Lecture 3

Limited Variables 2 of 5:
Binary Response Models I
Binary **Dependent** Variables: Linear Probability Model
(Dummy Variables 2)

Michael Curran

Trinity College Dublin

JS Econometrics

Lecture 3 Outline

Binary Dependent Variables

Linear Probability Model

Policy Analysis

Program Evaluation

Summary & References

Summary & References

Binary Dependent Variables

Linear Probability Model

Want multiple regression to *explain* a qualitative event. Simplest: binary outcome, $y = 0$ or $y = 1$. Examples?

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

$P(y = 1|\mathbf{x}) = E(y|\mathbf{x})$: the probability of success (i.e. $y = 1$) is same as expected value of y:

$$P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_k + \ldots + \beta_k x_k \tag{1}$$

which says that prob of success $p(\mathbf{x}) = P(y = 1|\mathbf{x})$ is a linear function of the x_j. (1) is a binary response model and $P(y = 1|\mathbf{x})$ is the **response probability**. **Linear probability model (LPM)** since the response probability is linear in the parameters β_j.

Linear Probability Model

Interpretation

In the LPM, β_j measures the change in the probability of success when x_j changes, holding other factors fixed:

$$\Delta P(y = 1|\mathbf{x}) = \beta_j \Delta x_j$$

So, multiple reg model allows us estimate effect of various explanatory variables on qualitative events. OLS is the same as before.

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

$\hat{y}$ is predicted probability of success so $\hat{\beta}_0$ is the predicted probability of success when each x_j is set to zero (may or may not be interesting) and slope coefficient $\hat{\beta}_1$ measures the predicted change in the probability of success when x_1 increases by one unit.

Linear Probability Model

Limitations

Negative:

1. Some combinations of values for independent variables lead to predictions (predicted probabilities) that are either less than 0 or greater than 1.
2. Constant partial effects.
3. Heteroscedasticity.

Positive: LPM works well for values of independent variables that are near the averages in the sample.

Can still use estimated probabilities to predict a zero-one outcome.

percent correctly predicted.

Linear Probability Model

Limitations

LPM violates GM ass. When y binary, its variance conditional on $\mathbf{x}$ is

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

Heteroscedasticity, unless probability doesn't depend on any of the independent variables. No bias, but t and F statistics rely on homogeneity, even when sample size is large. Corrections: heteroscedasticity-robust SE, t, F and LM statistics and tests for heteroscedasticity plus WLS, GLS and FGLS (chapter 8, especially 8.5).

Can also include dummy independent variables with DV dependent variables. Coefficient measures predicted difference in probability relative to the base group.

Lecture 3 Outline

Binary Dependent Variables

- Linear Probability Model

Policy Analysis

- Program Evaluation

Summary & References

- Summary & References

Policy Analysis

Program Evaluation

control group and **experimental group** or **treatment group**.

Except in rare cases, choice of groups is not random.

Is measured effect of qualitative variable causal? Examples.

Holzer et al (1993), effect of job training grants on worker productivity:

$$\log(scrap) = \beta_0 + \beta_1 grant + \beta_2 \log(sales) + \beta_3 \log(employ) + u$$

Controls.
Unobserved factors affecting worker productivity might be correlated with *grant*. Examples? Important to include factors that might be related to the binary independent variable of interest, even if policy analysis doesn't involve assigning units to control and treatment groups. Example: testing for racial discrimination.

$$approved = \beta_0 + \beta_1 nonwhite + \beta_2 income + \beta_3 wealth + \beta_4 credrate + otherfactors$$

Discrimination: rejection of $H_0 : \beta_1 = 0$ in favour of $H_0 : \beta_1 < 0$.

Copyright © 2013 Michael Curran

Policy Analysis

Program Evaluation

Self-selection problem: individuals self-select into certain behaviours or programs – participation is not randomly determined.

$$y = \beta_0 + \beta_1 partic + u \tag{2}$$

$E(u|partic = 1) \neq E(u|partic = 0)$ causes estimate of β_1 to be biased and so we will not uncover the true effect of participation.

- Endogenous explanatory variables.
- Multiple regression analysis can to some degree alleviate the self-selection problem.
- Factors in error term in (2) that are correlated with *partic* can be included in a multiple reg equation assuming we can collect data on these factors.
- Still may have unobserved factors related to participation, in which case multiple regression produces biased estimators.
- Panel methods/IV/2SLS.

Lecture 3 Outline

Binary Dependent Variables

- Linear Probability Model

Policy Analysis

- Program Evaluation

Summary & References

- Summary & References

Summary

- Linear probability model explains qualitative events (binary dependent variable).
    - Interpreting β_j: change in probability of success.
    - Be aware of drawbacks to LPM.
- Program evaluation is a special case of policy analysis.
    - Control group and treatment group plus controls.
    - Self-selection problem: non-random participation complicates issues.

References

- Linear Probability Model: Wooldridge 7.5.
- Policy Analysis & Program Evaluation: Wooldridge 7.6.