Binary Outcome Variables

Edicson Luna

University of Pennsylvania

Motivation

Linear Probability Model

Probit and Logit

Interpreting the coefficients

Motivation

A different kind of outcome

So far, our outcome variable y_i has always been *continuous*: a wage, a grade, the number of hours worked.

But many of the most interesting economic questions have a binary answer. The outcome is a "yes" or a "no", which we code as

$$y_i \in \{0, 1\}.$$

Today we ask: what changes when we want to explain a variable like this?

Binary outcomes are everywhere

In applied work, dummy outcomes appear constantly:

- Is the person employed? (1) or unemployed? (0)
- Did the household default on a loan? (1) or not? (0)
- Is the worker in the informal sector? (1) or formal? (0)
- Did the individual participate in the labor force? (1/0)
- Was the loan application approved? (1/0)

In every case we would like to know how some covariate x_i (education, income, age, ...) moves the outcome.

The expectation becomes a probability

Here is the key observation. If y_i takes only the values 0 and 1, then its expectation is

$$\mathbb{E}[y_i] = 1 \cdot \mathbb{P}(y_i = 1) + 0 \cdot \mathbb{P}(y_i = 0) = \mathbb{P}(y_i = 1).$$

Conditioning on the regressors, the same algebra gives

$$\mathbb{E}[y_i \mid x_i] = \mathbb{P}(y_i = 1 \mid x_i).$$

So modeling the conditional expectation of a dummy is exactly the same as modeling the conditional probability of success.

The response probability

This motivates our central object, the response probability

$$p(x) := \mathbb{P}(y_i = 1 \mid x_i = x) = \mathbb{E}[y_i \mid x_i = x].$$

Everything in this class is about choosing a functional form

for $p(x)$ and interpreting it. Two natural requirements appear immediately:

- As a probability, we would like $p(x) \in [0, 1]$.
- We want to read off how a change in x_j moves that probability.

We will study three models: the LPM, the Probit and the Logit.

Linear Probability Model

The Linear Probability Model

The simplest idea: just run OLS with the dummy y_i on the left-hand side.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i.$$

Taking the conditional expectation, and using that

$\mathbb{E}[y_i \mid x_i] = \mathbb{P}(y_i = 1 \mid x_i)$,

$$p(x_i) = \mathbb{P}(y_i = 1 \mid x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

We model the probability as a *linear* function of the covariates. This is why it is called the Linear Probability Model (LPM).

Interpreting the LPM

Because the probability is linear, the coefficients are wonderfully easy to read. For a continuous x_j,

$$\frac{\partial \, \mathbb{P}(y_i = 1 \mid x_i)}{\partial x_{ij}} = \beta_j.$$

So β_j is the change in the probability of success for a one-unit increase in x_j, holding the rest fixed.

Example. If $y_i = 1$ denotes being employed and x_j is years of schooling, $\beta_j = 0.03$ means one extra year of schooling raises the probability of employment by 3 percentage points.

The LPM has problems (I)

The convenience comes at a cost.

- **Predictions out of bounds.** Nothing forces $\beta_0 + \beta_1 x_{i1} + \dots$ to lie in $[0, 1]$. For extreme values of x, the fitted "probability" can be negative or above one.

- **Constant marginal effect.** The effect β_j is the same everywhere. But a change that moves someone from 0.05 to 0.08 should not have the "same room" as one starting from 0.98. Probabilities are bounded; a straight line is not.

The LPM has problems (II)

Heteroskedasticity is built in. Given x_i, the outcome y_i is Bernoulli with success probability $p(x_i)$. Hence

$$\text{Var}(y_i \mid x_i) = p(x_i)\big(1 - p(x_i)\big),$$

which *depends on* x_i. The error is heteroskedastic by construction.

The practical fix is the one you already know: use heteroskedasticity-robust standard errors. The estimator $\hat{\beta}$ is still consistent, and inference goes through by the CLT (errors need not be normal).

So why keep the LPM?

Despite its flaws, the LPM is used all the time, and for good reasons:

- It is just OLS, so it is simple and fast.
- Coefficients *are* the partial effects, directly in probability units.
- Near the center of the data it is often a good approximation.
- It handles many regressors and fixed effects with no extra effort.

Still, the out-of-bounds predictions push us to look for a functional form that respects $p(x) \in (0, 1)$.

Probit and Logit

Bounding the probability

We want a model where the probability can never escape $(0, 1)$.

The idea: keep the linear index $x_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, but pass it through a function $G(\cdot)$ that squashes the whole real line into $(0, 1)$:

$$\mathbb{P}(y_i = 1 \mid x_i) = G(x_i'\beta), \qquad G : \mathbb{R} \to (0, 1).$$

A natural choice for G is a cumulative distribution function (CDF): it is increasing and takes values in $(0, 1)$.

