Financial Econometrics

Binary dependent variable

Maximilian Rohrer

Overview

- Basics of binary dependent variables
- Linear probability model
- Logit and probit regression
- Comparing models and marginal effects
- Estimation and inference
- Measures of fit
- Hypothesis testing

Basics of binary dependent variables

Economic question

- Do banks discriminate on the basis of race when issuing loans?
- By law, they should not

Economic question

- Boston Fed HMDA Dataset (link): 2380 individual applications for single-family mortgages made in 1990 in the greater Boston area collected under the Home Mortgage Disclosure Act
- 28% of black applicants were denied mortgages, but only 9% of white applicants
- But, black and white applicants are not necessarily "identical but for their race"
- We need to compare rates of denials, *holding other applicant characteristics constant*
- Sounds like multiple regression, but with a twist
- **The dependent variable - loan denial - is binary**

Binary dependent variables

- Whether a loan application is accepted or not, is binary
- Many other important concerns have binary outcomes, such as
  - Whether a merger deal is paid in cash
  - Whether a household holds stocks
  - Whether a CEO is fired
  - ...

Discrimination in loan applications

- What other factors determine whether or not an individual loan application is denied?
- Size of the required loan payment relative to the applicant's income (P/I ratio)
- Load the data and investigate
- Scatterplot: x-axis = P/I ratio, y-axis = Denied

Packages you will need today

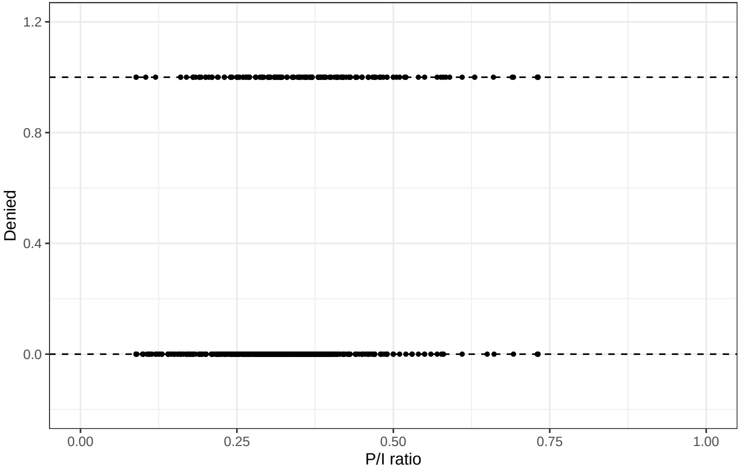
```r
require(stargazer) # regression table
require(lmtest) # robust SE
require(sandwich) # robust SE
require(DescTools) # winsorizing
require(mfx) # marginal effects
```

Set-up the data

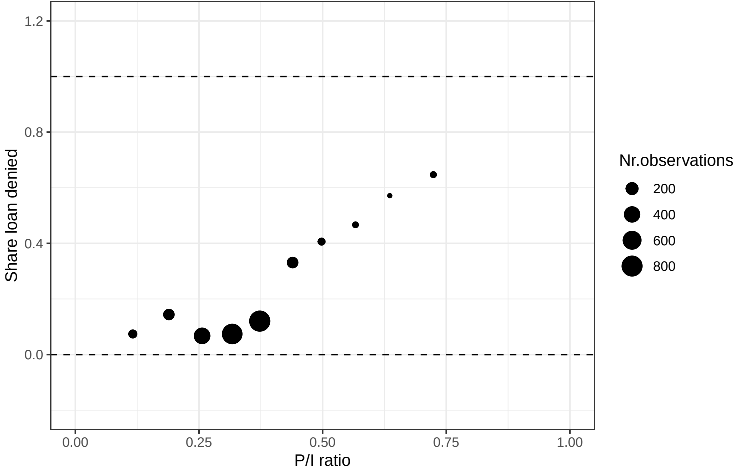
```
# read the data
mortgage <- read.csv("Mortgage.csv")

# clean observations with missing values
mortgage <- na.omit(mortgage)

# Re-scale and winsorize the main dependent variable
# Winsorizing leads to differences in the point estimates compared to the book
mortgage$debt_income_ratio <- Winsorize(mortgage$debt_income_ratio / 100,
                                        probs = c(0.005, 0.995), na.rm = T)

# inspect the data
str(mortgage)
```

```
## 'data.frame':    2379 obs. of  8 variables:
##  $ denied           : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ debt_income_ratio: num  0.221 0.265 0.372 0.32 0.36 ...
##  $ black            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ self_employed    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ score_consumer   : int  5 2 1 1 1 1 1 2 2 2 ...
##  $ score_mortgage   : int  2 2 2 2 1 1 2 2 2 1 ...
##  $ insurance_denied : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ loan_price_ratio : num  0.8 0.922 0.944 0.881 0.6 ...
##  - attr(*, "na.action")= 'omit' Named int 2246
##   ..- attr(*, "names")= chr "2246"
```

