Binary Choice Models

Jerome Dumortier

17 August 2023

Overview

Packages:

- mfx

Binary choice models

- Did you vote during the last election?
- Does an individual get arrested again after being released from prison?
- Participation in the labor market
- Purchasing a home
- Model: $Pr(y = 1|x)$

Dependent variable y takes one of two values: 0 or 1

Numerical Methods

Consider the following equation:

$$y = x^2$$

- What is the value of y if $x = 5$?
- What is the value of x if $y = 81$?

Next, consider the following equation:

$$y = x^2 + \sqrt{x}$$

- What is the value of y if $x = 9$?
- What is the value of x if $y = 84$?

Linear Probability Model

Most rudimentary model: Linear probability model (LPM)

- Use the linear regression model $y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon$
- Problem: Possibility of $E(y_i|x_i) > 1$ or $E(y_i|x_i) < 0$
- It can be shown that disturbance terms are not normally distributed and there is heteroscedastic.

Alternative: Model that calculates the probability of observing a 1.

- Logit and Probit models

Logit and Probit Models

General assumption about some function $G(\cdot)$: $0 \leq G(z) \leq 1$ for all values of z. Let

$$z = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$$

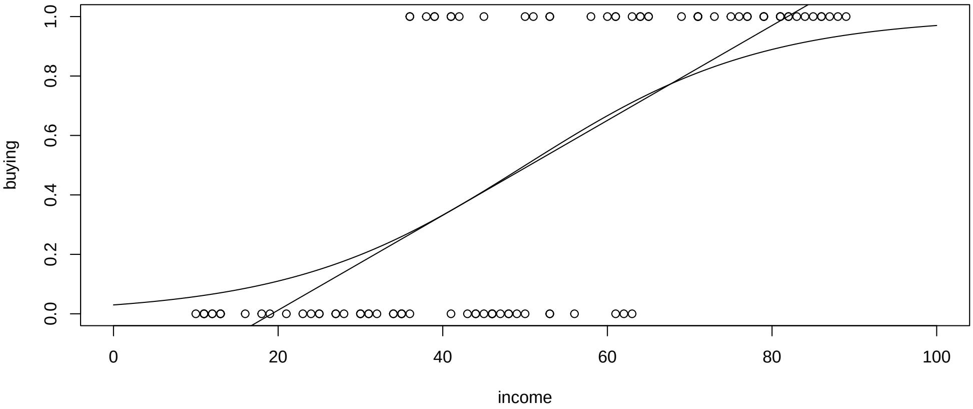
Then, we have

$$P(y = 1|x) = G(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)$$

Notes

- Estimation through Maximum Likelihood
- Difficulty interpreting the values of coefficient

Comparison LPM vs. Logit



Logit Model

Remember the Bernoulli distribution from statistics:

$$Pr(Y = 1) = p$$

$$Pr(Y = 0) = 1 - p$$

with $E(y) = p$. For the logit model we have the following:

$$Pr(y = 1) = G(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

where $z = \beta_0 + \beta_1 \cdot x$.

Probit Model

Instead of using the cumulative logistic distribution, the probit model uses the cumulative normal distribution:

$$G(z) = \Phi(z)$$

Both models lead to similar results (not similar coefficients!).

Example using `organic`

Data description

- *income* of the respondent in $ 1,000
- *buying* of organic food: yes (1) or no (0)

Results of interest for the binary choice model (for other models as well)

- Coefficient estimates
- Marginal effects
- Predicted probabilities

Coefficicent Estimates

Estimation with R

Coefficient estimates using the built-in R command:

```
bhatglm = glm(buying~income,
              family=binomial(link="logit"),
              data=organic)
library(mfx)
bhatmfx = logitmfx(buying~income,data=organic)
```

Obtaining summary from `bhatmfx`

```
summary(bhatmfx$fit)
```

Base Results

```
##
## Call:
## glm(formula = buying ~ income, family = binomial(link = "logit"),
##     data = organic)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.87557    1.13842  -5.161 2.45e-07 ***
## income       0.11709    0.02247   5.211 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.469  on 99  degrees of freedom
## Residual deviance:  70.931  on 98  degrees of freedom
## AIC: 74.931
##
## Number of Fisher Scoring iterations: 6
```

