Lecture notes to Stock and Watson chapter 11

Probit and logit regression

Tore Schweder

October 2009

Outline

- Binary regression - why?
- Probit and logit - cases of the generalized linear model, glm
- Maximum likelihood estimation
- Application to Boston Home Mortgage Data: is there a race bias in mortgage denials?

Exercise 10.10

a. In the model $Y_{it} = \alpha_i + u_{it}$

$$\begin{aligned} \widehat{\alpha}_i &= \overline{Y}_{i.} = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \\ Y_{it} &= \alpha_i + \lambda_t + u_{it} \Rightarrow \end{aligned}$$

b. No constency in $\widehat{\alpha}_i$ (when $var(u_{it}) > 0$), and only normal distribution for small T when data are normally distributed!

In the model $Y_{it} = \alpha_i + \lambda_t + u_{it}$

$$\widehat{\alpha}_i = \overline{Y}_{i.}, \quad \widehat{\lambda}_t = \overline{Y}_{.t} - \overline{Y}_{..}$$

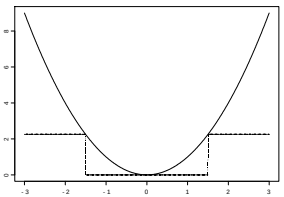
are the least squares estimators (check!) when $\lambda_1 = 0$ identifies the parameters. Same conclusion!

Consistency, and Chebychev's inequality

- An estimator $\widehat{\theta}$ is consistent whenever $\widehat{\theta} \xrightarrow{P} \theta$, i.e $P\left(|\widehat{\theta} - \theta| > \varepsilon\right) \to 0$ for all $\epsilon > 0$, as the data increases to ∞.
- $var\left(\widehat{\theta}\right) \to 0$ ensures consistencey for unbiased estimators, by Chebychev's inequality: $P\left(|\widehat{\theta} - E\widehat{\theta}| > \varepsilon\right) \leq var\left(\widehat{\theta}\right) / \varepsilon^2$

Proof of the inequality:
Let D be the dummy for $|\widehat{\theta} - E\widehat{\theta}| > \varepsilon$,
then

$$\begin{aligned} \varepsilon^2 D &\leq \left(\widehat{\theta} - E\widehat{\theta}\right)^2 \Rightarrow \\ \varepsilon^2 P\left(|\widehat{\theta} - E\widehat{\theta}| > \varepsilon\right) &= \varepsilon^2 E\left(D\right) \leq E\left(\widehat{\theta} - E\widehat{\theta}\right)^2 = var\left(\widehat{\theta}\right) \end{aligned}$$



Regression for binary response variables

Example questions

1. Is the probability of a firm going bust during 2009 dependent on branch?
   - Frame: firms registered 1. January 2009 in Norway
   - Explanatory variable: dummies for branch
   - Control variables: Size of firm (turnover; employment;...); Ownership;...
   - Response variable: binary Yes/No
   - What about external validity?
2. Will the probability of a country supporting the extention of the Kyoto protocol in the upcomming Copenhagen meeting depend on BNP, recent conomic growth, Region?
   - Frame: Nations represented in Copenhagen
   - Response variable: binary Yes/No
   - What about external validity?
3. Is the probability of soccer club i winning over club j dependent additively on club-specific fixed effect parameters, i.e. on $\beta_i - \beta_j$?

Example: Mortgage denials in Boston

TABLE 11.1 Variables Included in Regression Models of Mortgage Decisions

Variable	Definition	Sample Average
Financial Variables		
P/I ratio	Ratio of total monthly debt payments to total monthly income	0.331
housing expense-to-income ratio	Ratio of monthly housing expenses to total monthly income	0.255
loan-to-value ratio	Ratio of size of loan to assessed value of property	0.738
consumer credit score	1 if no "slow" payments or delinquencies 2 if one or two slow payments or delinquencies 3 if more than two slow payments 4 if insufficient credit history for determination 5 if delinquent credit history with payments 60 days overdue 6 if delinquent credit history with payments 90 days overdue	2.1
mortgage credit score	1 if no late mortgage payments 2 if no mortgage payment history 3 if one or two late mortgage payments 4 if more than two late mortgage payments	1.7
public bad credit record	1 if any public record of credit problems (bankruptcy, charge-offs, collection actions) 0 otherwise	0.074
Additional Applicant Characteristics		
denied mortgage insurance	1 if applicant applied for mortgage insurance and was denied, 0 otherwise	0.020
self-employed	1 if self-employed, 0 otherwise	0.116
single	1 if applicant reported being single, 0 otherwise	0.393
high school diploma	1 if applicant graduated from high school, 0 otherwise	0.984
unemployment rate	1989 Massachusetts unemployment rate in the applicant's industry	3.8
condominium	1 if unit is a condominium, 0 otherwise	0.288
black	1 if applicant is black, 0 if white	0.142
deny	1 if mortgage application denied, 0 otherwise	0.120