Visualize the data



Visualize the data



Linear probability model

Linear probability model

Natural starting point is the linear regression model with a single regressor

$$D(\textit{denied}) = \alpha + \beta \text{debt-to-income} + u$$

- What does the predicted value $\hat{y}$ mean when y is binary?

Lineary probability model

- In the LPM, the predicted value of y is interpreted as the predicted probability that $y = 1$, and β is the change in that predicted probability for a unit increase in x

$$y = \alpha + \beta x + u$$

- For binary y: $E(y|x) = 1 \times Prob(y = 1|x) + 0 \times Prob(y = 0|x) = Prob(y = 1|x)$
- Under zero conditional mean assumption ($E(u|x) = 0$): $E(y|x) = E(\alpha + \beta x + u) = \alpha + \beta x$
- So, $Prob(y = 1|x) = \alpha + \beta x$

Linear probability model

- When y is binary, the linear regression model $y = \alpha + \beta x + u$ is called linear probabiltiy model because $Prob(y = 1|x) = \alpha + \beta x$
- Predicted values from such models are probabilties and the relationship is linear
  - $\hat{y}$ is the predicted probability that $y = 1$, given x
  - β is the change in probability that $y = 1$ for a unit change in x

$$\beta_1 = \frac{Prob(y = 1|x + \Delta x) - Prob(y = 1|x)}{\Delta x}$$

Lineary probability model applied to HMDA data

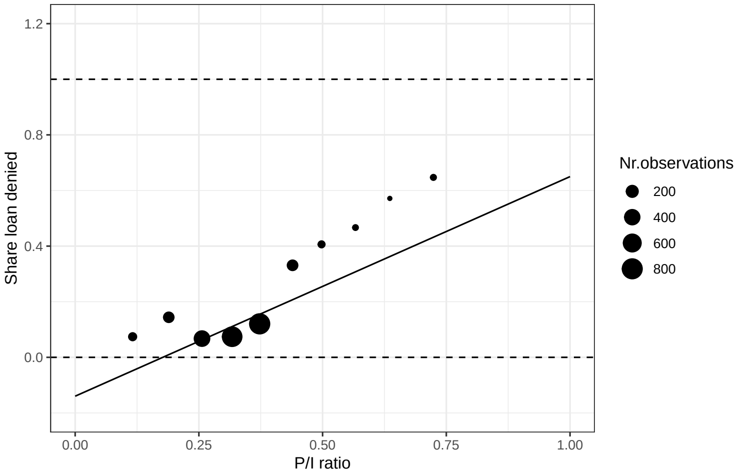
```r
# regression
fit.lpm <- lm(denied ~ debt_income_ratio, data = mortgage)

# robust standard errors
se.robust.fit.lpm <- coeftest(fit.lpm, vcov = vcovHC(fit.lpm))[,2]
# equivalent option: coeftest(fit.lpm, vcov = hccm)
```

```r
# display the results
stargazer(list(fit.lpm),
          se = list(se.robust.fit.lpm),
          type="text", keep.stat=c("n", "rsq", "adj.rsq"),
          report=('vc*t'))
```

