Maximum Likelihood and Binary Dependent Variable Models

Matteo Paradisi
(EIEF)

Applied Micro - Lecture 9

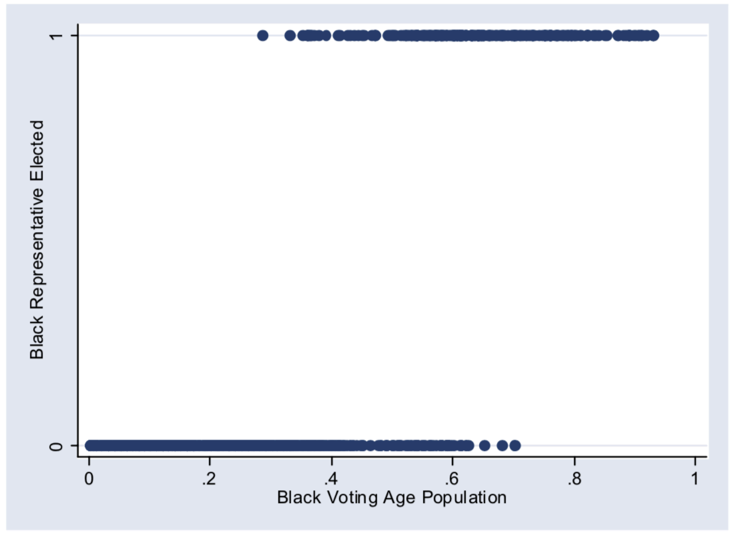
Limited Dependent Variable

- So far Y, the dependent variable, was continuous
- However, dependent variables could be dichotomous (dummy variables) or categorical
- Hence, we study non-linear estimation with dichotomous Y vars
- Some examples
  - Votes (Left vs Right)
  - Labor force participation (extensive margin)
  - College dropout

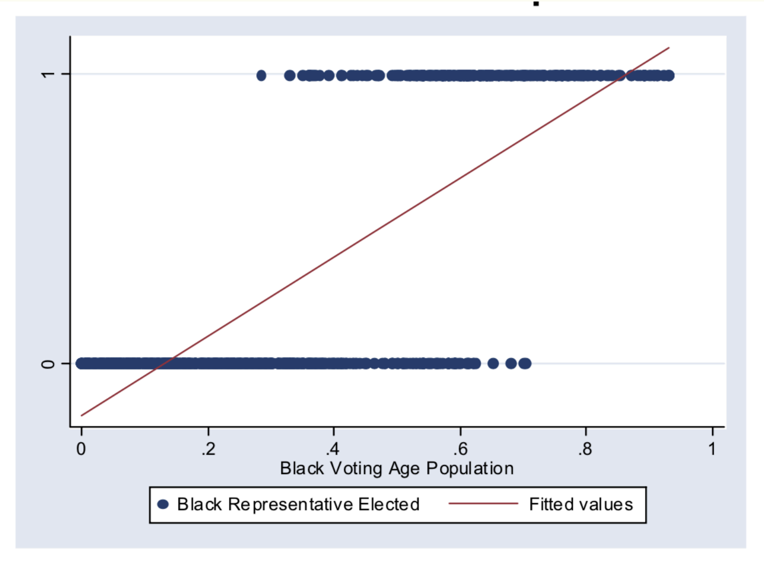
Example and Intuition

- Suppose you want to relate the share of blacks in the population to whether a black representative is elected
- Dependent variable: dummy =1 if elected
- Now plot the data to see the relationship

Example and Intuition



Example and Intuition - Linear Fit



Example and Intuition

- A line does not fit the data well
- We need something better
- AND something that will predict values between 0 and 1
- What can we do?

