Binary outcome regression models

B.D. in Business Administration and Economics
Course in Quantitative Methods III

Rosario Barone
University of Rome “Tor Vergata”

rosario.barone@uniroma2.it

Why Generalized Linear Models?

Let suppose to be in the ordinary linear regression framework, such that:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

in cases where the response variable Y is expected to be always positive and varying over a wide range, these assumptions turn out to be inappropriate.

The GLM is a flexible generalization of ordinary linear regression.

GLM routine

More specifically, GLMs are generalization of linear models for situations in which the outcome is not Gaussian, summarized as follows:

- specify distribution for the dependent variable $f(Y|\theta)$;
- specify a link function $g(\cdot)$;
- specify a linear predictor.

Assumption on Y

The distribution of the dependent variable $f(Y|\theta)$ is assumed to belong to the exponential family. Some examples:

- Normal
- Poisson
- binomial (with fixed n)
- multinomial (with fixed n)
- negative binomial (with fixed number of failures).

*Note tath the parameters which must be fixed determine a limit on the size of observations.

Model definition

We define the distribution $f(Y|X)$, with mean $\boldsymbol{\mu}$ of the depending on the independent variables, X, through:

$$\mathrm{E}(Y|X) = \boldsymbol{\mu} = g^{-1}(X\boldsymbol{\beta})$$

where:

- $E(Y|X)$ is the expected value of Y conditional on X;
- $X\boldsymbol{\beta}$ is the linear predictor;
- g is the link function.

The variance is typically a function, V, of the mean:

$$\mathrm{var}(Y|X) = \nu(g^{-1}(X\boldsymbol{\beta})).$$

However, by choosing ν as a distribution of the exponential family we get a more flexible model.

Binary Outcome dependent variable

Let Y denote a binary response variable ($Y \in \{0, 1\}$) and let $\mathbf{x} = (x_1, \ldots, x_k)$ be the vector of observed covariates.

We denote with

- $\pi(\mathbf{x})$ the mean $E(Y|X) = P(Y = 1)$, for underlining its dependence on the covariates $\mathbf{x}$;
- $\text{var}(Y) = \pi(\mathbf{x})(1 - \pi(\mathbf{x}))$.

We present three GLMs for binary data:

- Linear probability model
- Logit model
- Probit model

Linear probability model

A *linear probability model* is a GLM with binomial random component Y and *identity link* function

$$\pi(x) = \alpha + \beta x$$

A structural problem due to the identity link:

1. linear functions take values over the entire real line;
2. $\pi(x) \in [0, 1]$ (it is a probablity);
3. for sufficiently large or small x, $\pi(x)$ falls outside the $[0, 1]$ interval.

Linear probability model

In the multiple predicor extension ($\mathbf{x} = (x_1, \ldots, x_k)$)

$$\pi(\mathbf{x}) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

we have the same fitting problems as in the univariate case:

- $\hat{\pi}(\mathbf{x})$ may fall outside the range $[0, 1]$ for some observed individuals.
- The model can be valid over a restricted range of x values.

So, what's the andantage with this model?

The andantage is its **simple interpretation**: β represents the increment in $\pi(x)$ as x increases of one-unit.

Since the model is *linear*, one may think to estimate it via *Ordinary Least Squares* instead of *MLE*. Is it a good guess?

1. *OLS* assumes constant variance: condition not satisfied;
2. the binomial ML estimator is more efficient than OLS.

However, *OLS* and *MLE* estimates are similar when $\hat{\pi}(x)$ is in the range within which the variance is relatively stable.

Logit model

Usually, binary data result from a nonlinear relationship between $\pi(x)$ and x, that is for a fixed change in x, there is:

- lower impact if $\pi(x)$ is close to 0 or 1
- higher impact if $\pi(x)$ lies in a neighborhood of 0.5