Example: Mortgage denials in Boston

The probability of denial as function of Payment to Income ratio

- Linear regression does not work - the fitted curve exceeds the probability interval $[0, 1]$!
- Curved regression is required!

FIGURE 11.1 Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

Mortgage applicants with a high ratio of debt payments to income (*PI/ ratio*) are more likely to have their application denied (*deny* = 1 if denied, *deny* = 0 if approved). The linear probability model uses a straight line to model the probability of denial, conditional on the *PI/ ratio*.

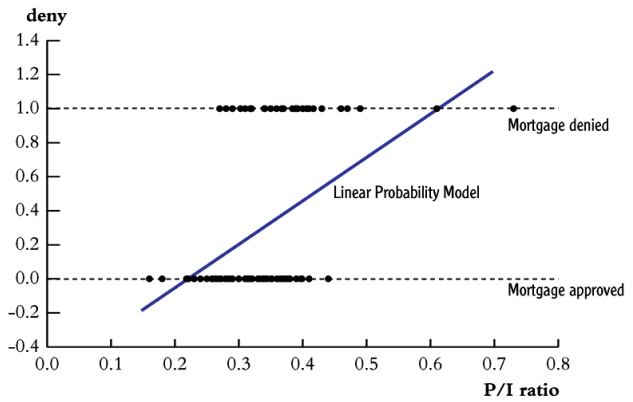
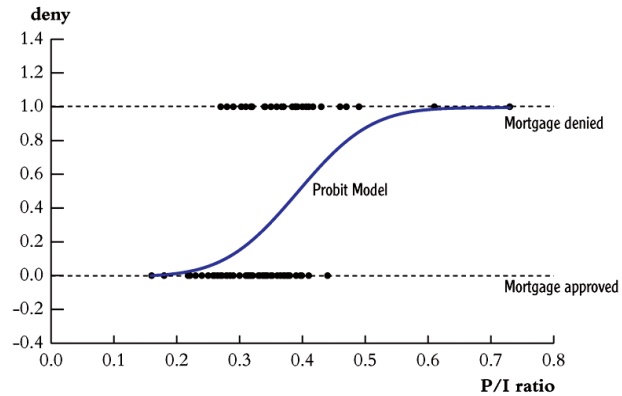


FIGURE 11.2 Probit Model of the Probability of Denial, Given the P/I Ratio

The probit model uses the cumulative normal distribution function to model the probability of denial given the payment-to-income ratio or, more generally, to model $\Pr(Y = 1|X)$. Unlike the linear probability model, the probit conditional probabilities are always between 0 and 1.



Probit and logit regression for

$P(Y = 1|X_1, \cdots, X_k)$

- The probit model consists of a linear regression within the standard normal cumulative distribution function $\Phi(z) = F_{N(0,1)}(z) = P(Z \leq z),\ Z \sim N(0,1)$:
  - $P(Y = 1|X_1, \cdots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$ (why no residual term u?)
- The logit model consists of a linear regression within the cumulative logistic distribution function $F(v) = F_{\text{logit}}(v) = \frac{e^v}{1+e^v}$:
  - $P(Y = 1|X_1, \cdots, X_k) = F(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$