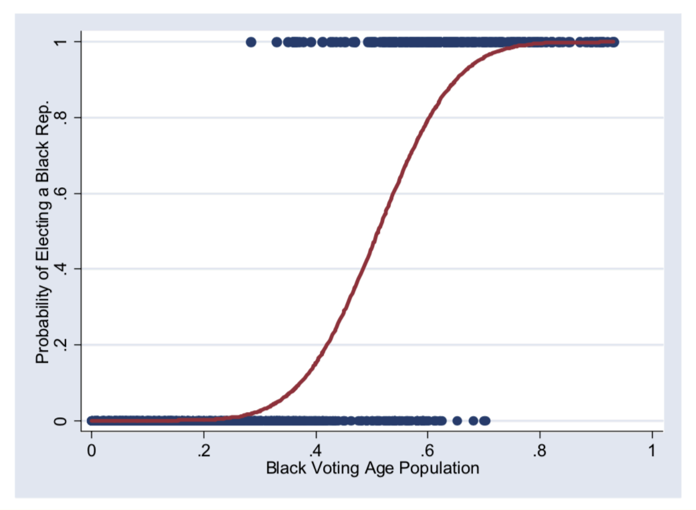
Example and Intuition

- Think of dependent variable as a probability of the event
- We need a function that takes continuous values and provides something in the $[0, 1]$ interval
- Which function does this?
- Example: the CDF of a Normal distribution!

$$Y = \Phi\left(X\beta + \varepsilon\right)$$

- Must be careful about interpretation of βs (we'll see it later)
- This model fits the data better!

Example and Intuition - Non-Linear Fit



Estimating the Model

- This model allows us to better fit the data
- However, how do we estimate it?
- We need to introduce the concept of maximum likelihood estimation

Maximum Likelihood Estimation

Maximum Likelihood - Introduction

- What is maximum likelihood?
- Estimation method: find values of parameters that maximize the likelihood of observing the sample at hand
- Two steps:
  1. Write a closed-form of the likelihood
     - function of data and parameters
  2. Maximize it to find estimates
- These methods are particularly useful in models where the dependent variable is a discrete choice

Probit Model

- We start from the simplest model
- Let's assume that our theory gives us a latent variable y^* determined by some x

$$y^* = x\beta + u$$

- However, y^* is not observed. It could be for instance the utility from a choice
- We instead observe a binary choice

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

- Notice that 0 is just a normalization
- Example: if y is choice about entering the labor force, then you enter if utility from entering is greater than alternative (normalized to 0)

Probit Model

- Suppose that $(y_i, x_i)_{i \in N}$ are i.i.d.
- We need to find a way to write the probability of observing a vector of choices and characteristics
- Assume that $u \sim N(0,1)$ then

$$
\begin{aligned}
\Pr(y_i = 1 | x_i, \beta) &= \Pr(x_i \beta + u_i > 0 | x_i) \\
&= \Pr(u_i > -x_i \beta | x_i, \beta) \\
&= 1 - \Phi(-x_i \beta) = \Phi(x_i \beta)
\end{aligned}
$$

  where $\Phi(\cdot)$ is the cdf of a Normal distribution
- Analogously

$$
\Pr(y_i = 0 | x_i, \beta) = 1 - \Phi(x_i \beta)
$$

Likelihood Function

- We can then write the likelihood of observing (y_i, x_i)

$$p(y_i|x_i, \beta) = [\Phi(x_i\beta)]^{y_i} [1 - \Phi(x_i\beta)]^{1-y_i}$$

  also called "likelihood contribution" of i

- Hence, the likelihood for the entire sample is

$$L(\beta) = \prod_{i=1}^{N} p(y_i|x_i, \beta)$$

- since y_i and x_i are observed, β is the only unknown

Maximizing the Likelihood

- In most cases, algorithms maximize a monotonic transformation of L

$$\hat{\beta}_{ML} = \arg\max_{\beta \in \Theta} \log \mathbf{L}(\beta) = \arg\max_{\beta \in \Theta} \sum_{i=1}^{N} \ln \mathbf{p}(\mathbf{y_i} | \mathbf{x_i}, \beta)$$

- the transformation takes the log of L and it is referred to as log-likelihood

- Notice that Θ is a generic set, so that one can add additional constraints on β

Linear Regression and Maximum Likelihood

- We can apply maximum likelihood to linear regressions too
- Consider the model
$$y_i = x_i \beta + u_i$$
- We add assumption to standard OLS assumptions:
$u \sim N\left(0, \sigma^2\right)$

