Lecture 3
Discrete Choice Models

1

Limited Dependent Variables

• So far, implicitly, we have assumed that the variable y_i is a continuous random variable.

• But, assumptions (A1)-(A4) in the CLM does not require continuity for y_i: y_i can have discontinuities, it can be discrete, follow counts, etc. Thus, we can use OLS with "*limited dependent variables*".

• Suppose, we have binary data, that is, y_i = (0, 1), for example, enroll/not enroll in an MBA program. We also have a vector of explanatory variables, $\boldsymbol{x}_i$, for example, work experience and age.

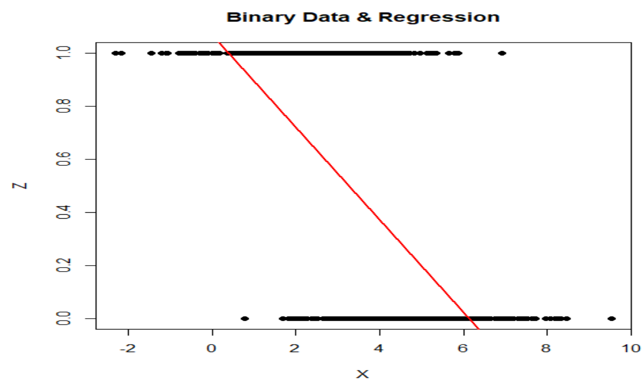
We use a linear model. Then, E[y_i] = $\boldsymbol{x}_i\boldsymbol{\beta}$. (We call this a *linear probability model*). This model has two main limitations:

1) Fitted values may get out of range.

2) Marginal effects are constant.

Limited Dependent Variables

Example: We simulate binary data (0, 1) for the dependent variable, y, & a continuous variable for x . We plot the regression (fitted) line in the scatter plot of the data.



Limited Dependent Variables

• With limited dependent variables, the conditional mean is rarely linear. We need to use adjusted models and adjust interpretations. For example, with binary data, we think of the the dependent variable, y, in terms of probabilities.

• Different types of discontinuities generate different models:

Discrete Dependent Variable → Discrete Choice Models (DCM)

Discrete Dependent Variable → Duration (Hazard) Models

Continuous dependent variable → Duration (Hazard) Models

Continuous dependent variable → Truncated/ Censored Regr. Models

Limdep: Discrete Choice Models (DCM)

• We usually study discrete data that represent a decision, a choice.

• Sometimes, there is a **single choice**. Then, the data come in binary form with a "1" representing a decision to do something and a "0" being a decision not to do something.

⇒ Single Choice (binary choice models): Binary Data

Data: y_i = 1 (yes/accept) or 0 (no/reject)

Examples: Trade a stock or not, for or against a board nominee, etc.

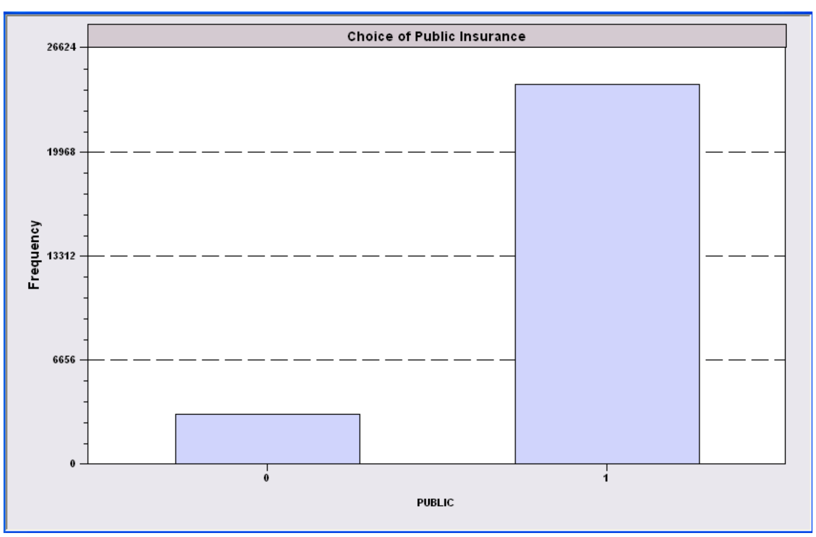
• Or we can have **several choices**. Then, the data may come as 1, 2, ..., J, where J represents the number of choices.

