Course Title: Quantitative Techniques for Economics
Course code: ECON6002
Topic: The Linear Probability Model (LPM)
Ph.D. Economics (1st Semester)

Dr. Kailash Chandra Pradhan

Mahatma Gandhi Central University,
Department of Economics

Linear Probability Model (LPM)

- We have studied so far, we have implicitly assumed that the regressand in all the regression models is a dependent variable, or the response variable.
- Y is quantitative, whereas the explanatory variables are either quantitative, qualitative (or dummy), or both.
- In this presentation, we consider several models in which the regressand itself is qualitative in nature.
- The qualitative response regression models pose interesting estimation and interpretation challenges in various areas of social sciences and medical research.
- Suppose we want to study the labor force participation (LFP) decision of adult males. Since an adult is either in the labor force or not, LFP is a **yes or no** decision. Hence, the response variable, or regressand, can take only **two values, say, 1 if the person is in the labor force and 0 if he or she is not.**
- In other words, the regressand is **a binary, or dichotomous** variable.
- Labor economics research suggests that the LFP decision is a function of the unemployment rate, average wage rate, education, family income, etc.
- The qualitative response regression models are often known as **probability models**.

Linear Probability Model (LPM)

- There are three approaches to developing a probability model for a binary response variable:

  1. The linear probability model (LPM)

  2. The logit model

  3. The probit model

- First, we will discuss the LPM.

The Linear Probability Model (LPM)

- Suppose the regression model is as follows:

  $$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (1)$$

  where X = family income and Y = 1 if the family owns a house and 0 if it does not own a house.

- The above model looks like a typical linear regression model. Here, the regressand is binary, or dichotomous, it is called a **linear probability model (LPM).**

Linear Probability Model (LPM)

- In this model, the conditional expectation of Y_i given X_i , $E(Y_i \mid X_i)$, can be interpreted as the conditional probability that the event will occur given X_i , that is, $Pr\ (Y_i = 1 \mid X_i)$.
- In our example, $E(Y_i \mid X_i)$ gives the probability of a family owning a house and whose income is the given amount X_i .
- The model assums $E(u_i) = 0$ to obtain unbiased estimators we obtain

  $E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i$ (2)
- If P_i = probability then $Y_i = 1$ (that is, the event occurs), and $(1 - P_i)$ = probability that $Y_i = 0$ (that is, that the event does not occur), the variable Y_i has the following (probability) distribution.

Y_i	Probability
0	$1 - P_i$
1	P_i
Total	**1**

Here, Y_i follows the **Bernoulli or Binomial probability distribution.**

Linear Probability Model (LPM)

- Now, by the definition of mathematical expectation, we obtain:

  $E(Y_i) = 0(1 - P_i) + 1(P_i) = P_i$ (3)

- Comparing (2) with (3), we can equate

  $E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i = P_i$ (4)

- The expectation of a Bernoulli random variable is the probability that the random variable equals 1.
- If there are n independent trials, each with a probability p of success and probability (1 − p) of failure, and X of these trials represent the number of successes, then X is said to follow the **binomial distribution.**
- The mean of the binomial distribution is **np** and its variance is **np(1 − p).** The term success is defined in the context of the problem.
- Since the probability P_i must lie between 0 and 1, we have the restriction

  $\mathbf{0 \leq E(Y_i \mid X_i) \leq 1}$

Linear Probability Model (LPM)

Non-Normality of the Disturbances u_i

- In the LPM, u_i cannot be assumed to be normally distributed; they follow the Bernoulli distribution.
- But the nonfulfillment of the normality assumption may not be so critical as it appears because we know that the OLS point estimates still remain unbiased (if the objective is point estimation, the normality assumption is not necessary).
- Besides, as the sample size increases indefinitely, statistical theory shows that the OLS estimators tend to be normally distributed.

Heteroscedastic Variances of the Disturbances

- Even if $E(u_i) = 0$ and cov $(u_i , u_j) = 0$ for $i = j$ (i.e., no serial correlation), it can no longer be maintained that in the LPM the disturbances are homoscedastic.
- In a Bernoulli distribution, the theoretical mean and variance are, respectively, p and $p(1 - p)$, where p is the probability of success (i.e., something happening), showing that the variance is a function of the mean. Hence the error variance is heteroscedastic.
- In the LPM, var $(u_i) = P_i(1 - P_i)$

Linear Probability Model (LPM)

- In the presence of heteroscedasticity, the OLS estimators, although unbiased, are not efficient; that is, they do not have minimum variance.
- But the problem of heteroscedasticity, like the problem of non-normality, is not insurmountable.
- To resolve the heteroscedasticity problem, the model (1) is to transform by dividing it through by

$$\sqrt{E(Y_i \mid X_i)[1-E(Y_i \mid X_i)} = \sqrt{P_i(1-P_i)} = say\sqrt{w_i}$$

$$\frac{Y_i}{\sqrt{w_i}} = \frac{\beta_1}{\sqrt{w_i}} + \beta_2 \frac{X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \text{.....(5)}$$

- The transformed error term in (5) is homoscedastic. Therefore, after estimating (1), we can now estimate (5) by OLS, which is nothing but the weighted least squares (WLS) with w_i serving as the weights.

