Logit

Probit and Logit Models in R

Copyright 2013 by Ani Katchova

离散选择模型(微观计量)的R实现，主要参考Econometrics Academy <https://sites.google.com/site/econometricsacademy/econometrics-models/probit-and-logit-models>

#读入数据  
mydata<- read.csv("d:/download/probit\_insurance.csv") %>%  
 tbl\_df %>%  
 print

## Source: local data frame [3,206 x 26]  
##   
## ins retire age hstatusg hhincome educyear married hisp personid private  
## 1 0 0 62 0 0.000 12 0 0 54963010 0  
## 2 0 0 59 0 0.000 12 0 0 21629010 0  
## 3 0 1 60 1 0.000 13 0 0 20421010 0  
## 4 0 0 62 0 0.000 10 0 0 38784010 0  
## 5 0 0 54 0 0.000 9 0 0 55630030 0  
## 6 0 1 62 1 0.000 12 1 0 15893020 0  
## 7 0 0 59 0 0.000 5 1 0 86263010 0  
## 8 0 0 59 0 0.000 11 0 0 51270010 0  
## 9 0 0 65 0 0.000 14 0 0 57341010 0  
## 10 0 0 58 0 0.101 12 0 0 39703010 0  
## .. ... ... ... ... ... ... ... ... ... ...  
## Variables not shown: eprhi (int), white (int), female (int), excel (int),  
## vegood (int), good (int), fair (int), poor (int), chronic (int), adl  
## (int), seprhi (int), sretire (int), age2 (int), agefem (int), agechr  
## (int), agewhi (int)

# 描述性统计  
summary(mydata$ins)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 0.387 1.000 1.000

summary(select(mydata,retire, age, hstatusg, hhincome, educyear, married, hisp))

## retire age hstatusg hhincome   
## Min. :0.000 Min. :52.0 Min. :0.000 Min. : 0.0   
## 1st Qu.:0.000 1st Qu.:65.0 1st Qu.:0.000 1st Qu.: 17.0   
## Median :1.000 Median :67.0 Median :1.000 Median : 31.1   
## Mean :0.625 Mean :66.9 Mean :0.705 Mean : 45.3   
## 3rd Qu.:1.000 3rd Qu.:69.0 3rd Qu.:1.000 3rd Qu.: 52.8   
## Max. :1.000 Max. :86.0 Max. :1.000 Max. :1312.1   
## educyear married hisp   
## Min. : 0.0 Min. :0.000 Min. :0.0000   
## 1st Qu.:10.0 1st Qu.:0.000 1st Qu.:0.0000   
## Median :12.0 Median :1.000 Median :0.0000   
## Mean :11.9 Mean :0.733 Mean :0.0727   
## 3rd Qu.:14.0 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :17.0 Max. :1.000 Max. :1.0000

table(mydata$ins)

##   
## 0 1   
## 1965 1241

table(mydata$ins)/sum(table(mydata$ins))

##   
## 0 1   
## 0.6129 0.3871

#定义回归公式  
f1<-as.formula(ins~retire+age+hstatusg+hhincome+educyear+married+hisp)  
# OLS回归  
olsreg <- lm(f1,data=mydata)  
summary(olsreg)

##   
## Call:  
## lm(formula = f1, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.123 -0.406 -0.229 0.530 1.034   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.127086 0.160563 0.79 0.42871  
## retire 0.040851 0.018220 2.24 0.02502  
## age -0.002896 0.002419 -1.20 0.23138  
## hstatusg 0.065558 0.019453 3.37 0.00076  
## hhincome 0.000492 0.000138 3.58 0.00035  
## educyear 0.023369 0.002867 8.15 5.1e-16  
## married 0.123470 0.019362 6.38 2.1e-10  
## hisp -0.121006 0.033666 -3.59 0.00033  
##   
## Residual standard error: 0.467 on 3198 degrees of freedom  
## Multiple R-squared: 0.0826, Adjusted R-squared: 0.0806   
## F-statistic: 41.1 on 7 and 3198 DF, p-value: <2e-16