Results from mfx

```
##
## Call:
## glm(formula = formula, family = binomial(link = "logit"), data = data,
##     start = start, control = control, x = T)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.87557    1.13842  -5.161 2.45e-07 ***
## income       0.11709    0.02247   5.211 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.469  on 99  degrees of freedom
## Residual deviance:  70.931  on 98  degrees of freedom
## AIC: 74.931
##
## Number of Fisher Scoring iterations: 6
```

Marginal Effects

Marginal Effects with mfx package

Advantage of mfx package: Estimation of marginal effects

```
bhatmfx$mfxest
```

```
##             dF/dx   Std. Err.        z        P>|z|
## income 0.02919553 0.005634262 5.181785 2.197728e-07
```

Important note:

- Marginal effects are estimated at the mean of the independent variable(s)!

Predicted Probabilities

Fitted Values in a Binary Choice Model

Example:

- What are the predicted probabilities of a person purchasing organic given their annual income (in $ 1,000) of 25, 50, and 75?

Solution in R:

```
datablock = data.frame(income=c(25,50,75))
test = predict(bhatglm,newdata=datablock,type="response")
```

Probit Model

Very similar results compared to Logit:

```
bhatmfx = probitmfx(buying~income,data=organic)
bhatmfx$mfxest
```

```
##              dF/dx   Std. Err.        z        P>|z|
## income 0.02771441 0.004753676 5.830101 5.539374e-09
```

Additional Example

Food Purchases `fpdata`

Food purchases data:

- `strawberries_org`: Frequency of strawberry purchases per month
- `tomatoes_org`: Frequency of strawberry purchases per month
- `age`: Age of the respondent
- `kidsunder12`: Presence of kids under the age of 12
- `rootsurban`: Urban (as opposed to rural) upbringing of respondent
- `education`: Education level
- `income`: Income

Data Preparation and Estimation

```
fpdata$strawberriesorg  = ifelse(fpdata$strawberriesorg==0,0,1)
fpdata$tomatoesorg      = ifelse(fpdata$tomatoesorg==0,0,1)
```

```
bhats = glm(strawberriesorg~age+kidsunder12+rootsurban+
            education+income,
            family=binomial(link="logit"),
            data=fpdata)
bhatt = glm(tomatoesorg~age+kidsunder12+rootsurban+
            education+income,
            family=binomial(link="logit"),
            data=fpdata)
```

Results Strawberries

```
##
## Call:
## glm(formula = strawberriesorg ~ age + kidsunder12 + rootsurban +
##     education + income, family = binomial(link = "logit"), data = fpdata)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.961e-02  6.909e-01   0.101  0.91974
## age         -8.478e-03  1.121e-02  -0.756  0.44947
## kidsunder12  8.526e-02  3.709e-01   0.230  0.81820
## rootsurban   3.507e-01  3.312e-01   1.059  0.28972
## education   -1.203e-01  1.329e-01  -0.905  0.36528
## income       1.524e-05  5.597e-06   2.722  0.00649 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 232.45  on 171  degrees of freedom
## Residual deviance: 222.77  on 166  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 234.77
##
## Number of Fisher Scoring iterations: 4
```

Results Tomatoes

```
##
## Call:
## glm(formula = tomatoesorg ~ age + kidsunder12 + rootsurban +
##     education + income, family = binomial(link = "logit"), data = fpdata)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.329e-01  7.010e-01  -0.190  0.84967
## age         -5.728e-03  1.138e-02  -0.503  0.61466
## kidsunder12 -1.104e-01  3.770e-01  -0.293  0.76956
## rootsurban   3.603e-01  3.364e-01   1.071  0.28417
## education   -4.158e-02  1.338e-01  -0.311  0.75606
## income       1.708e-05  5.892e-06   2.899  0.00374 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.06  on 171  degrees of freedom
## Residual deviance: 216.18  on 166  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 228.18
##
## Number of Fisher Scoring iterations: 4
```

Additional Questions

For the strawberries and tomatoes regression, do the following:

- Calculate the marginal effects of all independent variables
- Calculate the predicted probability for each observation