Linear Probability Model (LPM)

- To estimate w_i , we can use the following two-step procedure:

  Step 1. Run the OLS regression (1) despite the heteroscedasticity problem and obtain $\hat{Y}_i$ = estimate of the true $E(Y_i \mid X_i)$. Then obtain $\hat{w}_i = \hat{Y}_i(1 - \hat{Y}_i)$, the estimate of w_i.

  Step 2. Use the estimated w_i to transform the data as shown in (5) and estimate the transformed equation by OLS (i.e., weighted least squares).

Nonfulfillment of $0 \leq E(Y_i \mid X_i) \leq 1$

- Since E(Yi | X) in the linear probability models measures the conditional probability of the event Y occurring given X, it must necessarily lie between 0 and 1.
- There is no guarantee that $\hat{Y}_i$, the estimators of E(Yi | Xi), will necessarily fulfill this restriction, and this is the real problem with the OLS estimation of the LPM.
- There are two ways of finding out whether the estimated $\hat{Y}_i$ lie between 0 and 1. One is to estimate the LPM by the usual OLS method and find out whether the estimated $\hat{Y}_i$ lie between 0 and 1.

Linear Probability Model (LPM)

- If some are less than 0 (that is, negative), $\hat{Y}_i$ is assumed to be zero for those cases; if they are greater than 1, they are assumed to be 1.
- The second procedure is to devise an estimating technique that will guarantee that the estimated conditional probabilities $\hat{Y}_i$ will lie between 0 and 1. The logit and probit models discussed later will guarantee that the estimated probabilities will indeed lie between the logical limits 0 and 1.

Questionable Value of R^2 as a Measure of Goodness of Fit

- Corresponding to a given X, Y is either 0 or 1. Therefore, all the Y values will either lie along the X axis or along the line corresponding to 1.
- Therefore, generally no LPM is expected to fit such a scatter well, whether it is the unconstrained LPM or the truncated or constrained LPM .
- The LPM estimated in such a way that it will not fall outside the logical band 0–1. As a result, the conventionally computed R^2 is likely to be much lower than 1 for such models.
- In most practical applications the R^2 ranges between 0.2 to 0.6.

Linear Probability Model (LPM)

LPM: A Numerical Example

- Suppose, we have data on home ownership Y (1 = owns a house, 0 = does not own a house) and family income X (thousands of dollars) for 40 families.

  These data the LPM estimated by OLS was as follows:

  $\hat{Y}_i = -0.9457 + 0.1021X_i$

  $(0.1228) \quad (0.0082)$

  $t = (-7.6984) \quad (12.515) \qquad R^2 = 0.8048$

- The intercept of −0.9457 gives the "probability" that a family with zero income will own a house. Since this value is negative, and since probability cannot be negative, we treat this value as zero, which is sensible in the present instance.
- The slope value of 0.1021 means that for a unit change in income (here $1000), on the average the probability of owning a house increases by 0.1021 or about 10 percent at given a particular level of income,.
- We can estimate the actual probability of owning a house from the above equation. Thus, for X = 12 ($12,000), the estimated probability of owning a house is

  $(\hat{Y}_i \mid X = 12) = -0.9457 + 12(0.1021) = 0.2795$

Linear Probability Model (LPM)

- Most of cases, some estimated values are negative and some values are in excess of 1. This is one reason that the LPM is not the recommended model.
- Even if the estimated Y_i were all positive and less than 1, the LPM still suffers from the problem of heteroscedasticity.
- Therefore, we cannot trust the estimated standard errors reported in in the estimated model. We can use the weighted least-squares (WLS) to obtain more efficient estimates of the standard errors.
- As we know that, some Y_i are negative and some are in excess of one, the w_i hat corresponding to these values will be negative. Thus, we cannot use these observations in WLS , therefore, the number of observation will be reduced.
- The above estimated model can be re-estimated by using WLS method and the results is as follows.

$$\frac{\hat{Y}_i}{\sqrt{w_i}} = -1.2456\frac{1}{\sqrt{w_i}} + 0.1196\frac{X_i}{\sqrt{w_i}_i}$$

(0.1206) (0.0069)

t = (−10.332) (17.454) $R^2 = 0.9214$

Linear Probability Model (LPM)

- These results show that the estimated standard errors are smaller and, correspondingly, the estimated t ratios (in absolute value) larger compared with the previous results.

Alternative to LP

- The LPM is plagued by several problems, such as (1) nonnormality of u_i, (2) heteroscedasticity of u_i , (3) possibility of $\hat{Y}_i$ lying outside the 0–1 range, and (4) the generally lower R^2 values.
- The fundamental problem with the LPM is that it is not logically a very attractive model because it assumes that Pi = E(Y = 1 | X) increases linearly with X, that is, the marginal or incremental effect of X remains constant throughout.
- The **logit model and the probit (or normit) model** can be used to solve these uses.

Reference

Gujarati, D (1995), *Basic Econometrics*, 4th Edition, New York: McGraw Hill