Chapter 11 - Regression with a Binary Dependent Variable

Ercio Munoz

November 8, 2018

In this chapter we focus on models with binary dependent variable. First, we import and set up the data set.

```
library(foreign)
a="http://fmwww.bc.edu/ec-p/data/stockwatson/hmda_sw.dta"
d=read.dta(a)

d$deny=as.numeric(d$s7==3)
d$pi_rat=d$s46/100
d$black=as.numeric(d$s13==3)
attach(d)
# Descriptive stats
summary(d[,c("deny","black","pi_rat")])
```

```
##       deny             black            pi_rat      
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.2800  
##  Median :0.0000   Median :0.0000   Median :0.3300  
##  Mean   :0.1197   Mean   :0.1424   Mean   :0.3308  
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.3700  
##  Max.   :1.0000   Max.   :1.0000   Max.   :3.0000
```

```
# Looking the first 10 observations
head(d[,c("deny","black","pi_rat")],10)
```

```
##    deny black pi_rat
## 1     0     0  0.221
## 2     0     0  0.265
## 3     0     0  0.372
## 4     0     0  0.320
## 5     0     0  0.360
## 6     0     0  0.240
## 7     0     0  0.350
## 8     0     0  0.280
## 9     1     0  0.310
## 10    0     0  0.180
```

Linear probability model:

```
lpm=lm(deny~pi_rat)
summary(lpm)
```

```
##
## Call:
## lm(formula = deny ~ pi_rat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.73070 -0.13736 -0.11322 -0.07097  1.05577
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.07991    0.02116  -3.777 0.000163 ***
## pi_rat       0.60353    0.06084   9.920  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3183 on 2378 degrees of freedom
## Multiple R-squared:  0.03974,    Adjusted R-squared:  0.03933
## F-statistic: 98.41 on 1 and 2378 DF,  p-value: < 2.2e-16
```

Probit model:

```
probit=glm(deny~pi_rat,family=binomial(link="probit"))
summary(probit)
```

```
##
## Call:
## glm(formula = deny ~ pi_rat, family = binomial(link = "probit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4140  -0.5281  -0.4750  -0.3900   2.8159
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1941     0.1378 -15.927  < 2e-16 ***
## pi_rat        2.9679     0.3858   7.694 1.43e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1744.2  on 2379  degrees of freedom
## Residual deviance: 1663.6  on 2378  degrees of freedom
## AIC: 1667.6
##
## Number of Fisher Scoring iterations: 6
```

Probit model with two regressors:

```
p2 = glm(deny~pi_rat+black,family=binomial(link="probit"))
summary(p2)
```

```
##
## Call:
## glm(formula = deny ~ pi_rat + black, family = binomial(link = "probit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1208  -0.4762  -0.4251  -0.3550   2.8799
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.25879    0.13669 -16.525  < 2e-16 ***
```

```
## pi_rat       2.74178    0.38047   7.206 5.75e-13 ***
## black        0.70816    0.08335   8.496  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1744.2  on 2379  degrees of freedom
## Residual deviance: 1594.3  on 2377  degrees of freedom
## AIC: 1600.3
##
## Number of Fisher Scoring iterations: 5
```

Predicting the probability of deny=1 when p_irat=.3 and black=0, remember that we need to evaluate $(\beta_1 + \beta_2 * X)$ in a cumulative normal distribution:

```
coef=p2$coefficients
pnorm(coef[1]+coef[2]*.3)
```

```
## (Intercept)
##  0.07546516
```

Logit model:

```
attach(d)
```

```
## The following objects are masked from d (pos = 3):
##
##     bd, black, chval, deny, dnotown, dprop, mi, netw, old, pi_rat,
##     rtdum, s11, s13, s14, s15, s16, s17, s18, s19a, s19b, s19c,
##     s19d, s20, s23a, s24a, s25a, s26a, s27a, s3, s30a, s30c, s31a,
##     s31c, s32, s33, s34, s35, s39, s4, s40, s41, s42, s43, s44,
##     s45, s46, s47, s48, s49, s5, s50, s51, s52, s53, s54, s55,
##     s56, s57, s6, s7, s9, school, seq, uria, vr
```

```
l = glm(deny~pi_rat+black,family=binomial(link="logit"))
summary(l)
```

```
##
## Call:
## glm(formula = deny ~ pi_rat + black, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3709  -0.4732  -0.4219  -0.3556   2.8038
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.1256     0.2684 -15.370  < 2e-16 ***
## pi_rat        5.3704     0.7283   7.374 1.66e-13 ***
## black         1.2728     0.1462   8.706  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1744.2  on 2379  degrees of freedom
```

```
## Residual deviance: 1591.4  on 2377  degrees of freedom
## AIC: 1597.4
##
## Number of Fisher Scoring iterations: 5
```

Predicting the probability of deny=1 when p_irat=.3 and black=0, remember that we need to evaluate $(\beta_1 + \beta_2 * X)$ in a cumulative logistic distribution:

```
coef = l$coefficients
plogis(coef[1]+coef[2]*.3)
```

```
## (Intercept)
##  0.07485143
```