Linear probability model applied to HMDA data

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                denied
## ------------------------------------------------
## debt_income_ratio             0.790***
##                               t = 7.673
##
## Constant                      -0.140***
##                              t = -4.256
##
## ------------------------------------------------
## Observations                    2,379
## R2                              0.041
## Adjusted R2                     0.040
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Linear probability model applied to HMDA data



Is there discriminiation?

```
# regression
fit.lpm.multiple <- lm(denied ~ debt_income_ratio + black, data = mortgage)

# robust standard errors
se.robust.fit.lpm.multiple <-
  coeftest(fit.lpm.multiple,
           vcov = vcovHC(fit.lpm.multiple))[,2]
# equivalent option: coeftest(fit.lpm.multiple, vcov = hccm)
```

```
# display the results
stargazer(list(fit.lpm, fit.lpm.multiple),
          se = list(se.robust.fit.lpm, se.robust.fit.lpm.multiple),
          type="text",
          keep.stat=c("n", "rsq", "adj.rsq"),
          report=('vc*t'))
```

Linear probability model applied to HMDA data

```
##
## ==============================================
##                     Dependent variable:
##                 ------------------------------
##                             denied
##                       (1)            (2)
## ----------------------------------------------
## debt_income_ratio    0.790***       0.712***
##                     t = 7.673      t = 6.944
##
## black                               0.172***
##                                    t = 6.900
##
## Constant            -0.140***      -0.139***
##                     t = -4.256    t = -4.261
##
## ----------------------------------------------
## Observations          2,379          2,379
## R2                    0.041          0.075
## Adjusted R2           0.040          0.074
## ==============================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

Interpretation

$$\widehat{denied} = -0.139 + 0.712 \times debt\_income\_ratio + 0.172 \times black$$

- What is the predicted probability of loan denial for a black person with a debt_income_ratio = 0.30?
- $Prob(denied| \times debt\_income\_ratio = 0.30, black = 1) = -0.139 + 0.712 \times 0.3 + 0.172 \times 1 = 0.2466$
- What is the predcited probability of loan denial for a white person with the same debt_income_ratio?
- $Prob(denied| \times debt\_income\_ratio = 0.30, black = 1) = -0.139 + 0.712 \times 0.3 + 0.172 \times 0 = 0.0746$
- What's the effect of race? 17.2%
- Still plenty of room for omitted variable bias...

Summary linear probability model

- The LPM models $Prob(y = 1|x)$ as a linear function of x
- Advantages:
  - Simple to estimate and interpret
  - Inference the same as for multiple regression (note: LPM is inherently heteroskedastic)
- Disadvantages:
  - A LPM says that the change in the predicted probability for a given change in x is the same for all values of x, but that does not always make sense
  - Also, LPM predicts probabilities that can be < 0 and > 1
- Overall, we need a non-linear model: probit and logit regression

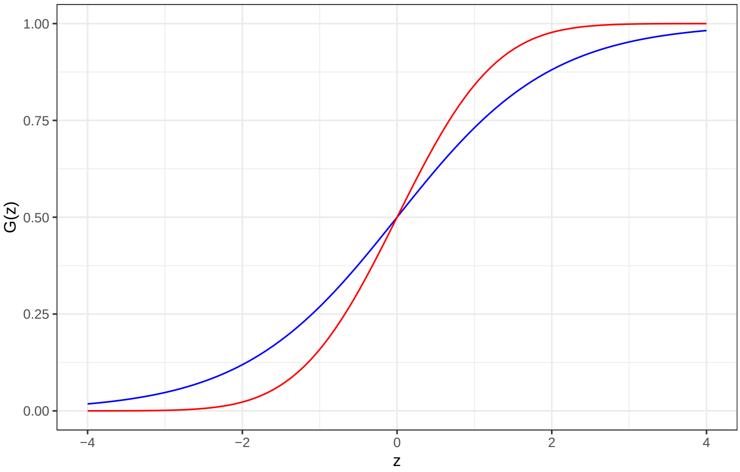
Logit and probit regression

Logit and probit regression

$$Prob(y = 1|x) = G(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)$$

- Probit model: $G(z) = \Phi(z) = \int_{-\infty} \phi(v)dv$ ($\Phi()$ is the standard normal cumulative distribution function)
- Logit model: $G(z) = \Lambda(z) = \frac{exp(z)}{1+exp(z)}$ ($\Lambda(z)$ is the logistic cumulative distribution function)

Logit (blue) and probit (red) linking function, visualization



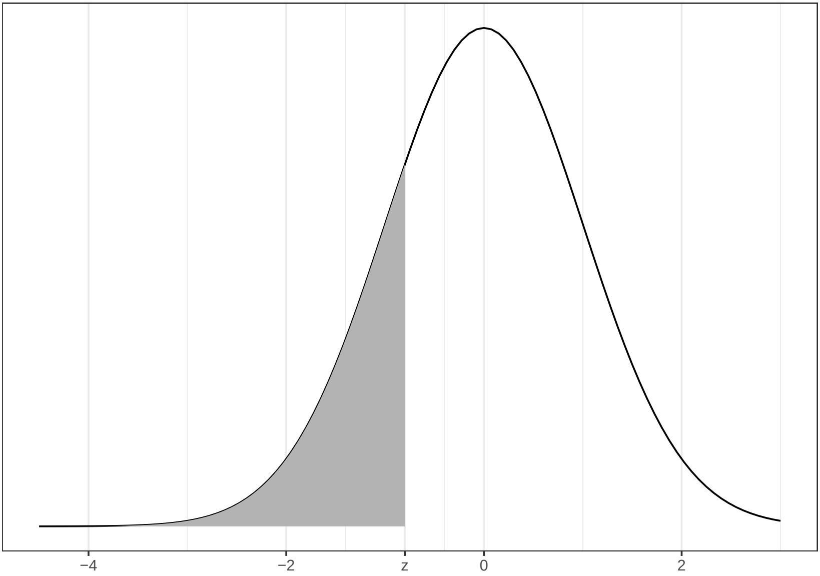
Probit regression

$$Prob(y = 1|x) = \Phi(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)$$

- Probit regression models the probability that $y = 1$ using the cummulative standard normal distribution, $\Phi(z)$ evaluated at $z = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$
- $\Phi(z) = \int_{-\infty}^{z} \phi(v) dv$
- $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ is the "z-value" or "z-index"
- Suppose
  - $Prob(y = 1|x) = \Phi(\beta_0 + \beta_1 x)$
  - $\beta_0 = -2$, $\beta_1 = 3$, and $x = 0.4$
  - $Prob(y = 1|0.4) = \Phi(-2 + 3 \times 0.4) = \Phi(-0.8)$
  - $Prob(y = 1|0.4) =$ area under the standard normal density to the left of $z = -0.8$, which is 0.2119
  - in R, $\Phi(z)$ equals:

