7.5 The Linear Probability Model

1. **Motivation**: What happens if we want to use multiple regression to explain a qualitative event? Till now we said the following:

   (a) What if the intercepts differ across the two categories

   $$y = \beta_0 + \delta_0 D_1 + \beta_1 x_1 + u$$

   When $D_1 = 0$, intercept is β_0, and when $D_1 = 1$ intercept is $\beta_0 + \delta_0$.

   (b) What if the slopes differ across two categories

   $$y = \beta_0 + (\beta_1 + \delta_0 D_1)x_1 + u$$

   when $D_1 = 0$, the slope is β_1 and when $D_1 = 1$ slope is $\beta_1 + \delta_0$.

   (c) What happens if slope and intercepts differ across the two categories:

   $$y = (\beta_0 + \delta_0 D_1) + (\beta_1 + \delta_1 D_1)x_1 + u$$

2. **Binary dependent variable**: $y = 1$ when an individual voted for the NDA and $y = 0$ otherwise. Under such assumption if we have the following model under MLR assumptions

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

then we have the probability of $y = 1$, called the "Response probability" is defined as:

$$P(y = 1|\mathbf{x}) = E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

3. **Linear probability models**: These belong to a class of linear probability models where the response probability is linear in β_j. In order to get the probability of $y = 0$ we use probability theory as

$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x})$$

4. Estimated equation where $\hat{y}$ represents the probability of success i.e. $y = 1$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

5. *Explaining Labor Force Participation Rates*: Consider the following model:

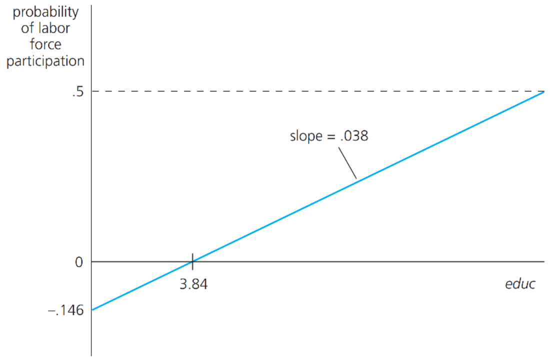
$$\begin{aligned}\widehat{inlf} =& 0.586^{**} - 0.0034^{**} nwifeinc + 0.038^{**} educ + 0.039^{**} exper \\ & - 0.00060^{**} exper^2 - 0.016^{**} age - 0.262^{**} kidslt6 + 0.013 kidsge6\end{aligned}$$

$inlf = 1$ if individual reports to be in labor force is success for this example. $nwifeinc$ is the income of partner, etc.

   (a) **Educ**: cetrius paribus, an additional year of education increases the probability of being in the labor force by 0.03 times, or a 10 year increase increases the probability of participating by 1/3rd.

(b) We could plot the probability of participation based on educ for arbitrary values of remaining variables as follows:

FIGURE 7.3 Estimated relationship between the probability of being in the labor force and years of education, with other explanatory variables fixed.



6. Limitations: (a) we cannot explain negative or greater than 1 probability (b) possible that probability is not linearly related and independent to other explanatory variables.

7. Correcting prediction errors in LPM models: $\hat{y}$ should have been $\in [0, 1]$. Re-define a new variable $\tilde{y} = 1$ if $\hat{y} \geq 0.5$ and 0 otherwise. This gives us a binary predicted variable which is a widely used goodness of fit measure.

8. Problem of heteroskedasticity: arises because the dependent variable is binary with a conditional variance as follows:

$$Var(p|\mathbf{x}) = p(\mathbf{x})(1 - p(\mathbf{x}))$$

   This would not be a problem if $p(\cdot)$ is not a function of $\mathbf{x}$. The problem is that standard errors are incorrect and inference is limited.

9. Example 7.12: Crime and arrests data.