Linear Regression and Maximum Likelihood

- Because data is continuous in this case, we cannot write a probability function
- We write a distribution function instead

$$f(y_i|x_i, \beta, \sigma) = f(u_i = y_i - x_i\beta|x_i)$$
$$= \varphi\left(\frac{y_i - x_i\beta}{\sigma}\right)$$

- Hence, log-likelihood is

$$L(\beta, \sigma) = \sum_{i=1}^{N} \ln \varphi\left(\frac{y_i - x_i\beta}{\sigma}\right)$$

Linear Regression and Maximum Likelihood

Rewrite the likelihood as

$$
\begin{aligned}
L(\beta, \sigma) &= \sum_{i=1}^{N} \ln \varphi\left(\frac{y_i - x_i\beta}{\sigma}\right) \\
&= \sum_{i=1}^{N} \ln\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - x_i\beta}{\sigma}\right)^2\right)\right] \\
&= \sum_{i=1}^{N}\left[-\ln\sigma - \frac{1}{2}\ln 2\pi - \frac{1}{2\sigma^2}(y_i - x_i\beta)^2\right] \\
&= -N\ln\sigma - \frac{N}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2
\end{aligned}
$$

Linear Regression and Maximum Likelihood

$$L(\beta, \sigma) = -N \ln \sigma - \frac{N}{2} \ln 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i \beta)^2$$

- FOCs are

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} x_i' (y_i - x_i \beta) = 0$$

$$-\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} (y_i - x_i \beta)^2 = 0$$

- Combining them

$$\hat{\beta}_{ML} = \left[\sum_{i=1}^{N} x_i' x_i \right]^{-1} \left[\sum_{i=1}^{N} x_i' y_i \right]$$

$$\hat{\sigma}^2_{ML} = N^{-1} \sum_{i=1}^{N} (y_i - x_i \beta)^2$$

Linear Regression and Maximum Likelihood

$$\hat{\beta}_{ML} = \left[\sum_{i=1}^{N} x_i' x_i\right]^{-1} \left[\sum_{i=1}^{N} x_i' y_i\right]$$

$$\hat{\sigma}^2_{ML} = N^{-1} \sum_{i=1}^{N} (y_i - x_i \beta)^2$$

- These are OLS formulas!
- However, the result depends on the assumption on u's distribution
- With a different distribution, OLS would not be a maximum likelihood estimator for this linear model

Formal Characterization of Maximum Likelihood

- Let's be a little more formal, derive the maximum likelihood, and discuss some properties
- Assumption 1: $\text{y}|\text{x} \sim \text{i.i.d. F}\left(\cdot|\theta\right)$
- Conditioning on θ emphasizes the fact that F is a function of the parameters to be estimated

Formal Characterization of Maximum Likelihood

- Log likelihood contribution of i

$$\ell_i(\theta) = \ln f(y_i|x_i, \theta)$$

- Total log likelihood is

$$L(\theta) = \sum_{i=1}^{N} \ell_i(\theta) = \sum_{i=1}^{N} \ln f(y_i|x_i, \theta)$$

- The maximum likelihood estimator is

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta)$$

- We must add assumptions to make $\hat{\theta}_{ML}$ consistent and asymptotically efficient

Additional Assumptions

- Additional assumptions:
  - a) Θ is closed and bounded
  - b) f is continuous and twice differentiable over Θ
  - c) $f(y|\theta)$ is such that $f(y|\theta_1) = f(y|\theta_2)$ if and only if $\theta_1 = \theta_2$
- The last assumption makes sure that L is never flat and there is a unique maximizer

Conditional VS Full Maximum Likelihood

- So far the formal model started from a conditional distribution of $y|x$
- But the probit model presented before relied on a joint distribution of x and y
- Notice that

$$f(y, x|\theta) = f(y|x, \theta)\, f(x|\theta)$$

- If the distribution of x does not depend on θ (i.e. $f(x|\theta) = f(x)$) then the θ maximizing $f(y|x, \theta)$ also maximizes $f(y, x|\theta)$
- If this is not the case, a full maximum likelihood is needed

Asymptotic Distribution and Properties

- What is the asymptotic distribution of $\hat{\theta}_{ML}$?

$$\hat{\theta}_{ML} \overset{a}{\sim} N\left[\theta, -E\left(\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}\right)^{-1}\right]$$

- First, the mean is θ, so the estimator is consistent
- Second, the variance is the negative of the inverse of the Hessian matrix. This is called Cramer-Rao lower bound and it is the smallest possible variance estimator. It follows that the estimator is also efficient!