```
pnorm(z, mean = 0, sd = 1)
```

Probit regression



Why cumulative normal regression?

- The "S-shape" gives us:
  - Marginal probabilities depend on the value of x
  - Predicted probabilities are bounded between 0 and 1
- Easy to use
- Relative straight forward interpretation
  - $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
  - β_1 is the change in the z-value of a unit change in x_1 holding everything else constant

Probit applied to HMDA data

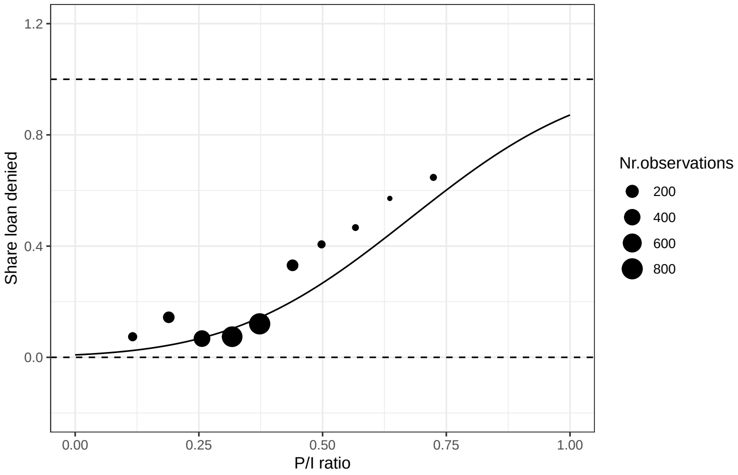
```r
# regression
fit.probit <- glm(denied ~ debt_income_ratio, data = mortgage,
                  family = binomial(link = probit))
```