⇒ Multiple Choice (multinomial choice models)

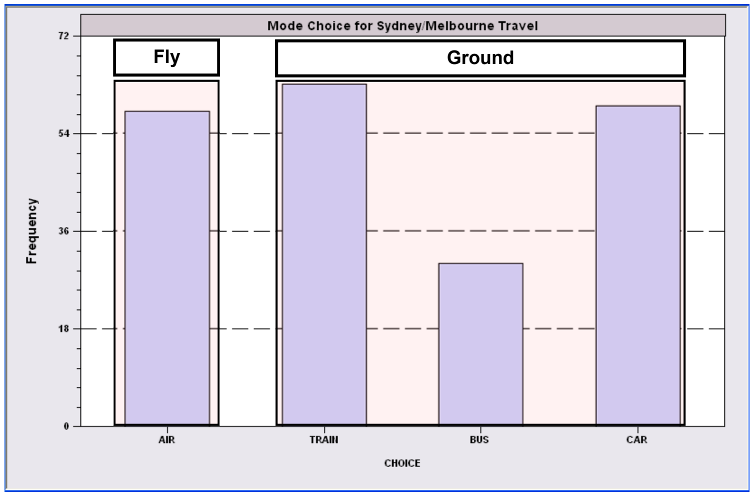
Data: y_i = 1(opt. 1), 2 (opt. 2), ..., J (opt. J)

Examples: CEO candidates, transportation modes, etc.

Limdep: DCM – Binary Choice - Example



Limdep: DCM – Mulinomial Choice - Example



Limdep: Truncated/Censored Models

• Truncated variables:

We only sample from (observe/use) a subset of the population. The variable is observed only beyond a certain threshold level ('*truncation point*').

Examples: Store expenditures, Capex, labor force participation, income below poverty line.

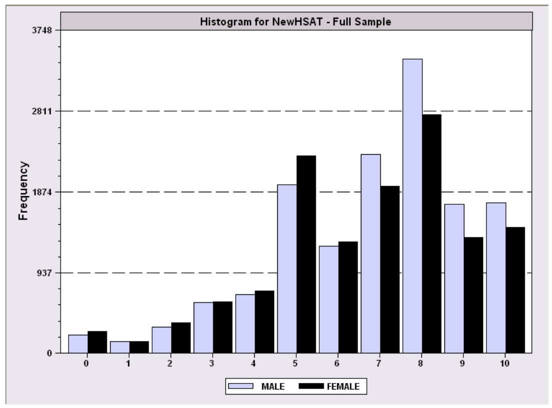
• Censored variables:

Values in a certain range are all transformed to/grouped into (or reported as) a single value.

Examples: hours worked, exchange rates under CB intervention.

Note: Censoring is a "defect" in the sample data. Presumably, if they were not censored, the data would be a representative sample from the population of interest.

Limdep: Censored Health Satisfaction Data



0 = Not Healthy **1 = Healthy**

Limdep: Duration/Hazard Models

• We model the time between two events.

Examples:
–Time between two trades.
–Time between cash flows withdrawals from a Mutual fund.
–Time until a consumer becomes inactive/cancels a subscription.
–Time until a consumer responds to direct mail or a questionnaire.

Microeconomics behind Discrete Choice

• Consumers maximize utility. The fundamental choice problem:

Max U(x_1, x_2, …)   s. t. prices and budget constraints

• A Crucial Result for the Classical Problem:

–Indirect Utility Function: V = V(**p**, I)

–Demand System of Continuous Choices

$$x_j^* = -\frac{\partial V(\mathbf{p}, \mathrm{I}) / \partial p_j}{\partial V(\mathbf{p}, \mathrm{I}) / \partial \mathrm{I}}$$

• The Integrability Problem: Utility is not revealed by demands.

Theory for Discrete Choice

• Theory is silent about discrete choices.

• Translation to discrete choice.

- Existence of well defined utility indexes: Completeness of rankings
- Rationality: Utility maximization
- Axioms of revealed preferences

• Choice and consideration sets: Consumers simplify choice situations

• Implication for choice among a set of discrete alternatives

• Commonalities and uniqueness

– Does this allow us to build “models?”

– What common elements can be assumed?

– How can we account for heterogeneity?

• Revealed choices do not reveal utility, only rankings which are scale invariant.

Discrete Choice Models (DCM)

• We will model discrete choice. We observe a discrete variable y_i and a set of variables connected with the decision x_i, usually called covariates. We want to model the relation between y_i and x_i.

• It is common to distinguish between covariates z_i that vary by units (individuals or firms), and covariates that vary by choice (and possibly by individual), w_{ij}.

Example of z_i's: individual characteristics, such as age or education.