Discussion on Maximum Likelihood VS OLS

- ML is consistent and efficient, so this is like the best we could ask for
- However, we have to put very strong assumptions to derive these estimators
- In particular, we have to take a stance on data distribution
- OLS only makes assumption on the conditional mean of the distribution ($E(u|X) = 0$), not on the entire conditional distribution
- However, the mild assumption of the OLS does not guarantee efficiency
- ML is attractive, but if wrong assumption on distribution then we lose efficiency AND consistency

Binary Choice Models

Binary Choice Models Formally

- Let's analyze binary choice models in greater details
- Let's go back to probit model

$$y^* = X\beta + u$$

- We observe

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

- To derive ML we need assumption on the distribution of u
- We assume it is normally distributed, this time with some variance σ^2

$$u \sim N\left(0, \sigma^2\right)$$

Probit Model

- Probability of observing $y = 1$ and $y = 0$

$$\begin{aligned}
\Pr\left(y = 1|X, \beta\right) &= \Pr\left(X\beta + u > 0|X\right) \\
&= \Pr\left(\frac{u}{\sigma} > -\frac{X\beta}{\sigma}|X, \beta\right) \\
&= \Phi\left(\frac{X\beta}{\sigma}\right)
\end{aligned}$$

$$\Pr\left(y = 0|X, \beta\right) = 1 - \Phi\left(\frac{X\beta}{\sigma}\right)$$

- Log likelihood is

$$L\left(\beta, \sigma\right) = \sum_{i=1}^{N}\left\{y_i \ln\Phi\left(\frac{X\beta}{\sigma}\right) + \left(1 - y_i\right)\ln\left[1 - \Phi\left(\frac{X\beta}{\sigma}\right)\right]\right\}$$

Cannot identify both β and σ

$$L(\beta, \sigma) = \sum_{i=1}^{N} \left\{ y_i \ln \Phi\left(\frac{X\beta}{\sigma}\right) + (1 - y_i) \ln \left[1 - \Phi\left(\frac{X\beta}{\sigma}\right)\right] \right\}$$

- Notice that β and σ always appear as a ratio
- Cannot identify both since if $\hat{\beta}$ and $\hat{\sigma}$ maximize L, then also $\tilde{\beta} = c\hat{\beta}$ and $\tilde{\sigma} = c\hat{\sigma}$ do
- For this reason, any assumption on σ is irrelevant
- Hence, as a convention probit models assume $\sigma = 1$

$$\hat{\beta}_{ML} \in \arg\max_{\beta} L(\beta, 1)$$

Interpreting βs

- In OLS, $\hat{\beta}_k = \frac{\partial y}{\partial x_k}$, is the effect of x_k on y (partial derivative interpretation)
- But remember how we setup the probit: $\hat{\beta}_{ML}$ is the effect on the latent variable y^*, NOT the effect on the probability that $y = 1$.
- For an economic interpretation we want to know the effect on the probability that $y = 1$
- Hence

$$Pr(y = 1|X) = F(X\beta) = F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K)$$

