Discrete Response Model
Lecture 2

datascience@berkeley

Introduction to Binary Response Models and Linear Probability Model

Response Probability

In a binary response model, we model the response probability

$$P(y_i = 1|\mathbf{x_i}) = P(y = 1|x_{1,i}, x_{k,i}, \ldots, x_{k,i})$$

where

y_i is a $n \times 1$ vector of response, which takes the value of 0 or 1

$x_{j,i}$ is a $n \times 1$ vector of covariates (or explanatory variables) associated with individual $i, i = 1, 2, \ldots, n$

Many authors use π to denote the response probability, and this is the notation used in our text. So, I will follow this notation in this course.

$$P(y_i = 1|\mathbf{x_i}) = \pi_i(\mathbf{x_i})$$

To the precise, this is a conditional response probability, conditional on the set of explanatory variables $\mathbf{x_i}$ of individual i.

General Formulation of the Class of Binary Response Model

Suppressing the individual subscript for the time being:

$$P(y = 1|\mathbf{x}) = F(\beta_0 + \mathbf{x}\beta)$$

where $F()$ is a function taking on values between 0 and 1: $0 < F(z) < 1$, for all real numbers z.

Various nonlinear transformation have been proposed for the function $F(.)$ to make sure that probabilities are between 0 and 1.

Logit and Probit Specification of Binary Response Models

The logit transformation uses the logistic function, resulting in the following form. We will cover logistic regression model in great details in this course.

$$F(z) = \Lambda(z) = \frac{exp(z)}{1 + exp(z)}$$

This is the for a standard logistic random variable.

The probit transformation, resulting in the probit model, uses the standard normal distribution and takes the following form:

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv$$

where $\phi(z)$ is the standard normal PDF:

$$\boxed{\phi(z) = (2\pi)^{-1/2} exp(-z^2/2)}$$

Linear Probability Model

Before studying the logistic regression model, let's examine using the classical linear regression covered in $w203$ to binary response variable.

Recall that linear regression model takes the following form (suppressing the individual subscript):

$$y = \mathbf{x}\beta + \epsilon = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

where

$\mathbf{x}$ is a $1 \times k$ covariate vector (associated with individual i) and y is a $k \times 1$ response vector.

When y is a binary variable taking on values 0 and 1, we have

$$\boxed{E(y|\mathbf{x}) = P(y|\mathbf{x})}$$

This means that the probability of success (we are defining that the event $y = 1$ as "success"), $\pi(\mathbf{x}) = P(y|\mathbf{x})$ is a *linear function* of the explanatory variables, $x's$.

Berkeley SCHOOL OF INFORMATION