```r
# display the results
stargazer(list(fit.probit),
          se = list(se.robust.fit.lpm),
          type="text", keep.stat=c("n", "rsq", "adj.rsq"),
          report=('vc*t'))
```

Probit applied to HMDA data

```
##
## =============================================
##                       Dependent variable:
##                   ---------------------------
##                             denied
## ---------------------------------------------
## debt_income_ratio          3.510***
##                            t = 8.597
##
## Constant                  -2.376***
##                           t = -16.232
##
## ---------------------------------------------
## Observations                 2,379
## =============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Probit applied to HMDA data



Is there discrimination?

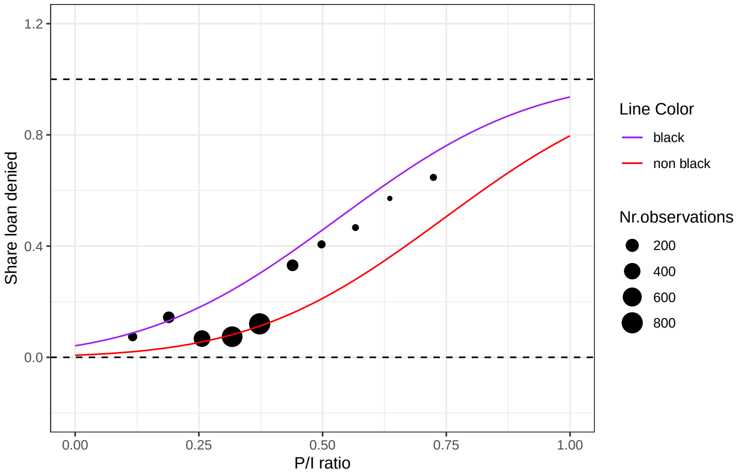
```
# regression
fit.probit.multiple <- glm(denied ~ debt_income_ratio + black,
                           data = mortgage,
                           family = binomial(link = probit))
```

```
# display the results
stargazer(list(fit.probit, fit.probit.multiple),
          se = list(se.robust.fit.lpm),
          type="text",
          keep.stat=c("n", "rsq", "adj.rsq"),
          report=('vc*t'))
```

Probit applied to HMDA data

```
##
## ==============================================
##                      Dependent variable:
##                  ----------------------------
##                             denied
##                      (1)            (2)
## ----------------------------------------------
## debt_income_ratio    3.510***       3.258***
##                    t = 8.597      t = 7.856
##
## black                               0.695***
##                                   t = 8.318
##
## Constant            -2.376***      -2.429***
##                   t = -16.232    t = -16.304
##
## ----------------------------------------------
## Observations          2,379          2,379
## ==============================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

Probit applied to HMDA data



Interpretation

$$\widehat{denied} = \Phi(-2.429 + 3.258 \times debt\_to\_income + 0.695 \times black)$$

- What is the predicted probability of loan denial for a black person with a debt_income_ratio = 0.30?
- $Prob(denied | \times debt\_income\_ratio = 0.30, black = 1) = \Phi(-2.429 + 3.258 \times 0.30 + 0.695 \times 1) = 0.225$
- What is the predcited probability of loan denial for a white person with the same debt_income_ratio?
- $Prob(denied | \times debt\_income\_ratio = 0.30, black = 1) = \Phi(-2.429 + 3.258 \times 0.30 + 0.695 \times 0) = 0.0733$
- What's the effect of race? 15.1%
- Is it the same for a debt_income_ratio = 0.10?
- Still plenty of room for omitted variable bias...

Logit regression

$$Prob(y = 1|x) = \Lambda(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)$$

- Logit regression models the probability that $y = 1$ using the cumulative standard logistic distribution function, $\Lambda(z)$ evaluated at $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- $\Lambda(z) = \frac{exp(z)}{1+exp(z)}$
- Suppose
  - $Prob(y = 1|x) = \Lambda(\beta_0 + \beta_1 x)$
  - $\beta_0 = -2$, $\beta_1 = 3$, and $x = 0.4$
  - $Prob(y = 1|0.4) = \Lambda(-2 + 3 \times 0.4) = \Lambda(-0.8)$
  - $Prob(y = 1|0.4) = 0.31$
- Note: because logit and probit use different probability functions, the coefficients (β's) are different in logit and probit
- In R, $\Lambda(z)$ equals:

```
plogis(z, location = 0, scale = 1)
```

Logit regression

$$Prob(y = 1|x) = \Lambda(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)$$

- Odds: $\frac{Prob(y=1|x)}{Prob(y=0|x)} = \frac{Prob(y=1|x)}{1-Prob(y=1|x)}$
- Odds in a logit model: $\exp(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)$
- The coefficients are given in units of log odds
- β_1 is the change in the log-odds for a unit change in x_1 holding everything else constant

Logit applied to HMDA data

```r
# regression
fit.logit <- glm(denied ~ debt_income_ratio,
                 data = mortgage,
                 family = binomial(link = logit))
```

```r
# display all models
stargazer(list( fit.logit),
          type="text",
          keep.stat=c("n", "rsq", "adj.rsq"),
          report=('vc*t'))
```

Logit applied to HMDA data

```
##
## ============================================
##                     Dependent variable:
##                   ---------------------------
##                           denied
## --------------------------------------------
## debt_income_ratio        6.892***
##                          t = 9.229
##
## Constant                 -4.377***
##                         t = -15.849
##
## --------------------------------------------
## Observations               2,379
## ============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Is there discrimination?

```r
# regression
fit.logit.multiple <- glm(denied ~ debt_income_ratio + black,
                  data = mortgage,
                  family = binomial(link = logit))
```

```r
# display all models
stargazer(list(fit.logit, fit.logit.multiple),
          type="text",
          keep.stat=c("n", "rsq", "adj.rsq"),
          report=('vc*t'))
```

Logit applied to HMDA data

```
##
## ==============================================
##                       Dependent variable:
##                   ----------------------------
##                              denied
##                        (1)            (2)
## ----------------------------------------------
## debt_income_ratio    6.892***      6.452***
##                     t = 9.229     t = 8.539
##
## black                              1.248***
##                                   t = 8.508
##
## Constant            -4.377***     -4.492***
##                    t = -15.849   t = -15.963
##
## ----------------------------------------------
## Observations          2,379         2,379
## ==============================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

Interpretation

$$\widehat{denied} = \Lambda(-4.492 + 6.452 \times debt\_to\_income + 1.248 \times black)$$

- What is the predicted probability of loan denial for a black person with a debt_income_ratio = 0.30?
- $Prob(denied| \times debt\_income\_ratio = 0.30, black = 1) = \Lambda(-4.492 + 6.452 \times 0.30 + 1.248 \times 1) = 0.213$
- What is the predicted probability of loan denial for a white person with the same debt_income_ratio?
- $Prob(denied| \times debt\_income\_ratio = 0.30, black = 1) = \Lambda(-4.492 + 6.452 \times 0.30 + 1.248 \times 0) = 0.072$
- What's the effect of race? 14.1%
- Is it the same for a debt_income_ratio = 0.10?
- Still plenty of room for omitted variable bias...

Logit versus Probit

- The main reason is historical: logit is computationally faster and easier, but that does not matter nowadays
- In practise, logit and probit are very similar
- Empirical results typically should not hinge on the logit versus probit choice

Quick Recap

- Basics of Binary Dependent Variables
- Linear Probability Model
- Logit and Probit Regression
- Comparing Models & Marginal Effects
- Estimation and Inference
- Measures of Fit
- Hypothesis Testing

Comparing Models & Marginal Effects

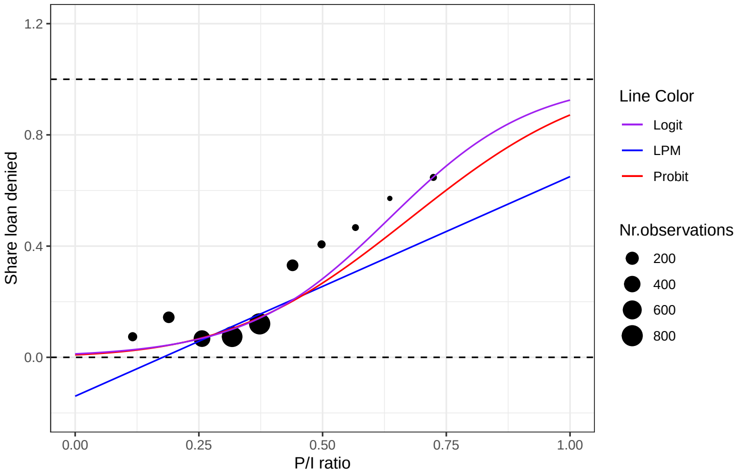
Comparing all three models