Comparing the standard normal (probit) and the logistic (logit) curves

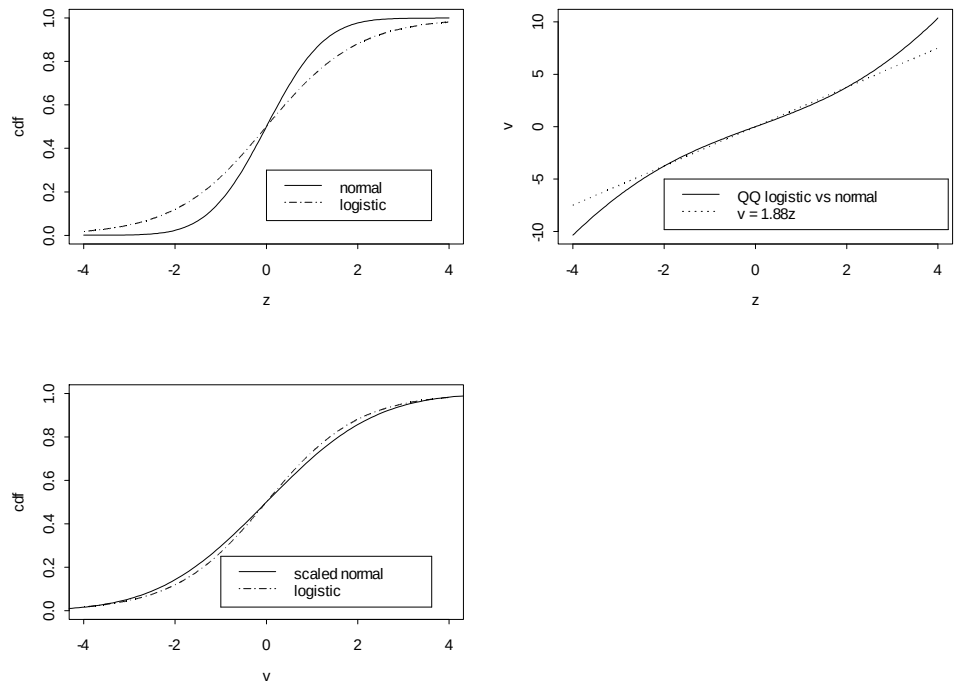


Figure: UL: The cumulative distribution functions (cdf) Φ and F_{logit}; UR: QQ-plot of the logistic versus the normal distribution, $\left(\Phi^{-1}(p), F_{\text{logit}}^{-1}(p)\right)$ $0 < p < 1$, the logistic has fatter tails than the normal!; LR: The scaled normal $\Phi(v/1.88)$ and $F_{\text{logit}}(v)$

Maximum likelihood estimation

Logistic regression

$p_i = p\left(X_{i1}, \cdots, X_k\right) = F_{\text{logit}}\left(X_{i1}, \cdots, X_{ik}\right)$ is the conditional "success" probability (denial of mortgage) for unit i

- For $y = 1$ and $y = 0$

$$P\left(Y_i = y | X_{i1}, \cdots, X_{ik}\right) = p_i^y \left(1 - p_i\right)^{1-y}$$

- By independence across units, the conditional joint outcome probability is

$$P\left(Y_1 = y_1, \cdots, Y_n = y_n\right) = \prod_{i=1}^{n} p_i^y \left(1 - p_i\right)^{1-y} = L\left(\beta_0, \beta_1, \cdots, \beta_k\right)$$

- $L\left(\beta_0, \beta_1, \cdots, \beta_k\right)$ is the likelihood function - for observed data it is a function of the parameters
- The (joint) maximum likelihood estimator $\left(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k\right)$ maximizes $L\left(\beta_0, \beta_1, \cdots, \beta_k\right)$
  - Maximum likelihood estimators are (under regularity conditions) consistent and asymptotically normally distributed with standard errors

A simple example of logistic regression

maximum likelihood estimation

Model: $P(Y = 1|x) = F_{\text{Logit}}(\beta x)$ no intercept. True value: $\beta = 0.5$,

