

O R^2_{MF} de McFadden é defino como

$$R^2_{MF} = 1 - \frac{\log(L_{UR})}{\log(L_R)},$$

onde L_{UR} é o máximo da função de verossimilhança da regressão não restrita e L_R é o máximo da função de verossimilhança da regressão restrita quando $\beta_1 = \beta_2 = \cdots = \beta_k$.

Para calcular o R^2-contado temos que inicialmente considerar os valores estimados das variáveis explicadas definidos pela seguinte regra:

$$Se \begin{cases} F(X_i\hat{\beta}) > 0.5 \Rightarrow \hat{y}_i = 1 \\ F(X_i\hat{\beta}) \leq 0.5 \Rightarrow \hat{y}_i = 0 \end{cases}$$

Então,

$$R^2_{contado} = \frac{\#\{i|\hat{y}_i = y_i\}}{n}.$$

Ou seja, é o número de previssões corretas dividido pelo número de observações. Em geral, dizemos que o modelo tem um bom ajuste se $R^2_{contado} > 0.75$.

Além disso, os coeficientes β's não mostram uma relação direta dos efeitos marginais das variáveis explicativas sob a variável explicada. Para calcular os efeitos

marginais derivamos a função $p = Prob(Y|X)$ parcialmente em relação à variável x_i. Através da regra da cadeia obtemos:

$$\frac{\partial p}{\partial x_i} = F'(X\beta)\beta_i.$$

Portanto, o efeito marginal sob a probabilidade de sucesso da variável explicada ($prob(y = 1|X)$) devido a uma mudança marginal da variável explicativa x_i depende dos valores fixados para todas as variáveis explicativas estatisticamente relevantes. Dessa forma, algums pesquisadores preferem reportar a chamada média dos efeitos amostrais definida por

$$\overline{\left(\frac{\partial p}{\partial x_i}\right)} := \frac{1}{n}\sum_{i=1}^{n} F'(X\hat{\beta})\hat{\beta}_i$$

2 Os Dados

Os dados correspondem a um total de 2798 observações sobre a decisão de 300 indivíduos sobre a escolha entre duas marcas de ketchup: Heinz e Hunts. Os dados podem ser facilmente obtidos no site people.few.eur.nl. As variáveis coletadas são:

"OBS": indica o número da observação.
"HOUSEHOLDID": indicador dos indivíduos.
"LASTPURCHASE": indica se é a última compra do indivíduo.
"HEINZ": indica se escolha foi pela marca Heinz.
"HUNTS": indica se escolha foi pela marca Hunts.
"PRICEHEINZ": preço da marca Heinz.
"PRICEHUNTS": preço da marca Hunts.
"DISPLHEINZ": marca presença de display da marca Heinz
"DISPLHUNTS": marca presença de display da marca Huntz
"FEATHEINZ": marca presença de "features"da marca Heinz
"FEATHUNTS": marca presença de "features"da marca Hunts
"FEATDISPLHEINZ": marca presença de display e "features"da marca Heinz
"FEATDISPLHUNTS": marca presença de display e "features"da marca Hunts

O modelo a ser estimado é

$$Prob(Y|X) = F(\beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7 + u),$$

onde $Y = HEINZ$, $X_1 = \log\left(\frac{PRICEHEINZ}{PRICEHUNTS}\right)$ e X_2,...,X_7 representam as variáveis DISPLHEINZ, DISPLHUNTS, FEATHEINZ, FEATHUNTS, FEATDISPLHEINZ, FEATDISPLHUNTS.

3 Rotina em R

Antes de elaborarmos qualquer rotina em R devemos instalar os pacotes que serão utilizados em nossa análise. No caso particular desta nota, devem estar instalados os pacotes `readxl` para ler dados no formato xlsx e `lmtest` para realizar testes estatíticos. O próximo passo é indicar que os pacotes serão utilizados. Faça isso através dos comandos:

```
library(readxl) #para ler dados em excel
library(lmtest) #para testes estatísticos
library(stargazer) #para criar tabelas em LaTeX
```

Agora iremos ler o banco de dados[1] e separá-los em dois. No primeiro, guarde todas as observações de todos os indivíduos, retirando a última observação de cada um. Chame este banco de *subd* e guarde as observações não utilizadas em *subd*2. Utilize o banco de dados *subd*.

```
d <- read_excel("seu diretorio/paap4.xlsx",
                sheet = 1, col_names=TRUE)

subd <- subset(d, LASTPURCHASE==0)
subd2 <- subset(d, LASTPURCHASE==1)

attach(subd)
```

Agora já podemos rodar nosso modelo de variável binária. Faremos um para regressão logit e outra probit.

```
Y <- HEINZ
logPHEPHU <- log(PRICEHEINZ/PRICEHUNTS)
X <- cbind(logPHEPHU , DISPLHEINZ , DISPLHUNTS ,
      FEATHEINZ , FEATHUNTS , FEATDISPLHEINZ , FEATDISPLHUNTS)
###############

logit<- glm(Y ~ X, family=binomial (link = "logit"))

summary(logit)

MX <- cbind((rep(1,length(Y))),X)
```