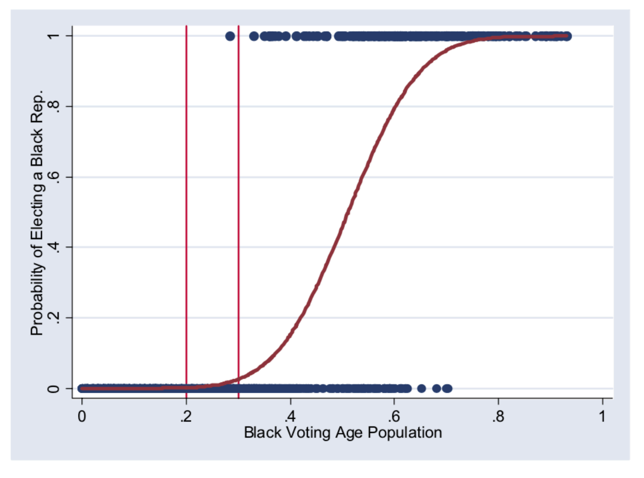
- The effect we are interested in is

$$\frac{\partial Pr(y = 1|X)}{\partial x_k} = f(X\beta)\beta_k$$

- These are called marginal effects

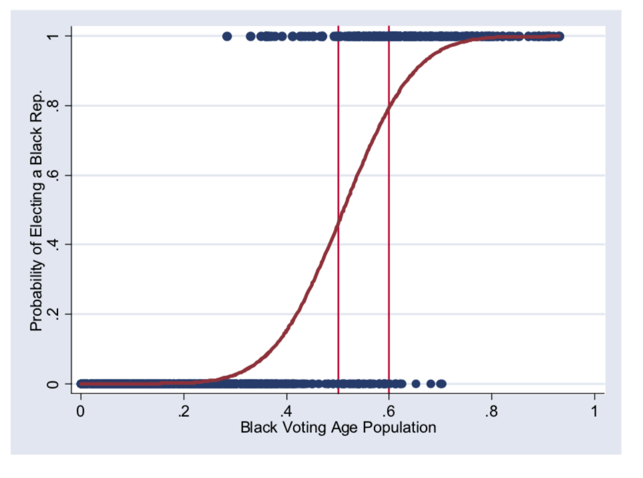
Marginal Effects - Graphical Intuition

The effect is small for small X



Marginal Effects - Graphical Intuition

The effect is larger for average X



Marginal Effects

- Marginal effects depend on $F(X\beta)$ and therefore are not constant
- Normally people look for meaningful points at which presenting marginal effects
- One option is the mean of Xs

$$\frac{\partial \Pr(y = 1|X)}{\partial x_k} = f(\bar{X}\beta)\,\beta_k$$

- where $\bar{X}$ is the vector of means for Xs
- If x_k is a dummy variable (e.g. gender), marginal effects should show

$$\Pr\left(y = 1|\bar{X}_{-k}, x_k = 1\right) - \Pr\left(y = 1|\bar{X}_{-k}, x_k = 0\right)$$

More on Marginal Effects

- The ML coefficients however already contain some useful info
- First, their sign shows us the direction of the effect of x_k
- Second, they tell us the relative importance of marginal effects

$$\frac{\partial \Pr(y = 1|X) / \partial x_k}{\partial \Pr(y = 1|X) / \partial x_j} = \frac{\beta_k}{\beta_j}$$

- If $\beta_k > \beta_j$, then x_k has a greater marginal effect than x_j

The Linear Probability Model

- An alternative to the probit is a simple linear probability model
- Linear model with dichotomous dependent variable y
- Hence

$$\Pr\left(y = 1\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K = X\beta$$

- What is the problem here?
- Nothing constraints $\widehat{\Pr\left(y = 1\right)}$ to be between 0 and 1
- Indeed

$$\begin{aligned} E\left(y|X\right) &= 1 \times \Pr\left(y = 1|X\right) + 0 \times \Pr\left(y = 0|X\right) \\ &= \Pr\left(y = 1|X\right) = X\beta \end{aligned}$$

The Linear Probability Model

- The error term is also a dichotomous variable

$$u = \begin{cases} 1 - X\beta & \text{if } y = 1 \text{ with probability } \Pr(y = 1|X) = X\beta \\ -X\beta & \text{if } y = 0 \text{ with probability } \Pr(y = 0|X) = 1 - X\beta \end{cases}$$

- The mean is zero

$$\begin{aligned} E(u|X) &= (1 - X\beta) \times \Pr(y = 1|X) - X\beta \times \Pr(y = 0|X) \\ &= (1 - X\beta) X\beta - X\beta (1 - X\beta) = 0 \end{aligned}$$