`x= -5 -4 -3 -2 -1 0 1 2 3 4 5`

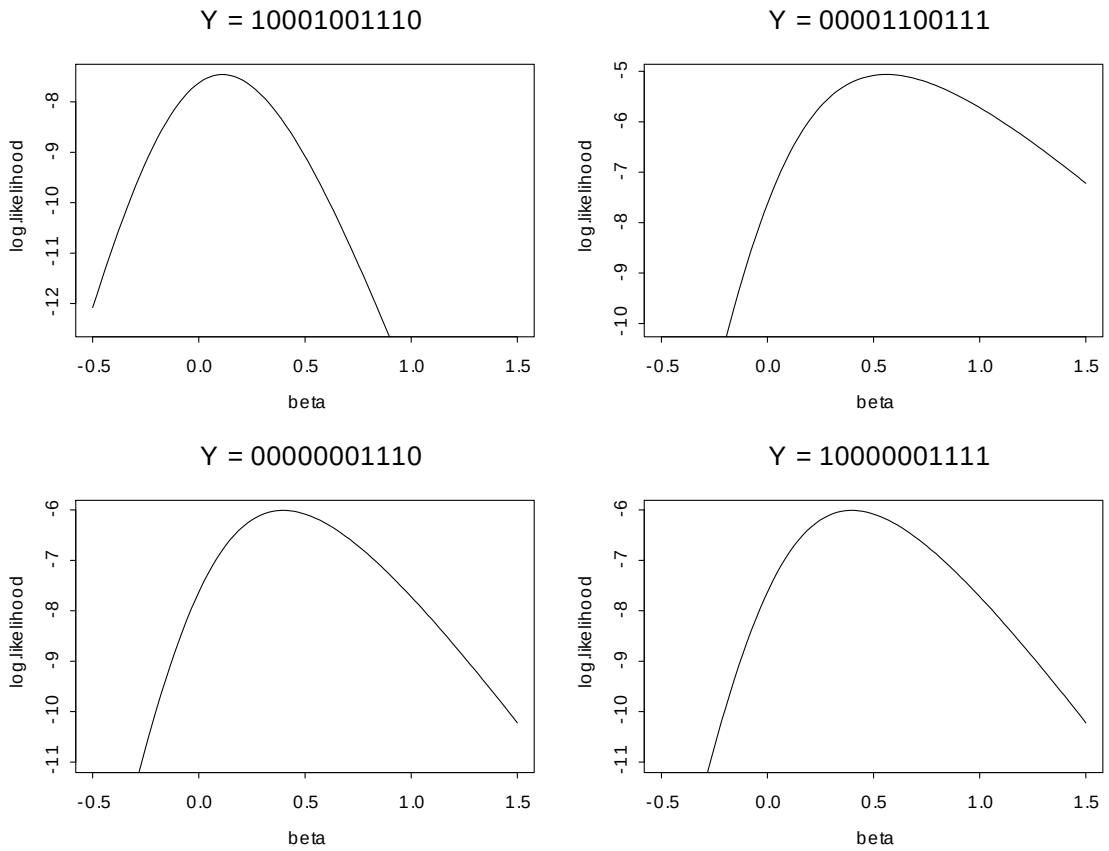


Figure: Log likelihood $\log(L)$ for four different realizations of $(Y_{-5}, \cdots, Y_5)$

Mortgage denials in Boston, logit and probit regressions

Table 11.2

TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: *deny* = 1 if mortgage application is denied, = 0 if accepted; 2380 observations.

Regression Model / Regressor	*LPM* (1)	*Logit* (2)	*Probit* (3)	*Probit* (4)	*Probit* (5)	*Probit* (6)
black	0.084** (0.023)	0.688** (0.182)	0.389** (0.098)	0.371** (0.099)	0.363** (0.100)	0.246 (0.448)
P/I ratio	0.449** (0.114)	4.76** (1.33)	2.44** (0.61)	2.46** (0.60)	2.62** (0.61)	2.57** (0.66)
housing expense-to-income ratio	−0.048 (.110)	−0.11 (1.29)	−0.18 (0.68)	−0.30 (0.68)	−0.50 (0.70)	−0.54 (0.74)
medium loan-to-value ratio (0.80 ≤ *loan-value ratio* ≤ 0.95)	0.031* (0.013)	0.46** (0.16)	0.21** (0.08)	0.22** (0.08)	0.22** (0.08)	0.22** (0.08)
high loan-to-value ratio (*loan-value ratio* ≥ 0.95)	0.189** (0.050)	1.49** (0.32)	0.79** (0.18)	0.79** (0.18)	0.84** (0.18)	0.79** (0.18)
consumer credit score	0.031** (0.005)	0.29** (0.04)	0.15** (0.02)	0.16** (0.02)	0.34** (0.11)	0.16** (0.02)
mortgage credit score	0.021 (0.011)	0.28* (0.14)	0.15* (0.07)	0.11 (0.08)	0.16 (0.10)	0.11 (0.08)
public bad credit record	0.197** (0.035)	1.23** (0.20)	0.70** (0.12)	0.70** (0.12)	0.72** (0.12)	0.70** (0.12)
denied mortgage insurance	0.702** (0.045)	4.55** (0.57)	2.56** (0.30)	2.59** (0.29)	2.59** (0.30)	2.59** (0.29)
self-employed	0.060** (0.021)	0.67** (0.21)	0.36** (0.11)	0.35** (0.11)	0.34** (0.11)	0.35** (0.11)
single				0.23** (0.08)	0.23** (0.08)	0.23** (0.08)
high school diploma				−0.61** (0.23)	−0.60* (0.24)	−0.62** (0.23)
unemployment rate				0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
condominium					−0.05 (0.09)	
black × *P/I ratio*						−0.58 (1.47)
black × *housing expense-to-income ratio*						1.23 (1.69)
Additional credit rating indicator variables	no	no	no	no	yes	no
constant	−0.183** (0.028)	−5.71** (0.48)	−3.04** (0.23)	−2.57** (0.34)	−2.90** (0.39)	−2.54** (0.35)