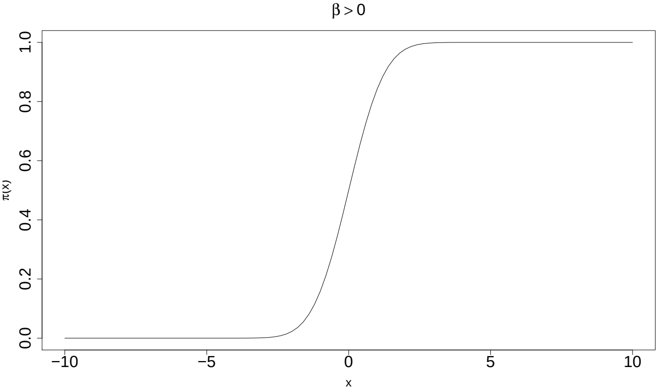
Logit model

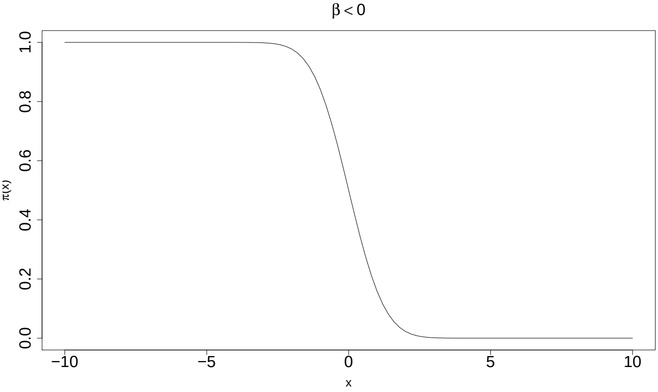
Let Y be a binary variable and let x be an observed covariate. We define the *Logistic regression model* or equivalentely *Logit model* as:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

As $x \to \infty$

- if $\beta < 0$, $\pi(x)$ gets closer to 0;
- if $\beta > 0$, $\pi(x)$ gets closer to 1.


β > 0
1.0
0.8
0.6
0.4
0.2
0.0
π(x)
−10
−5
0
5
10
x


β < 0
1.0
0.8
0.6
0.4
0.2
0.0
π(x)
−10
−5
0
5
10
x

Logit link

From the model definition, the *odds* of the logistic regression are:

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x)$$

thus, the *log-odds* has the linear relationship with the covariate

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x.$$

This is also called *logit link*.

Interpretation: hard challenge

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x.$$

How to interpret β?

- its sign determines whether $\pi(x)$ is increasing or decreasing as x increases;
- the rate of climb or descent increases as β increases;
- as $\beta \to 0$ the curve flattens to a horizontal straight line;
- when $\beta = 0$, Y is independent of X;

Interpretation: the odds increases multiplicatively by e^{β} as x increases of 1-unit.

"Approximating" the intepretation

Most scientists are not familiar with odds or logits. Two solutions proposed:

- Linear approximation (Berkson 1951);
- calculate $\pi(x)$ at certain x values.

Summary and properties

Logistic regression models (Logit models) are GLMs with:

- binary outcome variable;
- logit link function.

Advantages of logit models:

- the logit link is the natural parameter of the binomial distribution (canonical link);
- the logit link can be any real number and $\pi(x)$ always belong to $[0, 1]$.

Probit model

By defining a binary response having form $\pi(x) = F(x)$ for some cdf F permits the curve to be more flexibile.

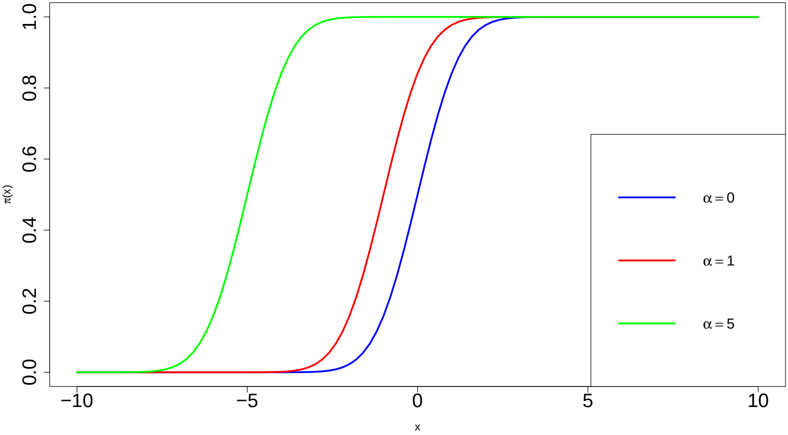
Let assume to use a standard Normal cdf Φ (N(0,1)) to define a model

$$\pi(x) = \Phi(\alpha + \beta x).$$

Shapes of different cdf's in the class occur as α and β vary.

