PPOL502-01, Spring 2016
Course Notes # 14: Binary Dependent Variable: The Linear Probability Model

---

I. BINARY *DEPENDENT* VARIABLES: LINEAR PROBABILITY MODELS (LPM)

- Now we'll talk about how *binary* variables (or indicator variables) can be used as *dependent* variables in a regression model.

  $Y= \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 .... \beta_k Xk + \eta$

  where Y can take on two possible values: 0 if the condition does not hold, and one if it does. Examples…..

- In all previous multiple regression models (with interval-ratio dependent variables), our interpretation of β_k was the predicted change in Y given a one-unit increase in X_k, holding all other variables in the model constant.

- With LPM models, the interpretation of β_k will be slightly different, due to the fact that the dependent variable is now a variable that can only range from 0 to 1 and is now interpreted as a probability:

  For LPM models, β_k is the predicted probability of "success" (i.e., the case when the dependent variable is equal to one) when X_k increases by one unit (examples on the next page)

  - $E(Y|\boldsymbol{X})$ has a particular interpretation:

  $E(Y|\boldsymbol{X})$ = *Probability* $(Y=1|\boldsymbol{X})$, *or* $P(Y=1|\boldsymbol{X})$.

  $P(Y=1|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 .... \beta_k X_k$

  Q: What is the interpretation of β_0 in LPM models?

- You can obtain predicted values from LPM models just as you can from regression models with an interval/ratio dependent variable. There will be one problem though: sometimes the prediction can be outside the range of 0 to 1. We'll talk about this more after an example.

Example (using the WAGE1 dataset)…

$SERVOCC = \beta_0 + \beta_1 EDUC + \beta_2 FEMALE + \beta_3 NONWHITE + \eta$

```
. describe servocc educ female nonwhite

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
servocc         byte   %8.0g                  =1 if in service occupation
educ            byte   %8.0g                  years of education
female          byte   %8.0g                  =1 if female
nonwhite        byte   %8.0g                  =1 if nonwhite

. sum servocc educ female nonwhite

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     servocc |       526    .1406844    .3480267          0          1
        educ |       526    12.56274    2.769022          0         18
      female |       526    .4790875     .500038          0          1
    nonwhite |       526    .1026616    .3038053          0          1

.
. reg servocc educ female nonwhite

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  3,   522) =    9.07
       Model |  3.15186229     3  1.05062076           Prob > F      =  0.0000
    Residual |  60.4374913   522  .115780635           R-squared     =  0.0496
-------------+------------------------------           Adj R-squared =  0.0441
       Total |  63.5893536   525  .121122578           Root MSE      =  .34027

------------------------------------------------------------------------------
     servocc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0194017   .0054025    -3.59   0.000     -.030015   -.0087883
      female |   .1013802   .0298115     3.40   0.001      .042815    .1599453
    nonwhite |  -.0461049   .0490656    -0.94   0.348    -.1424952    .0502854
       _cons |   .3405857   .0726943     4.69   0.000     .1977764     .483395
------------------------------------------------------------------------------
```

- What is the baseline category for each indicator or conceptual set of indicators?

- predicted probability of working in a service occupation for (e.g.):

  (a) a person with no education, male, who is white = **0.341**

  (b) a female with no education, who is white = 0.341 + 0.101 = **0.442**

  (c) a female with 10 years of education who is nonwhite =0.341 - (0.019*10) + 0.101 – 0.046 = **0.206**

  In other words, our best guess of the percent of the (c) group who work in a service occupation is 20.6%.

  (d) a male with 18 years of education who is nonwhite = 0.341 – (0.019*18) – 0.046 = **- 0.047**: an out-of-range prediction!

```
------------------------------------------------------------------------------
     servocc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0194017   .0054025    -3.59   0.000     -.030015   -.0087883
      female |   .1013802   .0298115     3.40   0.001      .042815    .1599453
    nonwhite |  -.0461049   .0490656    -0.94   0.348    -.1424952    .0502854
       _cons |   .3405857   .0726943     4.69   0.000     .1977764     .483395
------------------------------------------------------------------------------
```

- β_{EDUC}: This effect is statistically significant ($p<0.0001$):

  Holding gender and race constant, for each additional year of schooling, the predicted probability of working in a service occupation decreases by 0.019, on average. (Note: remember that the possible outcome scale is 0 to 1).

  In other words, there is a 1.9 *percentage point* effect of an additional year of education (i.e., almost 2 percentage points), conditional on the other variables in the model.

  NOTE: the "percentage point" interpretation refers to a possible outcome scale of 0 to 100 (instead of 0 to 1 as the variable was originally defined). Each 1-point increase is a percentage point increase. To convert from probability to percentage points, just move the decimal two places to the right; i.e., multiply by 100).

  Holding gender and race constant, what's the change in predicted probability of working in a service occupation for going to school an additional ten years?

- β_{FEMALE}: This effect is statistically significant ($p<0.0001$):

  Holding education and race constant, the predicted probability of working in a service occupation for females (compared to males) increases by 0.101, on average. In other words, there is a 10.1 *percentage point* effect on working in a service occupation of being female (holding the other vars in the model constant). Is this effect substantively significant?

- $\beta_{NONWHITE}$: This effect is not statistically significant ($p= 0.3478$). Statistically, we cannot distinguish it from zero.

- Advantages of the LPM model:

    1. It's a direct extension of the OLS model, and so doesn't require learning a new method to model binary dependent variables.

    2. Thus, the interpretation of the coefficients is straightforward.

- Drawbacks of the LPM model:

    1. It is possible that we could get a predicted probability for $\hat{Y}$ outside the range of the 0, 1 definition of the dependent variable.

        -- This is usually not a big deal, since we are often not interested in getting a prediction from the data but instead of isolating the effect of a particular variable or set of variables on the outcome.

        -- Nonetheless, some people hold the view that LPM models are just plain wrong and that they should never be used to model binary dependent variables. Instead, they argue, one should always use either a *logit* (or logistic) model, or a *probit* model.

        -- As a first cut, however, (and often for more than that), LPMs are just fine and suitable for many kinds of estimation tasks.

        -- As Wooldridge points out, the model usually works best when making predictions based on values of independent variables that are near their means in the sample.

    2. The model is heteroskedastic, violating one of the Gauss-Markov assumptions.