(Table 11.2 continued)

(Table 11.2 continued)

***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**

	(1)	(2)	(3)	(4)	(5)	(6)
Applicant single; HS diploma; industry unemployment rate				5.85 (< 0.001)	5.22 (0.001)	5.79 (< 0.001)
Additional credit rating indicator variables					1.22 (0.291)	
Race interactions and black						4.96 (0.002)
Race interactions only						0.27 (0.766)
Difference in predicted probability of denial, white vs. black (percentage points)	8.4%	6.0%	7.1%	6.6%	6.3%	6.5%

These regressions were estimated using the $n = 2380$ observations in the Boston HMDA data set described in Appendix 11.1. The linear probability model was estimated by OLS, and probit and logit regressions were estimated by maximum likelihood. Standard errors are given in parentheses under the coefficients and *p*-values are given in parentheses under the *F*-statistics. The change in predicted probability in the final row was computed for a hypothetical applicant whose values of the regressors, other than race, equal the sample mean. Individual coefficients are statistically significant at the *5% or **1% level.

Deviance for evaluating fit

- To see whether additional explanatory variables improve the fit significantly, we should for each model have the resulting

$$\textit{deviance} = -2 \log \left(\textit{likelihood} \right)$$

  - The deviance will never increase when additional explanatory variables are introduced
  - The deviance difference from one fitted model to an extended model is approximately χ^2-distributed with df = additional free regression parameters, provided none of them has a real effect (null distribution).

Mortgage denials in Boston

Are black applicants discriminated?

- The interactions in model 6 makes interpretation more involved
- Recalling that logit regression parameters are about 1.88 as large as probit regression parameters, models 2-5 give roughly the same answer: the probit estimate for black is about 0.38.
  - Everything else equal, the difference in predicted (fitted) value on the probit scale for a black applicant is moved 0.38 units to the right relative to that for a white applicants.
  - In terms of probability of denial, the difference depends on where on the scale the two points are. If the two applicants have average values on all other variables, the black has 6-7 percentage points higher probability of being denied.

Logistic regression is easier to interpret!

Let p_{black} and p_{white} be the denial probabilities for the black and the white applicant that otherwise are identical. In logistic regression the log odds is for the black person

$$\log\left(O_{black}\right) = \log\left(\frac{p_{black}}{1-p_{black}}\right) = \beta_{black} + \sum_j \beta_j X_j$$

where the sum is over all other regressors including the intercept.
The log odds for the white person is

$$\log\left(O_{white}\right) = \log\left(\frac{p_{white}}{1-p_{white}}\right) = \sum_j \beta_j X_j.$$

The log odds ratio is thus

$$\log\left(\frac{O_{black}}{O_{white}}\right) = \log\left(O_{black}\right) - \log\left(O_{white}\right) = \beta_{black}$$

Regardless of the other explanatory variables for the two otherwise identical black and white, the log odds ratio is 0.7 for black versus white (model 2).

Summing up

- OLS is stupid when the response variable is dichotomous (a dummy)
- Use probit or logit regression! Probit is fashionable in econometrics, but logistic regression is slightly preferable for statistical and interpretational reasons
- Probit regression and logit regression are cases of the generalized linear model (**glm**) with Bernoulli (binomial) **variational model** and with the probit or logit **link function** mapping the linear predictor to the expected response: $E(Y) = F_{\text{Logit}}\left(\sum_j \beta_j X_j\right)$ etc. See SW: Appendix 11.3, and http://en.wikipedia.org/wiki/Generalized_linear_model
- Individual coefficients can be tested by t-test. To test whether some parameters are all zero, compare the deviance difference to the appropriate χ^2-distribution.
- Do Exam questions from 2002.