Regression with a binary dependent variable

## Regression with a binary dependent variable

So far the dependent variable () has been continuous:

* district-wide average test score
* economic growth rate
* earnings

What if is binary?

* = get into college, or not; = years of education

- = person smokes, or not; = income

- = mortgage application is accepted, or not; = income, house characteristics, marital status, race

Examples of Binary Dependent Variables

| Topic | Dummy Dependent Variable | Description |
| --- | --- | --- |
| Labour Force Participation | Inlabourforce | 0 if out of LF  1 if in the LF |
| Choice of Occupation | Managerial | 0 if not managerial  1 if managerial |
| Firm Location | Shoppingmall | 0 if not in the shopping mall  1 if in the shopping mall |
| Union Membership | Union | 0 if not a union member  1 if a union member |
| Retirement | Retired | 0 if not retired  1 if retired |
| Use of Seat Belts | Seatbeltused | 0 if does not use seat belt  1 if uses seat belt |

## Main concepts

We review the following concepts:

* the linear probability model
* the Probit model
* the Logit model
* maximum likelihood estimation of nonlinear regression models

## Scatter plot of the Boston HMDA data

library(haven)  
hmda <- read\_stata("hmda\_sw.dta")  
# create the variables of interest  
hmda$deny = ifelse(hmda$s7==3, 1, 0)  
hmda$pi\_ratio = hmda$s46/100  
hmda$black = ifelse(hmda$s13==3, 1, 0)  
library(ggplot2)  
figure11\_1 = ggplot(data=hmda, aes(x=pi\_ratio, y=deny))  
figure11\_1 = figure11\_1 +  
geom\_point(alpha=0.2, size=1) +  
geom\_smooth(method="lm", se=F) +  
ylim(-0.4,1.4) + scale\_x\_continuous(breaks = seq(0,3,0.5)) +  
labs(x="P/I Ratio", y="Deny") +  
annotate("text",x=2.5,y=0.9,label="Mortgage denied") +  
annotate("text",x=2.5,y=-0.1,label="Mortgage approved") +  
theme\_bw()

## Trimming the data

# sample of the HMDA data with 150 observations  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

hmda\_s1 = dplyr::sample\_n(filter(hmda,s13==3 & s7==3 & s46/100<1), 35)  
hmda\_s2 = dplyr::sample\_n(filter(hmda,s13==3 & s46/100<1), 15)  
hmda\_s3 = dplyr::sample\_n(filter(hmda,s13==5 & s46/100<1), 50)  
hmda\_s = rbind(hmda\_s1, hmda\_s2, hmda\_s3)  
# create the variables of interest  
hmda\_s$deny = ifelse(hmda\_s$s7==3, 1, 0)  
hmda\_s$pi\_ratio = hmda\_s$s46/100  
hmda\_s$black = ifelse(hmda\_s$s13==3, 1, 0)  
figure11\_1mod = ggplot(data=hmda\_s, aes(x=pi\_ratio, y=deny))  
figure11\_1mod = figure11\_1mod +  
geom\_point(size=2, shape=1) +  
geom\_smooth(method="lm", se=F) +  
ylim(-0.4,1.4) + scale\_x\_continuous(breaks = seq(0,1,0.1)) +  
labs(x="P/I Ratio", y="Deny") +  
annotate("text",x=0.6,y=1.1,label="Mortgage denied") +  
annotate("text",x=0.6,y=-0.1,label="Mortgage approved") +  
annotate("label",x=0.3,y=0.5,label="LPM") +  
theme\_bw()

## LPM

Equation of the LPM regression line:

library(lmtest)

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

library(sandwich)  
LPM = lm(deny ~ pi\_ratio, data=hmda)  
coeftest(LPM, vcov = sandwich)

t test of coefficients:  
  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.079910 0.031953 -2.5008 0.01246 \*   
pi\_ratio 0.603535 0.098441 6.1309 1.02e-09 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## LPM inclduing race variable

LPM2 = lm(deny ~ pi\_ratio + black, data=hmda)  
coeftest(LPM2, vcov=sandwich)

t test of coefficients:  
  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.090514 0.028582 -3.1669 0.001561 \*\*   
pi\_ratio 0.559195 0.088610 6.3107 3.303e-10 \*\*\*  
black 0.177428 0.024931 7.1169 1.455e-12 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Probit Model, Predicted Probabilities and Estimated Effects

Assume that is a binary variable. The model

with $P(Y = 1 \vert X\_1, X\_2, \dots ,X\_k)\\ = \Phi(\beta\_0 + \beta\_1 + X\_1 + \beta\_2 X\_2 + \dots + \beta\_k X\_k)$ is the population Probit model with multiple regressors and is the cumulative standard normal distribution function.

## Logit regression

The population Logit regression function is

The idea is similar to Probit regression except that a different CDF is used:

is the CDF of a standard logistically distributed random variable.

