Lecture notes to Chapter 11, Regression with binary dependent variables - probit and logit regression

Tore Schweder

November 1, 2010

Outline

- Examples of binary respons variables
- Probit and logit - examples of generalized linear modelling (glm)
- Example: crime rates in Norway by country of birth - a comparisson
- The maximum likelihood method - a general method to fitting models to data
- Application to Boston Home Mortgage Data: is there a race bias in mortgage denials?

Binary respons variables, examples and issues A nominal respons variable with 2 categories is a binary variable. It could be coded YES/NO, SUCCESS/FAILURE, 0 1, ... In addition to the binary response variable Y there might be explanatory variables X_1, X_2, $\cdots$, X_k.

1. Does probability of bancrupcy in 2010 in Norway depend on branch?
   (a) Population: firms registred in Norway by 01.01.2010
   (b) A sample from the registry
   (c) explanatory variables: by 01.01.2010: branch, size of firm, type of firm,...
   (d) Problem: given the covariates, is outcome (binary Y) independent across firms?
   (e) Result: estimated bankrupcy probability by branch for a firm of typical size and type.
2. Will children of mothers that spend more time with child during its first year of life have less chance of dropping out of high school? ESOP seminar 01.11.2010: http://www.econ.yale.edu/seminars/labor/lap08/carneiro-081031.pdf

(a) Responsvariable: child drops out (YES/NO)

(b) Population: Norwegian children born around the reform in July 1977, when 18 weeks of paid maternity leave, in addition to 4 months of unpaid leave, was introduced for working (eligible) mothers.

(c) Sample: all children born before and after the reform (from registry)

(d) explanatory variable: birth before/after, mothers income, length of leave, eligible,...

(e) Result: predicted reduction in drop-out probability when mother took full leave

(f) Problems: Did some mothers postpone birth to after reform? Is their seasonal/period effects confounded with effect of reform?...

3. Does crime rate depend on country of bith for residents of Norway?

(a) Response: individual charged within 2009 (YES/NO)

(b) Population: Inhabitants in Norway

(c) Sample: Inhabitants in Norway 01.01.2009

(d) Covariates: Country of birth (categorical), sex, age, ...

(e) Result: Estimated charge probability by country of birth, for individuals of same sex, age, ...

(f) Problems: independence across individuals? Omitted variables?, ...

4. Is probability of mortgage denial dependent on the race of the applicant (black/caucasian...)? Stock and Watson 11.4

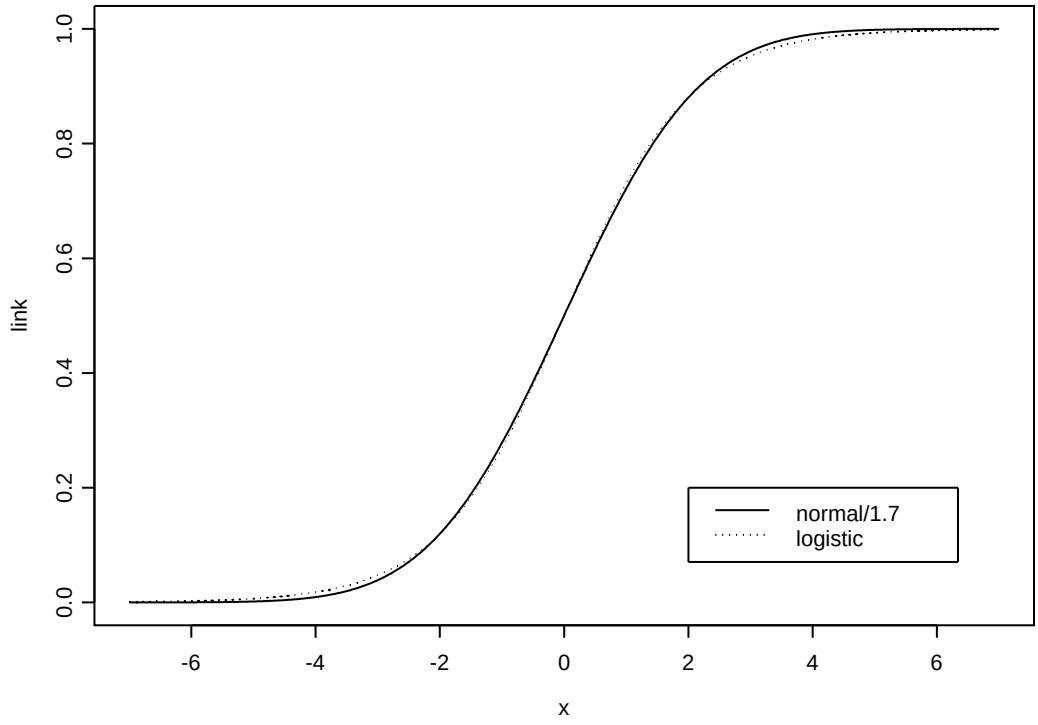
Generalized linear model: probit and logit regression A model where the expected response is a non-linear function of a linear function of the covariates

$$EY = g(L) = g\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\right)$$

When g is the identity, $g(L) = L$, we have ordinary linear regression that can be fitted by least squares. When Y is a Bernoulli variable, say coded 0 and 1, such that $P(Y = 1) = p = 1 - P(Y = 0)$, then $EY = p$, $0 \leq p \leq 1$, linear regression is no good (SW fig 11.1). A curved link function g is needed (SW fig 11.2). Any cumulative distribution function for a distribution on the whole real line can be used as link function.