- and variance

$$\begin{aligned} \text{Var}(u|X) = E\left(u^2|X\right) &= (1 - X\beta)^2 X\beta - (X\beta)^2 (1 - X\beta) \\ &= (1 - X\beta) X\beta \end{aligned}$$

- since the variance is not constant the model is necessarily heteroskedastic

The Linear Probability Model

- We can deal with heteroskedasticity the usual way (GLS)
- For consistency we only need $\mathrm{E}\left(X'u\right) = 0$
- The big advantage of this model however is that it requires milder assumptions than the probit
- Also, easier to interpret: β is already the effect on the probability
- Last but not least, it can easily be used with panel data unlike the probit

The Logit Model

- The logit model is another alternative to the probit
- It works the same way, but with a different assumption on the distribution of y
- Logit assumes that $u \sim F(u) = e^{-e^{-u}}$
- This is called Type I extreme distribution
- It seems ugly, but has nice properties that we will see in multinomial models

Some Applications

The Effects of Different Political Campaigns

- Wantchekon (2003) studies the effects of clientelism in political campaigns
- The focus is on developing countries
- Research question: is a purely clientelist political platform more effective than a purely public policy one?
- Experimental design in Benin to answer the question

Experimental Design

- Two interventions: present the same policies with a clientelist framing vs a "public policy" framing
- You need three groups: two treatment arms and one control
- Each group is composed of different villages to avoid spillovers

Experimental Design

TABLE 1
DESCRIPTION OF THE EXPERIMENTAL DISTRICTS

District	*Exp. Candidate*	*Exp. Villages*	*Treatment*	*Ethnicity*
Kandi	Kerekou	Kassakou	clientelism	Bariba (92%)
		Keferi	public policy	Bariba (90%)
Nikki	Kerekou	Ouenou	clientelism	Bariba (89%)
		Kpawolou	public policy	Bariba (88%)
Bembereke	Saka Lafia	Bembereke Est	clientelism	Bariba (86%)
		Wannarou	public policy	Bariba (88%)
Perere	Saka Lafia	Tisserou	clientelism	Bariba (93%)
		Alafiarou	public policy	Bariba (94%)
Abomey-Bohicon	Soglo	Agnangnan	clientelism	Fon (99%)
		Gnidjazoun	public policy	Fon (99%)
Ouidah-Pahou	Soglo	Acadjame	clientelism	Fon (99%)
		Ahozon	public policy	Fon (99%)
Aplahoue	Amoussou	Boloume	clientelism	Adja (99%)
		Avetuime	public policy	Adja (96%)
Dogbo-Toviklin	Amoussou	Dékandji	clientelism	Adja (99%)
		Avedjin	public policy	Adja (99%)
Parakou	Ker./Lafia	Guema	competition	Bariba (80%)
		Thiam	competition	Bariba (82%)
Come	Am./Soglo	Kande	competition	Adja (90%)
		Tokan	competition	Adja (95%)

Probit Analysis

- Estimate the following probit model:

$$y^1_{ik} = \alpha + \beta X_i + \lambda y^0_{ik} + \gamma CL_k + \delta PB_k + \varepsilon_i$$

- We observe $y_{ik} = 1$ if $y^*_{ik} > 0$ and $y_{ik} = 0$ if $y^*_{ik} \leq 0$
- CL_k: clientist treatment; PB_k: public policy treatment
- y^0_{ik}: past vote

