DATA 5610

Logistic Regression (LPM, Probit)

Sources:

- Wooldridge chps 7, 17
- James et al chp. 4
- Alpaydin chp. 5

- Start by looking at linear regression with a binary dependent variable

- In Econometrics we call this the Linear Probability Model (LPM)

- What happens when we write down a regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

when y is a binary random variable?

- B/c Y only takes on values $\{0,1\}$ we cannot interpret β_j as the change in y given a 1 unit change in x_j (holding all x's fixed)

- Y either changes from $\emptyset$ to 1 or from 1 to $\emptyset$ (or doesn't change)

- Still β_j may still have a useful interpretation

- If we assume if the zero conditional mean holds, i.e.

$$E[u \mid x_1, \ldots, x_k] = \emptyset$$ then we have

$$E[y \mid X] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- When $y \in \{0,1\}$ it is true that

$$P(y=1 \mid \mathcal{X}) = E(y=1 \mid \mathcal{X})$$

- In other words: the probability of "success" $[P(y=1)]$ is the same as the expected value of y, $E(y)$

so that

$$P(y=1 \mid \mathcal{X}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- The prob. of success $P(y) = P(y=1 \mid \mathcal{X})$ is a linear function of the x_j's

- $P(y=1 \mid x)$ is also called the response probability

  - NB: that $P(y=\emptyset \mid x) = 1 - P(y=1 \mid x)$

  - so $P(y=0)$ is also a linear function of the x_j's

- This is called the Linear Probability Model when estimated via OLS

- In the LPM β_j measures the change in response prob. ~~given~~ give a 1 unit change in x_j

$$\Delta P(y=1 \mid x) = \beta_j \Delta x_j \qquad (\text{holding all else fixed})$$

- The mechanics of estimation via OLS are the same as for regression with a continuous response variable

- The estimated equation is

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

  - NB: $\hat{y}$ is the predicted prob. of success

  - $\hat{\beta}_0$ is the pred. prob of success when $\beta_j = 0$ for $j = 1, \ldots, k$

  - $\hat{\beta}_{j \neq 0}$ is the change in response prob. given a 1 unit change in x_j

LPM's Short comings

- For different values of the x_j's it is possible to get predictions $\hat{y}$ that are less then $\emptyset$ or greater than 1

- That's non-sensical if $\hat{y}$ is a response probability, which needs to be on [0,1] to be a valid probability.

- A related problem is the magnitude of $\hat{\beta}_j$

  - Let's say y is labor-force participation
  - And x_j is the number of children
  - And $\hat{\beta}_j = .262$
  - Then $\Delta \widehat{inlf} = .262\,(\Delta \text{kids})$

$.262(1) = .262$ but

$.262(4) = 1.048$ which is nonsensical!!

- The LPM is still used quite a bit in Econometrics despite these drawbacks

- Even though a response prob. less than $\emptyset$ or greater than 1 is logically troublesome, it may not matter much for prediction

- Let $\hat{y}_i$ denoted the fitted ~~values~~ values

  - Define a predicted value as

$$\tilde{y} = 1 \text{ if } \hat{y} \geq .5$$
$$\tilde{y} = \emptyset \text{ if } \hat{y} < .5$$

  - Then for $i = 1, \ldots, n$ we have $\tilde{y}_i$ that will be either $\emptyset$ or 1

- NB: LPM violates the homoskedasticity assumption of the GM theorem

Logit and Probit models (as known in Econometrics)

- Binary response variable
- Again, we want to model the response probability

$$P(y=1 \mid x) = P(y=1 \mid x_1, x_2, \dots, x_k)$$

- when x denotes the full set of explanatory variables (features)

- To avoid the shortcomings of the LPM we consider models of the form

$$P(y=1 \mid x) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

$$= G(\beta_0 + x\beta)$$

- G is a function taking on ~~[illegible]~~ values between zero and one :

$$0 < G(z) < 1$$

i.e. G is a "squisher" function of some kind

- This ensures that the estimated response prob's are strictly between 0 and 1 for all real numbers z

- There are various functions we can use for G

- Two considered here :

  1. the logit model : $G(z) = \dfrac{\exp(z)}{1 + \exp(z)} = \Lambda(z)$

  2. the probit model : $G(z) = \Phi(z) = \displaystyle\int_{-\infty}^{z} \phi(v)\,dv$

     where $\phi(z)$ is the standard normal pdf

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$$

- $\Lambda(z)$ is the logistic CDF
- $\Phi(z)$ is the std. normal CDF (expressed as an integral)
- B/c these are "valid" CDFs they will ensure that the response probability will be strictly between $\emptyset$ and 1.
- Both are increasing functions

$$G(z) \to \emptyset \quad \text{as } z \to -\infty$$

$$G(z) \to 1 \quad \text{as } z \to \infty$$

- Logit and probit can be thought of as coming from an underlying "Latent variable model"

- Let y^* be an unobserved (latent) variable

- Suppose that

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$$

$$y = 1[y^* > 0]$$

- NB: $1[\cdot]$ is called an "indicator function"
  - It produces a binary variable given a continuous input and a boolean conditional