- β controls the rate of increasing (if $\beta > 0$) or decreasing (if $\beta < 0$) of the cdf;
- α controls the location of the curve.

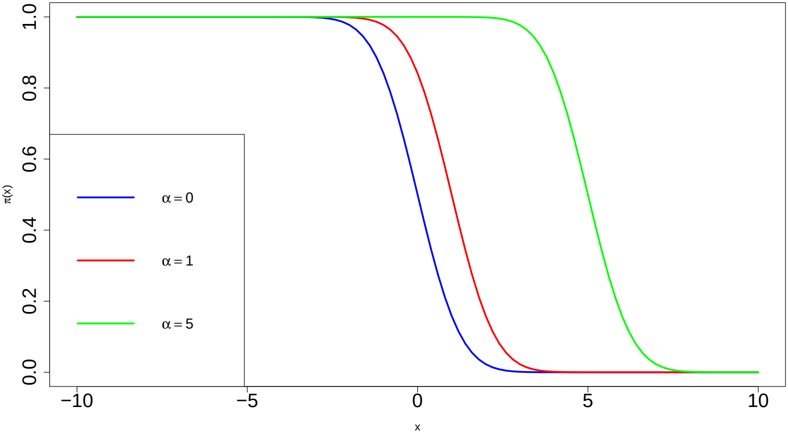
With $\beta > 0$, as α varies...



With $\beta < 0$, as α varies...

Binary outcome regression models

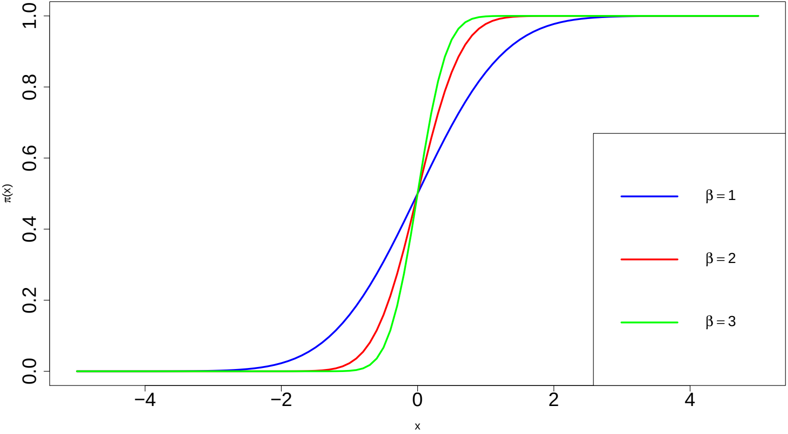
Rosario Barone



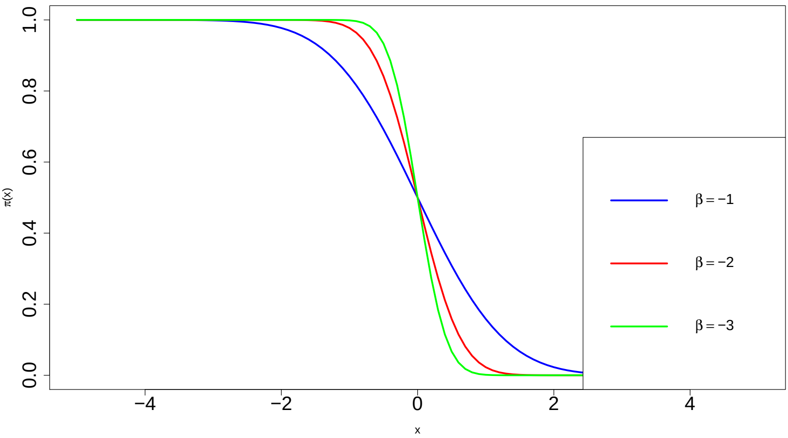
With $\alpha = 0$, as β varies in $[0, \infty]$...