Example of w_{ij}: the cost associated with the choice, for example the cost of investing in bonds/stocks/cash, or the price of a product.

• This distinction is important for the interpretation of these models using utility maximizing choice behavior. We may put restrictions on the way covariates affect utilities: the characteristics of choice i should affect the utility of choice i, but not the utility of choice j.

Discrete Choice Models (DCM)

• The modern literature goes back to the work by Daniel McFadden in the seventies and eighties (McFadden 1973, 1981, 1982, 1984).

• Usual Notation:

n = decision maker

i, j = choice options

y = decision outcome

x = explanatory variables/covariates

β = parameters

ε_i = error term

I[zz] = indicator function (= 1 if zz is true, 0 otherwise).

Example: I[$y = j \mid x$] = 1 if j was selected (given x)

= 0 otherwise

DCM – What Can we Learn from the Data?

• Q: Are the characteristics of consumers/firms relevant?

• Predicting behavior

- Individual – for example, will a person buy the add-on insurance?

- Aggregate – for example, what proportion of the population will buy the add-on insurance?

• Analyze changes in behavior when attributes change. For example, how will changes in education change the proportion of who buy the insurance?

Application: Health Care Usage (Greene)

German Health Care Usage Data, N = 7,293, Varying Numbers of Periods
Data downloaded from Journal of Applied Econometrics Archive. This is an unbalanced panel with 7,293 individuals. This is a large data set. There are altogether 27,326 observations. The number of observations ranges from 1 to 7. (Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987). (Downloaded from the JAE Archive)

Variables in the file are

DOCTOR	=	**1(Number of doctor visits > 0)**
HOSPITAL	=	**1(Number of hospital visits > 0)**
HSAT	=	**health satisfaction, coded 0 (low) - 10 (high)**
DOCVIS	=	**number of doctor visits in last three months**
HOSPVIS	=	**number of hospital visits in last calendar year**
PUBLIC	=	**insured in public health insurance = 1; otherwise = 0**
ADDON	=	**insured by add-on insurance = 1; otherswise = 0**
HHNINC	=	**household nominal monthly net income in German marks / 10000**. (4 observations with income=0 were dropped)
HHKIDS	=	**children under age 16 in the household = 1; otherwise = 0**
EDUC	=	**years of schooling**
AGE	=	**age in years**
FEMALE	=	**1 for female headed household, 0 for male**
EDUC	=	**years of education**

Application: Binary Choice Data (Greene)

Listing of raw data (Current sample)

Line	Observ.	ID	DOCTOR	AGE	HHNINC	FEMALE
1	29	9	0	40	.75000	0
2	31	10	0	36	.92000	1
3	34	11	1	43	.20000	1
4	38	12	1	51	.45000	1
5	42	13	1	36	.62500	0
6	49	14	0	46	.70000	0
7	52	15	1	38	.70000	1
8	58	16	0	46	.20000	1
9	83	21	0	48	.55000	0
10	90	22	0	48	.41200	1
11	109	28	0	47	.25000	0
12	116	30	0	62	.18000	0
13	125	32	0	47	.62000	1
14	132	33	1	64	.09700	0
15	154	44	1	49	.48000	0
16	158	45	1	35	.48000	1
17	177	51	0	39	.38000	0
18	184	52	1	38	.38000	1
19	191	53	0	57	.61000	0
20	201	55	1	57	.36000	1
21	209	57	1	55	.23000	1
22	215	58	1	60	.52000	0
23	220	59	1	53	.44000	1
24	223	60	1	31	.62000	1
25	233	62	0	56	.33000	1

Application: Health Care Usage (Greene)

Q: Does income affect doctor's visits? What is the effect of age on doctor's visits? Is gender relevant?

27,326 Observations –

- 1 to 7 years, panel
- 7,293 households observed
- We use the 1994 year => 3,337 household observations

Descriptive Statistics

Variable	Mean	Std.Dev.	Minimum	Maximum
DOCTOR	.657980	.474456	.000000	1.00000
AGE	42.6266	11.5860	25.0000	64.0000
HHNINC	.444764	.216586	.340000E-01	3.00000
FEMALE	.463429	.498735	.000000	1.00000

DCM: Setup – Choice Set

1. Characteristics of the choice set

- Alternatives must be mutually exclusive: No combination of choice alternatives. For example, no combination of different investments types (bonds, stocks, real estate, etc.).

- Choice set must be exhaustive: all *relevant* alternatives included. If we are considering types of investments, we should include all: bonds; stocks; real estate; hedge funds; exchange rates; commodities, etc. If relevant, we should include international and domestic financial markets.