```
stargazer(list(fit.lpm.multiple,
               fit.logit.multiple,
               fit.probit.multiple),
          se = list(se.robust.fit.lpm.multiple,
                    NULL, NULL),
          type="text",
          keep.stat=c("n", "rsq", "adj.rsq"),
          report=('vc*t'))
```

Comparing all three models

```
##
## =====================================================
##                           Dependent variable:
##                     ---------------------------------
##                                  denied
##                        OLS      logistic     probit
##                        (1)        (2)         (3)
## -----------------------------------------------------
## debt_income_ratio  0.712***   6.452***    3.258***
##                   t = 6.944  t = 8.539   t = 7.856
##
## black              0.172***   1.248***    0.695***
##                   t = 6.900  t = 8.508   t = 8.318
##
## Constant          -0.139***  -4.492***   -2.429***
##                   t = -4.261 t = -15.963 t = -16.304
##
## -----------------------------------------------------
## Observations         2,379      2,379       2,379
## R2                   0.075
## Adjusted R2          0.074
## =====================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Predicted path for all models



Marginal effects

- Marginal effect: the effect on the dependent variable that results from changing an independent variable by a small amount
- Probit and logit are non-linear functions
- Hence, the ultimate effect of one unit change in a regressor (x) on predicted probabilities is different from different values of x
- It is common to report marginal (or partial) effects instead of coefficients
  - Marginal effect at the means
  - Average marginal effect

Marginal effects

```r
# marginal effect at the mean
fit.probit.mfx.at.mean <-
  probitmfx(denied ~ debt_income_ratio + black,
            data = mortgage,
            atmean = T)

# average marginal effect
fit.probit.mfx.average <-
  probitmfx(denied ~ debt_income_ratio + black,
            data = mortgage,
            atmean = F)
```

Marginal effects

```
fit.probit.mfx.at.mean
```

```
## Call:
## probitmfx(formula = denied ~ debt_income_ratio + black, data
##     atmean = T)
##
## Marginal Effects:
##                       dF/dx Std. Err.      z     P>|z|
## debt_income_ratio 0.588885  0.073852 7.9738 1.538e-15 ***
## black             0.166484  0.024505 6.7939 1.091e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## dF/dx is for discrete change for the following variables:
##
## [1] "black"
```

Marginal effects

```
fit.probit.mfx.average
```

```
## Call:
## probitmfx(formula = denied ~ debt_income_ratio + black, data
##     atmean = F)
##
## Marginal Effects:
##                      dF/dx Std. Err.      z     P>|z|
## debt_income_ratio 0.597761  0.075365 7.9316 2.164e-15 ***
## black             0.165578  0.024077 6.8769 6.116e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## dF/dx is for discrete change for the following variables:
##
## [1] "black"
```

Estimation and Inference

Estimation of Logit and Probit

- Focus on probit with one explanatory variable
- Minimize squared error
  - Recall OLS: min $\sum_{i=1}^{n}(y_i - \alpha - \beta x)^2 \rightarrow \alpha, \beta$
  - Non-linear least squares min $\sum_{i=1}^{n}(y_i - \Phi(\alpha - \beta x))^2 \rightarrow \alpha, \beta$
  - How to solve minimization problem?
    - specialized minimization algorithm
    - Instead of non-linear least squares, a more efficient estimator is used...