Probit regression The link function $g(L) = \Phi(L)$, where Φ is the cumulative distribution function of the standard normal distriburtion. The observed responses $Y_1, \cdots, Y_n$ are assumed independen and Bernoulli distributed given the covariate variables. The regression coefficients $\beta_0 \cdots \beta_k$ are (optimally) estimated by the maximum likelihood method.

Logit regression The link function is $g(L) = p = \frac{e^L}{1+e^L}$, which is the cdf of the logistic distribution with density $f(x) = \frac{e^x}{(1+e^x)^2}$. Independent responses, given covariates. Logistic regression is more easily interpreted than probit regression since the regression coefficients are log odds ratios, but the models are very similar, see graph. The tails are slightly heavier in the logistic distribution than the normal. In the central part, $\frac{e^L}{1+e^L} \simeq \Phi(L/1.7)$



The odds for an individual is

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}$$

If, say covariate X_1 is changed with an ammount dx, $X_1 \to X_1 + dx$, while everything else is kept constant, the response probability is changed to p_d and the odds is changed to

$$\frac{p_d}{1-p_d} = \frac{p}{1-p} e^{\beta_0+\beta_1(X_1+dx)+\cdots+\beta_k X_k} / e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k} = \frac{p}{1-p} e^{\beta_1 dx}.$$

Thus,

$$\beta_1 dx = log\left(\frac{p_d/(1-p_d)}{p/(1-p)}\right)$$

is the log odds ratio of the change.

Maximum likelihood Independent observations $Y_1, \cdots, Y_n$has in general joint probability density/probability $\prod f_i(y_i, \beta)$ when f_i is the probability density of Y_i depending on the common parameter (vector) β. The maximum likelihood estimator (the MLE) is the value of β that maximizes the likelihood $\prod f_i(y_i, \beta)$ as a function of β given the observed data $(y_1, \cdots, y_n)$. It is the value of the parameter that makes the observations maximally probable.

Instead of maximizing the likelihood function, it is convenient to maximizing the log-likelihood function

$$l(\beta; y_1, \cdots, y_n) = \ln\left(\prod f_i(y_i, \beta)\right) = \sum_{i=1}^{n} \ln\left(f_i(y_i, \beta)\right)$$

Example: normally distributed data For $Y_1, \cdots, Y_n$independent normally distributed $N(\mu, \sigma)$ where σ is known and μ is the parameter to estimate. The likelihood function is $\prod_i^n \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right)\right)$ making the log-likelihood

$$l(\mu) = n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2.$$

Maximizing log-likelihood is for normally distributed data equivalent to minimizing the sum of squares $\sum_i^n (y_i - \mu)^2$!

In the linear normal regression model with $EY_i = \mu_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$ and OLS is MLE.

Optimality, consitency and asyptotic normality Under mild assumptions (as those in SW for least linear regression and logit/probit regression) the MLE is consistent and asymptotically normally distributed around the true value of the (vector-)parameter β with standard errors obtained from the matrix of second derivatives of the log-likelihood function. No estimator has smaller standard deviations than the MLE (in a certain sense) in the limit ($n \to \infty$, β fixed).

Logit and probit regression is easy and fast to run in Stata, R and other statistical systems. The maximum of the log-likelihood function is found by numerical optimization, and standard errors are obtained directly. Other types of glm's, such as Poisson regression (Poisson ditributed response, $EY_i = \exp\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\right)$) are also easy and fast to run.

Pseudo-R^2 and deviance The deviance function $D(\beta) = 2(l(\hat{\beta}) - l(\beta))$ measures how far away the value β is from the MLE $\hat{\beta}$. Let β satisfy r restrictions, and let $\hat{\beta}_{rest}$ be the restricted MLE, while $\hat{\beta}$ is the unrestricted MLE. The deviance difference

$$D_{res}(\beta) - D(\beta) = 2(l\left(\hat{\beta}\right) - l\left(\hat{\beta}_{res}\right))$$

is close to χ^2 distributed with r degrees of freedom. The deviance difference is also called twice the log likelihood ratio.

In the pseudo-R^2formula SW (11.18), $\ln\left(f^{max}_{Bernoulli}\right)$ is the log-likelihood with only an intercept parameter $L_i = \beta_0$.

Logistic regression With no covariates, $L = \beta_0$and logit and probit (regression) is identical. The problem is then to estimate a common p. See SW page 416.

MLE is a bit simpler for logit than for probit regression since as in SW (11.16),

$$f_i(y_i; p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = (1-p_i)\left(\frac{p_i}{1-p_i}\right)^{y_i}.$$

The log-likelihood function is then

$$l(\beta) = \sum_i \left[-\ln\left(1+e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}}\right) + y_i\left(\beta_0+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}\right)\right].$$

Problem: Does crime rate depend on country of bith for residents of Norway? Based on registre data for residents in Norway, with observed respons (YES, the individual was charged in 2009, or NO) and with observed covariates (sex, age) for each individual, which logistic regression model would you fit, and how would you use the estimated regression parameters to obtain a fair comparison of charge probabilities for natives and imigrants by country of birth. (Hint: with dummies for country, except for natives, the country effect is the log odds rate between individuals from that contry and native Norwegians given that the individuals are equal with respect to covariate values.)