- Finite (countable) number of alternatives.

DCM: Setup – RUM

2. Random utility maximization (**RUM**)

Assumption: Revealed preference. The decision maker selects the alternative that provides the highest utility. That is,

Decision maker n selects choice i if $U_{ni} > U_{nj}\ \ \forall\, j \neq i$

Decomposition of utility: A deterministic (observed), V_{nj}, and random (unobserved) part, ε_{nj}:

$$U_{nj} = V_{nj} + \varepsilon_{nj}$$

- The deterministic part, V_{nj}, is a function of some observed variables, $\boldsymbol{x}_{nj}$ (age, income, sex, price, etc.):

$$V_{nj} = \alpha + \beta_1\, Age_n + \beta_2\, Income_{nj} + \beta_3\, Sex_n + \beta_4\, Price_{nj}$$

- The random part, ε_{nj}, follows a distribution. For example, a normal.

DCM: Setup – RUM

2. RUM (continuation)

• We think of an individual's utility as an unobservable variable, with an observable component, V_n, and an unobservable (tastes?) random component, ε_n.

• The deterministic part is usually intrinsic linear in the parameters:

$$V_{nj} = \alpha + \beta_1 Age_n + \beta_2 Income_{nj} + \beta_3 Sex_n + \beta_4 Price_{nj}$$

- In this formulation, the parameters, β, are the same for all individuals. There is no heterogeneity. This is a useful assumption for estimation. It can be relaxed.

DCM: Setup - RUM

2. RUM (continuation)

<u>Probability Model</u>: Since both U's are random, the choice is random. Then, n selects i over j if:

$$\begin{aligned} P_{ni} &= \text{Prob } (U_{ni} > U_{nj} \ \forall j \neq i) \\ &= \text{Prob } (V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall j \neq i) \\ &= \text{Prob } (\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \ \forall j \neq i) \end{aligned}$$

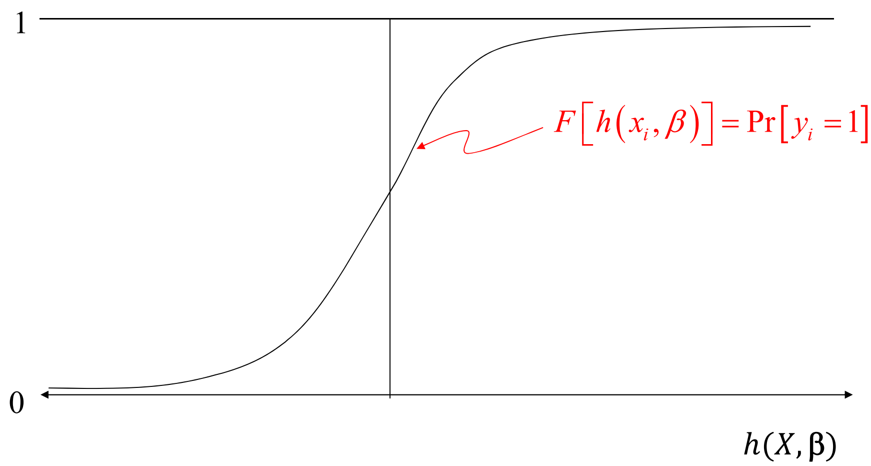
$$P_{ni} = \int I\left[\varepsilon_{nj} - \varepsilon_{ni} > V_{ni} - V_{nj}, \forall i \neq j\right] f(\varepsilon_n)\, d\varepsilon_n$$

$\Rightarrow V_{nj} = F(X, \beta)$ is a CDF.

• $V_{ni} - V_{nj} = h(X, \beta)$. $h(.)$ is usually referred as the *index function*.

• To evaluate the CDF, $F(X, \beta)$, $f(\varepsilon_n)$: needs to be specified.

DCM: Setup - RUM



DCM: Setup - RUM

• To evaluate the integral, $f(\varepsilon_n)$ needs to be specified. Many possibilities:

– Normal: **Probit Model**, natural for behavior.