Probit and Logit

Two standard choices for G give the two workhorse models:

- **Probit:** G is the standard normal CDF Φ,

$$\mathbb{P}(y_i = 1 \mid x_i) = \Phi(x_i'\beta)\,.$$

- **Logit:** G is the logistic CDF Λ,

$$\mathbb{P}(y_i = 1 \mid x_i) = \Lambda(x_i'\beta)\,, \qquad \Lambda(z) = \frac{e^z}{1+e^z}.$$

Both Φ and Λ are smooth, increasing, S-shaped curves bounded between 0 and 1. The probability now *cannot* leave $(0, 1)$.

Where does this come from? A latent variable

A useful story. Imagine an unobserved latent variable, e.g. a propensity or a net utility,

$$y_i^* = x_i'\beta + e_i,$$

and we only observe the decision

$$y_i = \mathbf{1}\{y_i^* > 0\}.$$

Then, if e_i is symmetric with CDF G,

$$\mathbb{P}(y_i = 1 \mid x_i) = \mathbb{P}(e_i > -x_i'\beta \mid x_i) = G(x_i'\beta).$$

A *normal* e_i gives the Probit; a *logistic* e_i gives the Logit.

Estimation: maximum likelihood

Since $p(x) = G(x_i'\beta)$ is *nonlinear* in β, we no longer use OLS. We use maximum likelihood.

Each observation contributes its probability of what we actually saw,

$$G(x_i'\beta)^{y_i} \left(1 - G(x_i'\beta)\right)^{1-y_i},$$

and we choose $\hat{\beta}$ to maximize the log-likelihood

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \Big[y_i \log G(x_i'\beta) + (1 - y_i) \log\big(1 - G(x_i'\beta)\big) \Big].$$

The resulting $\hat{\beta}$ is consistent and asymptotically normal.

Software does the optimization for us.

Interpreting the coefficients

The coefficient is not the effect

In the LPM, β_j *was* the partial effect. In Probit/Logit this is no longer true, because of the nonlinear G. For a continuous x_j, the chain rule gives

$$\frac{\partial \, \mathbb{P}(y_i = 1 \mid x_i)}{\partial x_{ij}} = g(x_i'\beta) \; \beta_j, \qquad g = G'.$$

Here $g = \phi$ (normal density) for Probit and

$g(z) = \Lambda(z)\big(1 - \Lambda(z)\big)$ for Logit.

Two consequences:

- The **sign** of the effect is the sign of β_j (since $g > 0$).
- The **magnitude** depends on x_i: the same β_j implies a different effect for different individuals.

A dummy regressor

If x_j is itself discrete (say a dummy), a derivative makes no sense. We instead take the difference in predicted probabilities:

$$G(\beta_0 + \cdots + \beta_j \cdot 1 + \dots) - G(\beta_0 + \cdots + \beta_j \cdot 0 + \dots),$$

holding the other regressors fixed.

Either way, the partial effect is not a single number: it varies across individuals. So how do we report "the" effect?

Two ways to summarize the effect

Since the partial effect $g(x_i'\hat{\beta})\,\hat{\beta}_j$ changes from person to person, we collapse it into one number. Two conventions dominate:

- **PEA** — Partial Effect at the Average. Plug in the average individual:

$$\text{PEA}_j = g\left(\bar{x}'\hat{\beta}\right)\,\hat{\beta}_j.$$

- **APE** — Average Partial Effect. Compute each person's effect, then average:

$$\text{APE}_j = \frac{1}{n}\sum_{i=1}^{n} g\left(x_i'\hat{\beta}\right)\,\hat{\beta}_j.$$

PEA vs. APE

They answer slightly different questions.

- **PEA**: "the effect for the *average* individual." Fast, but the average individual may not exist. The mean of a dummy (say, 0.4 female) is not anybody.

- **APE**: "the *average* of the effects across real individuals." Generally preferred, and it handles discrete regressors cleanly.

In practice both are reported and are usually close. In Stata, `margins, dydx(*)` gives the APE, and adding `atmeans` gives the PEA.

Comparing across models

A warning and a comfort.

- Warning: the raw coefficients of LPM, Logit and Probit are *not* comparable. They live on different scales (roughly, logit $\approx 1.6\times$ probit, and both are larger than the LPM slopes).
- Comfort: once you translate everything into partial effects (APE), the three models typically deliver very similar numbers.

This is why the APE is the natural quantity to report: it is comparable and interpretable in probability units.

Takeaway

- A binary outcome turns the conditional expectation into a probability: $\mathbb{E}[y_i \mid x_i] = \mathbb{P}(y_i = 1 \mid x_i)$.
- **LPM**: simple OLS, coefficients are partial effects directly. But predictions can leave $[0, 1]$, the effect is constant, and errors are heteroskedastic (use robust SE).
- **Probit / Logit**: pass the index through a CDF G, so $p(x) \in (0, 1)$; estimated by maximum likelihood. Coefficients give the sign, not the magnitude.
- Report the APE (or PEA) to recover an interpretable, comparable effect in probability units.