Maximum Likelihood

- The likelihood function is the conditional density of $y_1, \ldots, y_n$ given $x_1, \ldots, x_n$ treated as a function of the unknown parameters α, β
- Maximizing the probability of observing the data given the assumed model
- For probit (with one explanatory variable):

$$f(y_i|x_i, \alpha, \beta) =$$

$$\Phi(\alpha + \beta x_i)^{y_i} \times (1 - \Phi(\alpha + \beta x_i)^{1-y_i})$$

$$ln(L(\alpha, \beta)) = \ln(\prod_{i=1}^{n} f(y_i|x_i, \alpha, \beta))$$

$$\max ln(L(\alpha, \beta)) \rightarrow \alpha, \beta$$

Maximum likelihood

- The maximum likelihood estimator (MLE) is the value of $\alpha, \beta_1, \ldots, \beta_k$ that maximinize the likelihood function
- The MLE is the value of $\alpha, \beta_1, \ldots, \beta_k$ that best describe the full distribution of the data
- In large samples, MLE is:
  - consistent
  - normally distributed
  - efficient (has the smallest variance of all estimators)

Measures of fit

Measures of fit for logit and probit

- The R^2 and *Adjusted* R^2 do not make sense here (even for LPM). So, two other specialised measures are used
  - The fraction correctly predicted = fraction of y's for which the predicted probability is $> 50\%$ when y_i is 1, or is $< 50\%$ when y_i is 0
    - What about rare events (rare disease)? Ease to predict $y = 0$, but hard to predict $y = 1$. How does the aforementioned measure behave?
  - The pseudo R^2 (also known as McFadden R^2) measures the imporvement in the value of the log likelihood, relative to having no x's: $1 - \frac{\ln(L(\alpha, \beta_1, \ldots, \beta_k))}{\ln(L(\alpha))}$

Measures of fit

```r
# Probit
PseudoR2(fit.probit)
```

```
##   McFadden
## 0.04813484
```

```r
# Equivalent approach
1 - fit.probit$deviance / fit.probit$null.deviance
```

```
## [1] 0.04813484
```

```r
# Logit
PseudoR2(fit.logit)
```

```
##   McFadden
## 0.05159848
```

Hypothesis testing

Hypothesis test

- Usual t-tests and confidence intervals can be used testing H0: β=0
- For joint hypothesis test, use likelihood ratio test that compares likelihood functions of the restricted and unrestricted models
- For q restrictions:

$$LR = 2(\ln(L(unrestricted)) - \ln(L(restricted))) \sim \chi_q^2$$

- Our example:
  - we augment previous model with following variables: score_consumer and score_mortgage
  - subsequently we ask, whether coefficients on both variables are jointly zero

Hypothesis test

```
# unrestricted
fit.probit.unrestricted <- glm(
  denied ~ debt_income_ratio + black + score_consumer + score_mortgage,
  data = mortgage, family = binomial(link = probit))
# restricted
fit.probit.restricted <- glm(denied ~ debt_income_ratio + black,
                             data = mortgage,
                  family = binomial(link = probit))
# likelihood ratio test (requrie lmtest)
lrtest(fit.probit.restricted, fit.probit.unrestricted)
```

```
## Likelihood ratio test
##
## Model 1: denied ~ debt_income_ratio + black
## Model 2: denied ~ debt_income_ratio + black + score_consumer + score_mortgag
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -796.75
## 2   5 -746.07  2 101.35  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

- If y is binary, then $E(y|x) = Prob(y = 1|x)$
- Three models
  - linear probability model (linear multiple regression)
  - probit (cumulative standard normal distribution)
  - logit (cumulative standard logistic distribution)
- All models produce predicted probabilities
- Probit and logit are estimated using maximum likelihood
- Marginal effects depend on initial values of x for logit and probit

Textbook

- Binary dependent variable: Chapter 11, Stock and Watson, Introduction to Econometrics, Global Edition, 4th edition