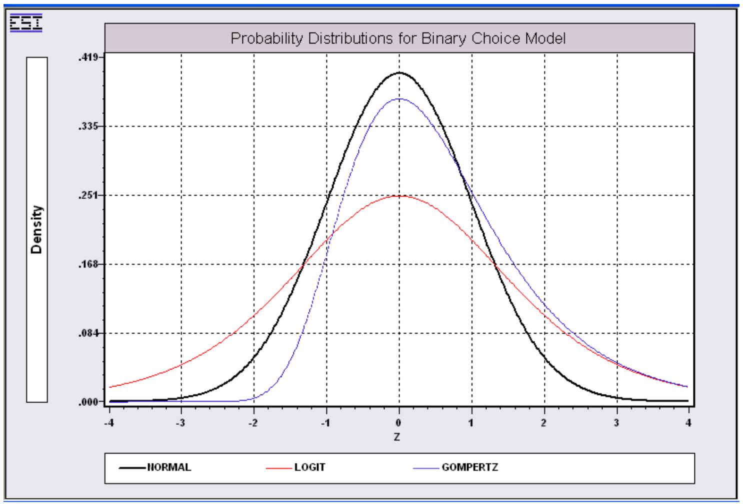
– Logistic: **Logit Model**, allows "thicker tails."

– Gompertz: **Extreme Value Model**, asymmetric distribution.

• We can use non-parametric or semiparametric methods to estimate the CDF F(X, β). These methods impose weaker assumptions than the fully parametric model described above.

• In general, there is a trade-off: Less assumptions, weaker conclusions, but likely more robust results.

DCM: Setup – RUM – Different $f(\varepsilon_n)$



DCM: Setup - RUM

• Note: Probit? Logit?

A one standard deviation change in the argument of a standard Normal distribution function is usually called a "Probability Unit" or *Probit* for short. "Probit" graph papers have a normal probability scales on one axis.

The Normal qualitative choice model became known as the *Probit* model. The "it" was transmitted to the Logistic Model (Logit) and the Gompertz Model (Gompit).

DCM: Setup - Distributions

• Many candidates for CDF –i.e., $P_n(x_n\beta)$ = F(Z_n),:
  – Normal (**Probit Model**) = $\Phi(Z_n)$
  – Logistic (**Logit Model**) = 1/[1+exp(-Z_n)]
  – Gompertz (**Gompit Model**) = 1 – exp[-exp(Z_n)]

• Suppose we have binary (0, 1) data. Assume β > 0.

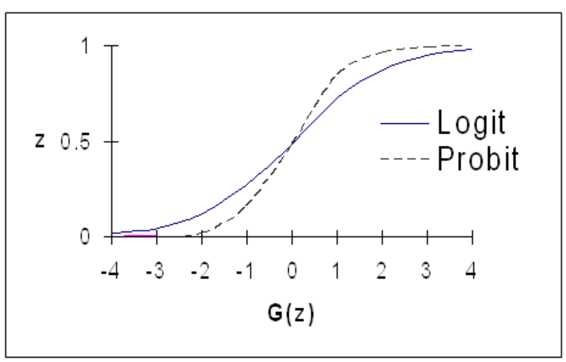
- **Probit Model**: Prob(y_n = 1) approaches 1 very rapidly as X and therefore Z increase. It approaches 0 very rapidly as X & Z decrease.

- **Logit Model**: It approaches the limits 0 and 1 more slowly than does the Probit.

- **Gompit Model**: Its distribution is strongly negatively skewed, approaching 0 very slowly for small values of Z, and 1 even more rapidly than the Probit for large values of Z.

DCM: Setup - Distributions

• Comparisons: Probit vs Logit



DCM: Setup - Normalization

Note: Not all the parameters may be identified.

• Suppose we are interested in whether an agent chooses to visit a doctor or not –i.e., (0, 1) data.

If $U_{visit} > 0$, an agent visits a doctor, set $Y = 1$ if $U_{visit} > 0$. Then,

$U_{visit} > 0 \Leftrightarrow \alpha + \beta_1\, Age + \beta_2\, Income + \beta_3\, Sex + \varepsilon > 0$

$\Rightarrow \varepsilon > -(\alpha + \beta_1 \text{ Age} + \beta_2 \text{ Income} + \beta_3 \text{ Sex})$

where ε has zero mean and Var[ε] = σ^2.

• Now, divide everything by σ.

$U_{visit} > 0 \quad \Leftrightarrow \frac{\varepsilon}{\sigma} > -[\frac{\alpha}{\sigma} + \frac{\beta_1}{\sigma} Age + \frac{\beta_2}{\sigma} Income + \frac{\beta_3}{\sigma} Sex] > 0$

or $\quad w > -[\alpha + \beta_1\, Age + \beta_2\, Income + \beta_3\, Sex] > 0$

DCM: Setup - Normalization

• $Y = 1$ if $U_{visit} > 0$

$U_{visit} > 0 \quad \Leftrightarrow \frac{\varepsilon}{\sigma} > -[\frac{\alpha}{\sigma} + \frac{\beta_1}{\sigma} Age + \frac{\beta_2}{\sigma} Income + \frac{\beta_3}{\sigma} Sex] > 0$

or $\quad w > -[\alpha + \beta_1\, Age + \beta_2\, Income + \beta_3\, Sex] > 0$

where Var[w] = 1.