- Therefore

$$y = \begin{cases} 1 & \text{if } y^* > 0 \\ \emptyset & \text{if } y^* \leq 0 \end{cases}$$

- Assume that e is independent of x and either has a standard logistic (logit) or standard normal (probit) distribution

- In either case, e is symmetrically distributed about $\emptyset$

$$1 - G(-z) = G(z)$$

- From this we can derive the response probability for y:

$$P(y=1|x) = P(y^*>0|x) = P[e > -(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)|x]$$

$$= 1 - G[-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)]$$

$$= G(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)$$

- In some applications (econometrics) the focus is on properly interpreting the $\hat{\beta}_j$ as the effects of x_j on $P(y=1|x)$

- The latent variable formulation gives the impression that we are interested in the effects of x_j on y^*

- For Logit/Probit the direction of the effects of each x_j on $E(y^*|x) = \beta_0 + x\beta$ is always the same

- The latent variable y^* rarely has a well-defined unit of measurement

- The magnitudes of β_j's are not (by themselves) of much use (in contrast to the LPM)

- We want to estimate the effect of x_j on the response prob., $P(y=1|x)$ but this is complicated by the nonlinearity of $G(\cdot)$

- If x_j is a (roughly) continuous variable, its partial effect on $p(x) = P(y=1 \mid x)$ is obtained from the partial derivative

$$\frac{\partial p(x)}{\partial x_j} = g(\beta_0 + X\beta)\,\beta_j$$

where $g(z) \equiv \frac{dG}{dz}(z)$

- Because G is a CDF, g is a PDF

- For Logit/Probit $G(\cdot)$ is strictly increasing CDF, so $g(z) > 0$ for $\forall\, z$

- The partial effect of x_j on $p(x)$ depends on x through the positive quantity $g(\beta_0 + x\beta)$, which means the partial effect always has the same sign as β_j

- The relative effects of any two features do not depend on x

  - The ratio of the partial effects for x_j and x_h is β_j / β_h

- If $G(\cdot)$ is a symmetric density about zero (with unique mode at zero) the largest effect happens when $\beta_0 + x\beta = \emptyset$

  - Probit: $g(z) = \phi(z)$, so $g(0) = 1/\sqrt{2\pi} \approx .40$

  - Logit: $g(z) = \dfrac{\exp(z)}{[1+\exp(z)]^2}$, so $g(0) \approx .25$

- If x_1 is a binary feature/predictor then the partial effect from changing x_1 from zero to one (holding all else fixed) is

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

- This depends on the values of the other x_j
- For example, let y be an employment indicator and x_1 a dummy indicating job program participation then the above is
  - The change in the prob of employment due to the job program
  - Depends on the other x_j, such as:
    - AGE
    - EDUC
    - Experience
    - etc

- Note: the sign of β_1 is enough to detect if the program has a positive or negative effect

- If we want the magnitude of the effect, then we have to estimate

$$G(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k) - G(\hat{\beta}_0 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)$$

- We can do the same for other discrete variables, such as the number of children

    - If x_k is the variable, then the effect on the prob of x_k going from c_k to c_k+1 is simply

$$G[\beta_0 + \beta_1 x_1 + \cdots + \beta_k (c_k + 1)] -$$

$$G[\beta_0 + \beta_1 x_1 + \cdots + \beta_k c_k]$$

MLE

- Under the classical regression assumptions, OLS is the MLE
- B/c of the nonlinearity of $E(y|x)$ for Logit/Probit we have to use MLE directly
- Assume we have a random sample of size n
- For the MLE, we need the density of y_i given x_i

$$f(y|x_i;\beta) = [G(x_i\beta)]^{y}[1-G(x_i\beta)]^{1-y} \quad \text{for } y=0,1$$

- For $y=1$: we get $G(x_i\beta)$
- For $y=0$: we get $1-G(x_i\beta)$

- The Log-likelihood for observation i is a function of the parameters and the data (X_i, Y_i) and is obtained by taking logs

$$\ell_i(\beta) = y_i \log[G(x_i\beta)] + (1-y_i)\log[1 - G(x_i\beta)]$$

- B/c $G(\cdot)$ is strictly between $\emptyset$ and 1 for logit/probit, $\ell_i(\beta)$ is well defined for all values of β

- The log-likelihood for a sample size of n is obtained by summing

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} l_i(\beta)$$

- The MLE of β, denoted by $\hat{\beta}$, maximizes this value
- If $G(\cdot)$ is the standard logistic CDF, then $\hat{\beta}$ is the logit estimator
- If $G(\cdot)$ is the standard normal CDF, then $\hat{\beta}$ is the probit estimator

- B/c of the nonlinearity of $G(\cdot)$ we cannot obtain closed-form solutions for $\hat{\beta}$

- We have to rely on numerical optimization

  - This makes the statistical aspects of logit/probit much more complex than ~~OLS~~ OLS

  - But b/c it's an MLE, $\hat{\beta}$ has "good" sampling properties (consistency, asymptotic normality, efficiency, etc)

- Numerical methods used in practice: Newton's method, gradient descent, etc.

- Q: because $\hat{\beta}$ is an MLE, does that mean we can interpret it as a MAP?