Results of Probit Analysis

TABLE 5
PROBIT ANALYSIS OF VOTE FOR TYPE OF CANDIDATES IN TREATMENT VILLAGES

	Southern	*Northern*	*Local*	*National*	*Incumbent*	*Opposition*
Constant	−0.946**	−0.513	−0.367	−0.741	−0.186	0.222
	(0.395)	(0.374)	(0.306)	(0.469)	(0.415)	(0.271)
Sex	−0.513*	−0.516***	−0.424	−0.828**	−0.415	0.024
	(0.200)	(0.194)	(0.330)	(0.332)	(0.370)	(0.231)
Age	0.006	−0.003	−0.009*	0.011*	0.004	0.002
	(0.006)	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)
Past	2.139***	.865***	1.555***	2.057***	1.893***	0.966***
	(0.203)	(0.235)	(0.201)	(0.271)	(0.180)	(0.215)
Public policy	0.309**	−0.372***	−0.594*	0.429	−0.287	0.512*
	(0.333)	(0.365)	(0.318)	(0.427)	(0.387)	(0.290)
Clientelist	1.004**	0.264	0.444	0.550	0.344	0.754**
	(0.447)	(0.391)	(0.342)	(0.457)	(0.468)	(0.319)
Sex*Client.	−0.502	−0.191	−0.348	0.489	0.208	−0.324
	(0.505)	(0.435)	(0.379)	(0.548)	(0.539)	(0.364)
Sex*Public Pol.	0.167	−1.050**	0.147	−0.572	−0.111	−0.773**
	(0.402)	(0.414)	(0.358)	(0.482)	(0.450)	(0.345)
N	524	543	596	510	472	602
log-L	−145.250	−208.538	−284.0500	−115.986	−146.161	−244.583

Changes in Production Inputs

- Conley and Udry (2010) study how information changes choices of inputs
- They study effect of good and bad news for information neighbors
- See whether farmers adjust to align with those of neighbors
- Study adoption of fertilizer (new technology) in Ghana

Changes in Production Inputs: Logit

- They run the following specification

$$\Pr\left(\Delta x_{it} \neq 0\right) = \Lambda \left[\begin{array}{c} \alpha_1 s\left(\text{good}, x = x_{it_p}\right) + \alpha_2 s\left(\text{good}, x \neq x_{it_p}\right) \\ \alpha_3 s\left(\text{bad}, x = x_{it_p}\right) + \alpha_4 s\left(\text{bad}, x \neq x_{it_p}\right) \\ + \alpha_5 \tilde{\Gamma}_{it} + z_{it}' \alpha_6 \end{array} \right]$$

- $s\left(\text{good}, x = x_{it_p}\right)$: share of neighbors with same technology with good news
- $\tilde{\Gamma}_{it}$: difference with growing conditions nearby

Results of Logit Analysis

TABLE 4—DETERMINANTS OF CHANGING INPUT USE

	A	B	C
	Dependent variable: Indicator for change between zero and positive	Dependent variable: Indicator for $\|change\|$ $> 1Cedi/Plant$	Dependent variable: Indicator for nonzero change in fertilizer
Good news at previous input use $s(good, x = x_{i,previous})$	−0.94 (1.24) [−0.04]	−0.08 (0.95) [−0.01]	−0.34 (0.84) [−0.03]
Good news at alternative fertilizer use $s(good, x \neq x_{i,previous})$	1.15 (0.81) [0.03]	1.64 (0.78) [0.09]	2.35 (1.80) [0.14]
Bad news at lagged fertilizer use $s(bad, x = x_{i,previous})$	6.38 (2.86) [0.15]	4.32 (1.93) [0.20]	4.16 (1.80) [0.22]
Bad news at alternative fertilizer use $s(bad, x \neq x_{i,previous})$	−6.72 (3.04) [−0.09]	−5.90 (2.57) [−0.15]	−3.05 (1.85) [−0.09]
Ave. abs. dev. from geog. neighbors' fertilizer use [$\Gamma_{i,t}$]	0.09 (0.10) [0.07]	0.15 (0.07) [0.24]	0.08 (0.04) [0.15]
Novice farmer	2.32 (0.75) [0.26]	1.97 (0.89) [0.43]	1.22 (0.92) [0.30]
Talks with extension agent	−0.48 (0.61) [−0.05]	−1.35 (0.67) [−0.29]	−1.38 (0.76) [−0.34]
Wealth (million cedis)	0.20 (0.10) [0.06]	0.18 (0.13) [0.10]	0.10 (0.12) [0.06]
Clan 1	1.62 (1.14) [0.18]	1.59 (1.10) [0.35]	2.15 (1.03) [0.54]
Clan 2	4.54 (1.45) [0.51]	2.15 (1.23) [0.47]	2.51 (0.99) [0.63]
Church 1	1.84 (0.93) [0.21]	−0.29 (0.73) [−0.06]	−0.24 (0.77) [−0.06]