Same data. The data contain no information about the variance. We could have assigned the values (1, 2) instead of (0, 1) to y_n. It is possible to produce any range of values in y_n.

• Normalization: Assume Var[ε] = 1.

DCM: Setup - Aggregation

Note: Aggregation can be problematic

- Biased estimates when aggregate values of the explanatory variables are used as inputs:

$$E[P_1(x_i)] \neq P_1[E[x_i]]$$

- But, when the sample is exogenously determined, consistent estimates can be obtained by sample enumeration:
  - Compute probabilities/elasticities for each decision maker
  - Compute (weighted) average of these values.

$$P_1 = \frac{\sum_{i=1}^{N} P_1(x_i)}{N}$$

• More on this later.

DCM: Setup – Aggregation

Example (from Train (2002)): Suppose there are two types of individuals, a and b, equally represented in the population, with

$$V_a = \beta' x_a$$
$$V_b = \beta' x_b$$

then

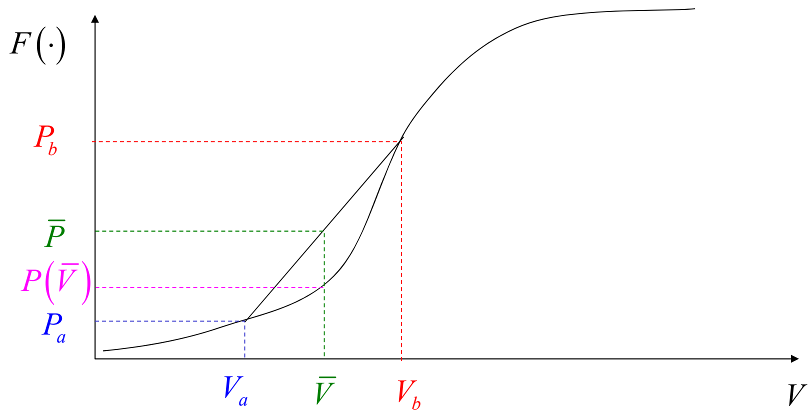
$$P_a = \Pr\left[y_i = 1 \middle| x_a\right] = F\left[\beta' x_a\right] \qquad P_b = \Pr\left[y_i = 1 \middle| x_b\right] = F\left[\beta' x_b\right]$$

but,

$$\bar{P} = \frac{1}{2}(P_{\boldsymbol{a}} + P_{\boldsymbol{b}}) \neq P(\bar{x}) = F[\beta' \bar{\boldsymbol{x}}]$$

DCM: Setup – Aggregation

In general, $P(\bar{V})$ will tend to (underestimate) overestimate $\bar{P}$ when probabilities are (high) low.



DCM: Setup - Aggregation

Graph: Average probability (2.1) vs. Probability of the average (2.2)

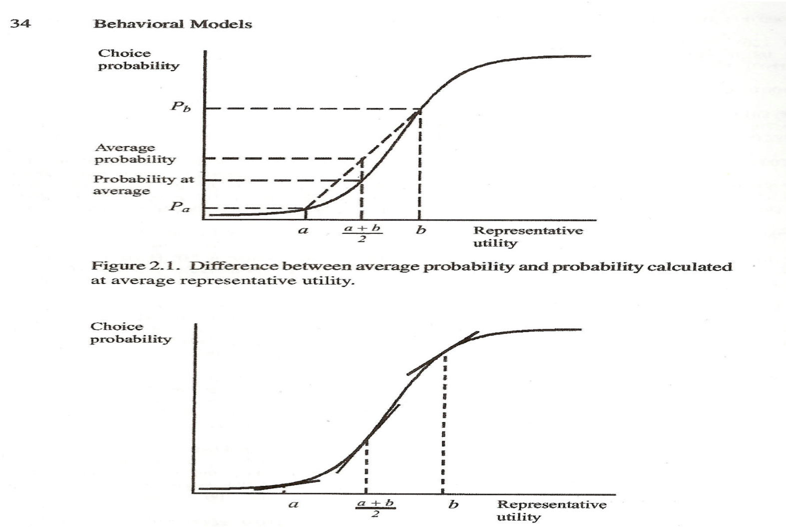


Figure 2.1. Difference between average probability and probability calculated at average representative utility.