## Probit regression

in (11.4) plays a role of a quantile .Remember that such that the Probit coefficient in (11.4) is the change in associated with a one unit change in . Although the effect on of a change in is linear, the link between and the dependent variable is nonlinear since is a nonlinear function of . . ## Probit chart code

figure11\_3mod = ggplot(data=hmda\_s, aes(x=pi\_ratio, y=deny))  
figure11\_3probit = figure11\_3mod +  
geom\_point(shape=1, size=2) +  
geom\_smooth(method="glm",  
  
method.args=list(family=binomial(link="probit")),  
se=F) +  
ylim(-0.4,1.4) +  
scale\_x\_continuous(breaks = seq(0,1,0.1)) +  
labs(x="P/I Ratio", y="Deny") +  
annotate("label",x=0.3,y=0.5,label="Probit Regression") +  
theme\_bw()

## Probit regression

probit = glm(  
deny ~ black + pi\_ratio,  
data=hmda,  
family=binomial(link="logit"))  
coeftest(probit, vcov=sandwich)

z test of coefficients:  
  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -4.12556 0.34576 -11.9320 < 2.2e-16 \*\*\*  
black 1.27278 0.14607 8.7136 < 2.2e-16 \*\*\*  
pi\_ratio 5.37036 0.96315 5.5758 2.464e-08 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## LPM , Logit, Probit

# rename and create some variables for regression  
hmda$hse\_inc = hmda$s45/100  
hmda$loan\_val = hmda$s6/hmda$s50  
hmda$ccred = hmda$s43  
hmda$mcred = hmda$s42  
hmda$pubrec = ifelse(hmda$s44>0,1,0)  
hmda$denpmi = ifelse(hmda$s53==1,1,0)  
hmda$selfemp = ifelse(hmda$s27a==1,1,0)  
hmda$married = ifelse(hmda$s23a=="M",1,0)  
hmda$single = ifelse(hmda$married==0,1,0)  
hmda$hischl = ifelse(hmda$school>=12,1,0)  
hmda$probunmp = hmda$uria  
hmda$condo = ifelse(hmda$s51 == 1,1,0)

# LPM  
model\_1 = lm(  
deny ~ black + pi\_ratio + hse\_inc + loan\_val + ccred +  
mcred + pubrec + + denpmi + selfemp,  
data=hmda)  
# Logit  
model\_2 = glm(  
deny ~ black + pi\_ratio + hse\_inc + loan\_val + ccred +  
mcred + pubrec + + denpmi + selfemp,  
data=hmda,  
family=binomial(link="logit"))  
# Probit  
model\_3 = glm(  
deny ~ black + pi\_ratio + hse\_inc + loan\_val + ccred +  
mcred + pubrec + + denpmi + selfemp,  
data=hmda,  
family=binomial(link="probit"))

models<-list(model\_1,model\_2,model\_3)  
library(huxtable)

Attaching package: 'huxtable'

The following object is masked from 'package:dplyr':  
  
 add\_rownames

The following object is masked from 'package:ggplot2':  
  
 theme\_grey

library(modelsummary)  
modelsummary(models,estimate = "{estimate}{stars}", output="huxtable")

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| (Intercept) | -0.241\*\*\* | -6.830\*\*\* | -3.519\*\*\* |
|  | (0.032) | (0.541) | (0.266) |
| black | 0.090\*\*\* | 0.726\*\*\* | 0.408\*\*\* |
|  | (0.017) | (0.175) | (0.096) |
| pi\_ratio | 0.461\*\*\* | 4.743\*\*\* | 2.472\*\*\* |
|  | (0.087) | (1.046) | (0.549) |
| hse\_inc | -0.054 | -0.191 | -0.238 |
|  | (0.096) | (1.234) | (0.653) |
| loan\_val | 0.092\*\* | 1.784\*\*\* | 0.764\*\* |
|  | (0.034) | (0.498) | (0.251) |
| ccred | 0.031\*\*\* | 0.290\*\*\* | 0.154\*\*\* |
|  | (0.004) | (0.040) | (0.021) |
| mcred | 0.023\* | 0.307\* | 0.161\* |
|  | (0.011) | (0.139) | (0.073) |
| pubrec | 0.201\*\*\* | 1.221\*\*\* | 0.701\*\*\* |
|  | (0.023) | (0.203) | (0.117) |
| denpmi | 0.712\*\*\* | 4.518\*\*\* | 2.542\*\*\* |
|  | (0.041) | (0.550) | (0.281) |
| selfemp | 0.060\*\*\* | 0.672\*\* | 0.354\*\* |
|  | (0.018) | (0.210) | (0.111) |
| Num.Obs. | 2380 | 2380 | 2380 |
| R2 | 0.258 |  |  |
| R2 Adj. | 0.255 |  |  |
| AIC | 711.7 | 1302.1 | 1305.7 |
| BIC | 775.2 | 1359.8 | 1363.4 |
| Log.Lik. | -344.855 | -641.046 | -642.847 |
| F | 91.479 | 32.298 | 35.294 |
| RMSE | 0.28 | 0.28 | 0.28 |