Binary outcome regression models

Rosario Barone



With $\alpha = 0$, as β varies in $[-\infty, 0]$...



Probit model

When Φ is stricly increasing over $\mathbb{R}$ its inverse function exists and

$$\Phi^{-1}(\pi(x)) = \alpha + \beta x$$

is called *Probit model*. Here Φ^{-1} (the quantile function of the standard Normal distribution) is the link function.

With this model setting, β indicates how much the (conditional) probability of the outcome variable changes when you change the value of x.

Probit to logit

As for probit models, we may consider the logistic regression curve as the cdf of the *logistic distribution*, having expression

$$F(x) = \frac{\exp\left((x-\mu)/\tau\right)}{1+\exp\left((x-\mu)/\tau\right)},$$

where μ is the mean and $\tau > 0$ is the dispersion parameter.

The standardized logistic distribution ($\mu = 0$ and $\tau = 1$) is then:

$$\Phi(x) = \frac{e^x}{1+e^x}$$

Probit to logit

Hence, the logit model is

$$\pi(x) = \Phi(\alpha + \beta x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

Therefore, the logit transofrmation is simply the quantile function (inverse cdf) for the standard logistic distribution

$$x = \Phi^{-1}(\pi(x)) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right).$$

Moments for Binomial GLM

Let consider $n_i Y_i \sim \text{Bin}(n_i, \pi_i)$. Then, y_i is the sample proportion of successes for n_i trials. We define the moments of the Binomial GLM as

- $E(Y_i) = \pi_i$
- $var(Y_i) = \pi_i(1 - \pi_i)/n_i$

Reminder: likelihood equations for GLM

For n independent observations, the likelihood function is:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \log(f(y_i; \theta_i, \psi))$$

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)$$

After some analytics, we get the *likelihood equations*:

$$\frac{\mathcal{L}(\beta)}{\partial\beta} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)} \frac{\partial\mu_i}{\partial\eta_i} = 0.$$

Likelihood equations for Binomial GLM

Let consider $n_i Y_i \sim \text{Bin}(n_i, \pi_i)$. Then, y_i is the sample proportion of successes for n_i trials. Then,

$$\pi_i = \Phi\left(\sum_{j=1}^{k} \beta_j x_{ij}\right)$$

with Φ beeing the standard cdf of some classes of continuous distributions.

We know that

$$\mu_i = \pi_i = \Phi\left(\sum_{j=1}^{k} \beta_j x_{ij}\right) = \Phi(\eta_i).$$

Likelihood equations for Binomial GLM

Then

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \Phi(\eta_i)}{\partial \eta_i} = \frac{\partial \Phi\left(\sum_{j=1}^{k} \beta_j x_{ij}\right)}{\partial \eta_i} = \phi\left(\sum_{j=1}^{k} \beta_j x_{ij}\right)$$

where $\phi = \partial \Phi(u)/\partial u$.

Therefore, the likelihood equations for the binomial GLM are:

$$\frac{\mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{n_i(y_i - \pi_i)x_{ij}}{\pi_i(1-\pi_i)} \phi\left(\sum_{j=1}^{k} \beta_j x_{ij}\right) = 0.$$

Overdispersion for Binomial GLMs and Quasi-likelihood

The quasi-likelihood approach can handle overdispersion for counts based on binary data.

As shown before,

- $E(Y_i) = \pi_i$;
- $var(Y_i) = \pi_i(1 - \pi_i)/n_i$.

A simple quasi-likelihood approach uses the alternative variance function

$$\nu(\pi_i) = \phi\pi_i(1 - \pi_i),$$

overdispersion occours when $\phi > 1$. Estimates are equal to the *ML* case for the Binomial response (ϕ drops out from likelihood equations and it is estimated separately) and the standard errors multiply by $\sqrt{\phi}$.

Binary GLMs diagnostics

- Deviance of the model
- Likelihood ratio
- Statistics on the residuals (RSS-like statistics):
    - deviance residuals
    - Pearson residuals

- Logit: `glm(formula, family = binomial(link = "logit"), data, ...)`

- Probit: `glm(formula, family = binomial(link = "probit"), data, ...)`