Figure 2.2. Difference between average response and response calculated at average representative utility.

DCM: Setup - Identification

3. Identification problems

a. Only differences in utility matter

Choice probabilities do not change when a constant is added to each alternative's utility.

Implication: Some parameters cannot be identified/estimated.

Alternative-specific constants; coefficients of variables that change over decision makers but not over alternatives.

b. Overall scale of utility is irrelevant

Choice probabilities do not change when the utility of all alternatives are multiplied by the same factor.

Implication: Coefficients of different models (data sets) are not directly comparable.

Normalization of parameters and/or Var[ε] done for identification.

DCM: Estimation

• Since we specify a pdf, ML estimation seems natural to do. But, it can get complicated.

• In general, we assume the following distributions:

– Normal: **Probit Model** = $\Phi(x_n{}'\beta)$

– Logistic: **Logit Model** = $\dfrac{\exp(x_n{}'\beta)}{1+\exp(x_n{}'\beta)}$

– Gompertz: **Extreme Value Model** = $1 - \exp[-\exp(x_n{}'\beta)]$

• Methods

- ML estimation (Numerical optimization)
- Bayesian estimation (MCMC methods)
- Simulation-assisted estimation

DCM: ML Estimation

Example: Logit Model

Suppose we have binary (0, 1) data. The logit model follows from:

$$P[y_n = 1 | \ x] = \frac{exp(x_n'\boldsymbol{\beta})}{1 + exp(x_n'\boldsymbol{\beta})} = F(\boldsymbol{x}_n'\boldsymbol{\beta})$$

$$P[y_n = 0 | \ x] = \frac{1}{1 + exp(x_n'\boldsymbol{\beta})} = 1 - F(\boldsymbol{x}_n'\boldsymbol{\beta})$$

- **Likelihood function**

$$L(\beta) = \prod_n \ (1 - P[y_n = 1 | \ x]) * P[y_n = 1 | \ x]$$

- **Log likelihood**

$$\text{Log } L(\beta) = \sum_{n\ (with\ y=0)} log(1 - F(\boldsymbol{x}_n'\boldsymbol{\beta})) + \sum_{n\ (with\ y=1)} log(F(\boldsymbol{x}_n'\boldsymbol{\beta}))$$

- Numerical optimization to get β.

DCM: ML Estimation

• The usual problems with numerical optimization apply. The computation of the Hessian, **H**, may cause problems.

• Recall, ML estimators are consistent, asymptotic normal and efficient. These properties are the big appeal of MLE.

DCM: ML Estimation – Covariance Matrix

• How can we estimate the covariance matrix, $\boldsymbol{\Sigma}_{\beta_1}$?
Using the usual conditions, we can use the information matrix:

$$\text{In general: } \Sigma_{\beta 1} = \left[-E \frac{\partial^2 L}{\partial \beta \partial \beta'} \right]^{-1} = I(\beta)^{-1}$$

$$\text{Newton-Raphson: } \Sigma_{\beta 1} = \left[- \frac{\partial^2 L}{\partial \beta \partial \beta'} \right]^{-1}_{\beta = \beta_1}$$

$$\text{BHHH: } \Sigma_{\beta 1} = \left[\sum_{i=1}^{T} \frac{\partial L_i}{\partial \beta} \frac{\partial L_i}{\partial \beta'} \right]^{-1}_{\beta = \beta_1}$$

• The NR and BHHH are asymptotically equivalent, but, in small samples they often produce different estimates for the same model.

39

DCM: ML Estimation

• Numerical optimization - Steps:

(1) Start by specifying the likelihood for one observation: $F_n(\text{X}, \boldsymbol{\beta})$

(2) Get the joint likelihood function: $\text{L}(\boldsymbol{\beta}) = \prod_n F_n(\text{X}, \boldsymbol{\beta})$

(3) It is easier to work with the log likelihood function:

$$\text{Log L}(\boldsymbol{\beta}) = \sum_n log(F_n(\text{X}, \boldsymbol{\beta}))$$

(4) Maximize Log L($\boldsymbol{\beta}$) with respect to $\boldsymbol{\beta}$

- Set the score equal to 0 $\Rightarrow$ no closed-form solution.

- Numerical optimization, as usual:

 (i) Starting values $\boldsymbol{\beta}_0$.