# Logit回归，假定Logistic残差  
logit<- glm(f1, ,data=mydata,family=binomial (link = "logit"))  
summary(logit)

##   
## Call:  
## glm(formula = f1, family = binomial(link = "logit"), data = mydata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.457 -1.009 -0.703 1.224 2.373   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.715578 0.748621 -2.29 0.02193  
## retire 0.196930 0.084207 2.34 0.01935  
## age -0.014596 0.011287 -1.29 0.19597  
## hstatusg 0.312265 0.091674 3.41 0.00066  
## hhincome 0.002304 0.000762 3.02 0.00250  
## educyear 0.114263 0.014201 8.05 8.6e-16  
## married 0.578636 0.093320 6.20 5.6e-10  
## hisp -0.810306 0.195751 -4.14 3.5e-05  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4279.5 on 3205 degrees of freedom  
## Residual deviance: 3989.8 on 3198 degrees of freedom  
## AIC: 4006  
##   
## Number of Fisher Scoring iterations: 4

# Probit 回归，假定正态残差  
probit<- glm(f1, data=mydata,family=binomial (link="probit"))  
summary(probit)

##   
## Call:  
## glm(formula = f1, family = binomial(link = "probit"), data = mydata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.422 -1.014 -0.698 1.224 2.474   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.069342 0.455310 -2.35 0.01884  
## retire 0.118353 0.051340 2.31 0.02115  
## age -0.008869 0.006863 -1.29 0.19623  
## hstatusg 0.197741 0.055491 3.56 0.00037  
## hhincome 0.001233 0.000437 2.82 0.00480  
## educyear 0.070749 0.008492 8.33 < 2e-16  
## married 0.362338 0.056052 6.46 1.0e-10  
## hisp -0.473114 0.110232 -4.29 1.8e-05  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4279.5 on 3205 degrees of freedom  
## Residual deviance: 3987.2 on 3198 degrees of freedom  
## AIC: 4003  
##   
## Number of Fisher Scoring iterations: 6

#odds ratios  
exp(logit$coefficients)

## (Intercept) retire age hstatusg hhincome educyear   
## 0.1799 1.2177 0.9855 1.3665 1.0023 1.1210   
## married hisp   
## 1.7836 0.4447

exp(probit$coefficients)

## (Intercept) retire age hstatusg hhincome educyear   
## 0.3432 1.1256 0.9912 1.2186 1.0012 1.0733   
## married hisp   
## 1.4367 0.6231

# OLS边际效应，即beta  
coef(olsreg)

## (Intercept) retire age hstatusg hhincome educyear   
## 0.1270857 0.0408508 -0.0028955 0.0655583 0.0004921 0.0233686   
## married hisp   
## 0.1234699 -0.1210059

# Logit回归平均边际效应  
LogitScalar <- mean(dlogis(predict(logit, type = "link")))  
LogitScalar \* coef(logit)

## (Intercept) retire age hstatusg hhincome educyear   
## -0.3725233 0.0427616 -0.0031693 0.0678058 0.0005002 0.0248111   
## married hisp   
## 0.1256459 -0.1759510

# Probit 回归平均边际效应  
ProbitScalar <- mean(dnorm(predict(probit, type = "link")))  
ProbitScalar \* coef(probit)

## (Intercept) retire age hstatusg hhincome educyear   
## -0.3792719 0.0419770 -0.0031458 0.0701344 0.0004372 0.0250932   
## married hisp   
## 0.1285131 -0.1678031

# OLS回归预测值（此处为概率）  
polsreg<- predict(olsreg)  
summary(polsreg)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.156 0.305 0.407 0.387 0.474 1.200

# Logit 模型预测概率  
plogit<- predict(logit, type="response")  
summary(plogit)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.034 0.285 0.399 0.387 0.478 0.965

# Probit 模型预测概率  
pprobit<- predict(probit, type="response")  
summary(pprobit)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0206 0.2870 0.4020 0.3860 0.4770 0.9650

# 模型预测正确率  
table(true = mydata$ins, pred = round(fitted(probit)))

## pred  
## true 0 1  
## 0 1660 305  
## 1 906 335

table(true = mydata$ins, pred = round(fitted(logit)))

## pred  
## true 0 1  
## 0 1657 308  
## 1 896 345

# McFadden's Pseudo R-squared 伪R方  
PseudoRsq <- function(m1,m2) {  
 McFadden<- 1-as.vector(logLik(m1)/logLik(m2))  
 return(McFadden)  
   
}  
  
   
probit0<-update(probit, formula= ins ~ 1)  
PseudoRsq(probit,probit0)

## [1] 0.0683

logit0<-update(logit, formula= ins ~ 1)  
PseudoRsq(logit,logit0)

## [1] 0.06771