[1] Disponível em repositório github

```
ZZ <- MX %*% coefficients(logit)

probit <- glm(Y ~ X,family=binomial (link = "probit"))

## Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred

summary(probit)
```

As funções **summary** criam uma tabela de resultados como mostramos no apêndice. Como os resultados foram próximos, iremos nos dedicar de agora em diante ao modelo logit. Para calcular o R^2 de McFadden, faça:

```
logit0<-update(logit, formula= Y ~ 1)
McFadden<- 1-as.vector(logLik(logit)/logLik(logit0))
```

O resultado é 0.301 que nos parece razoável para um painel com tantas observações. Para a medida R^2-contado, faremos uma tabela Booleana que verifica quando os valores estimados para a variável independente são iguais aos valores apresentados nos dados.

```
Y <- HEINZ
table(true = Y, pred = round(fitted(logit)))

##     pred
## true    0    1
##    0   71  201
##    1   28 2198
```

Assim,

$$R^2_{contado} = \frac{2198 + 71}{2498} = 0.908$$

Agora verificaremos se o modelo é capaz de prever razoavelmente as observações colecionadas em *subd*2 que corresponde a última compra para cada indivíduo.

```
attach(subd2)
Y <- HEINZ
X <- cbind(log(PRICEHEINZ/PRICEHUNTS) , DISPLHEINZ , DISPLHUNTS ,
           FEATHEINZ , FEATHUNTS , FEATDISPLHEINZ , FEATDISPLHUNTS)
MX <- cbind((rep(1,length(Y))),X)
Z <- MX %*% coefficients(logit)
fitZ <- round(exp(Z)/(1+exp(Z)))
table(true = Y, pred = fitZ)
```

```
##     pred
## true   0   1
##    0   4  31
##    1   1 264
```

Neste caso,

$$R^2_{contado} = \frac{264+4}{300} = 0.893$$

que corresponde a uma boa capacidade preditiva.

Para finalizar, vamos calcular a média dos efeitos marginais amostrais para a variável X_1. Mas antes vamos construir manualmente a função logística:

```
lal <- function(x) {exp(x)/(1+exp(x))}
```

Agora utilizaremos o vetor ZZ criado anteriormente:

```
MarEff <- mean(lal(ZZ)*(1-lal(ZZ)))*coefficients(logit)[2]
```

O valor encontrado é -0.426. Isso significa que o aumento de 1 ponto na variável X_1 diminui em (em média) -42.6% a probabilidade do consumidor optar pela marca Heinz. Portanto, temos um mercado que é muito sensível à uma mudança de preços relativos.

4 Observações

Além de Franses and Paap (2001), estas notas também foram inspiradas em Maddala (2001) e Johnston et al. (1997). A programação em R foi baseada nos exemplos encontrados no site Econometric Academy mantido por Ani Katchova. Para aprender mais sobre esta linguagem computacional, direcionamos o leitor ao repositório R-Bloggers. Também indicamos dois livros para consulta: Sheather (2009) e Christian Kleiber (2008).

Referências

Christian Kleiber, A. Z. a. (2008). *Applied Econometrics with R*. Use R. Springer-Verlag New York, 1 edition.

Franses, P. H. and Paap, R. (2001). *Quantitative Models in Marketing Research*. Cambridge University Press.

Johnston, J., Johnston, J., and DiNardo, J. (1997). *Econometric Methods*. Economics series. McGraw-Hill.

Maddala, G. (2001). *Introduction to Econometrics*. Wiley.

Sheather, S. (2009). *A Modern Approach to Regression with R*. Springer Texts in Statistics. Springer New York.

5 Apêndice:

Tabela 1: Resultado das regressões

	Dependent variable:	
	Y	
	logistic	*probit*
	(1)	(2)
log (price heinz/hunts)	-5.987^{***} (0.401)	-3.274^{***} (0.212)
display heinz only	0.526^{**} (0.254)	0.271^{**} (0.129)
display hunts only	-0.651^{**} (0.254)	-0.376^{**} (0.150)
feat. heinz only	0.474 (0.320)	0.188 (0.158)
feat. hunts only	-1.033^{***} (0.361)	-0.573^{***} (0.199)
feat. and display heinz	0.473 (0.489)	0.255 (0.248)
feat. and display hunts	-1.981^{***} (0.479)	-1.094^{***} (0.275)
Constante	3.290^{***} (0.151)	1.846^{***} (0.075)
Observations	2,498	2,498
Log Likelihood	−601.238	−598.528
Akaike Inf. Crit.	1,218.477	1,213.057
Note:		*p<0.1; **p<0.05; ***p<0.01