 (ii) Determine new value $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t$ + update, such that

$$\text{Log L}(\boldsymbol{\beta}_{t+1}) > \text{Log L}(\boldsymbol{\beta}_t).$$

 Say, N-R's updating step: $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \lambda_t \boldsymbol{H}^{-1} \nabla f(\boldsymbol{\beta}_t)$

 (iii) Repeat step (ii) until convergence.

DCM: Bayesian Estimation

• The Bayesian estimator will be the mean of the posterior density:

$$f(\boldsymbol{\beta},\gamma \mid \mathbf{y},\mathbf{X}) = \frac{f(\mathbf{y} \mid \mathbf{X},\boldsymbol{\beta},\gamma) f(\boldsymbol{\beta},\gamma)}{f(\mathbf{y} \mid \mathbf{X},\boldsymbol{\beta},\gamma)} = \frac{f(\mathbf{y} \mid \mathbf{X},\boldsymbol{\beta},\gamma) f(\boldsymbol{\beta},\gamma)}{\int f(\mathbf{y} \mid \mathbf{X},\boldsymbol{\beta},\gamma) f(\boldsymbol{\beta},\gamma) d\boldsymbol{\beta} d\gamma}$$

- $f(\boldsymbol{\beta},\gamma)$ is the prior density for the model parameters
- $f(y \mid \mathbf{X},\boldsymbol{\beta},\gamma)$ is the likelihood.

• As usual we need to specify the prior and the likelihood:
  - The priors are usually non-informative (flat), say $f(\boldsymbol{\beta},\gamma) \propto 1$.
  - The likelihood depends on the model in mind. For a Probit Model, we will use a normal distribution. If we have binary data, then,

$$f(y \mid \mathbf{X},\boldsymbol{\beta},\gamma) = \Pi_n \,(1 - \Phi[y_n \mid x,\boldsymbol{\beta},\gamma])\, \Phi[y_n \mid x,\boldsymbol{\beta},\gamma]$$

DCM: Bayesian Estimation

• Let $\boldsymbol{\theta} = (\boldsymbol{\beta},\gamma)$. Suppose we have binary data with

$$P_n[y_n = 1 \mid x, \boldsymbol{\theta}] = F_n(\mathrm{X}, \boldsymbol{\theta}).$$

The estimator of $\boldsymbol{\theta}$ is the mean of the posterior density.

Under a flat prior assumption:

$$E[\theta \mid \mathbf{y},\mathbf{X}] = \frac{\int \theta f(\mathbf{y} \mid \mathbf{X},\theta) f(\theta) d\theta}{\int f(\mathbf{y} \mid \mathbf{X},\theta) f(\theta) d\theta} = \frac{\int \theta \prod_{n=1}^{N} (1 - F(\mathbf{X},\theta))^{y_n} (F(\mathbf{X},\theta))^{y_n} d\theta}{\int \prod_{n=1}^{N} (1 - F(\mathbf{X},\theta))^{y_n} (F(\mathbf{X},\theta))^{y_n} d\theta}$$

• Evaluation of the integrals is complicated. We evaluate them using MCMC methods. Much simpler.

MP Model – Simulation-based Estimation

• ML Estimation is likely complicated due to the multidimensional integration problem. Simulation-based methods approximate the integral. Relatively easy to apply.

• Simulation provides a solution for dealing with problems involving an integral. For example:

$$\mathrm{E}[h(u)] = \int h(u)\, f(u)\, du.$$

• All GMM and many ML problems require the evaluation of an expectation. In many cases, an analytic solution or a precise numerical solution is not possible. But, we can always simulate E[$h(u)$]:

- Steps
  - Draw R *pseudo*-*RV* from $f(u)$: $u^1, u^2, \ldots, u^R$ (R: repetitions)
  - Compute Ê[$h(u)$] = $(1/R)\sum_{n=1}^{R} h(u^n)$

MP Model – Simulation-based Estimation

• We call Ê[$h(u)$] a *simulator*.

• If $h(.)$ is continuous and differentiable, then Ê[$h(u)$] will be continuous and differentiable.

• Under general conditions, Ê[$h(u)$] provides an unbiased (& most of the times consistent) estimator for $\mathrm{E}[h(u)]$.

• The variance of Ê[$h(u)$] is equal to Var[$h(u)$] $/R$.

• There are many simulators. But the idea is the same: compute an integral by drawing pseudo-RVs, never by integration.

DCM: Partial Effects

• In general, the β's do not have an interesting interpretation. β_k does not have the usual *marginal effect* interpretation.

• To make sense of β's, we calculate:

$$\text{Partial effect} = \frac{\delta P(\alpha + \beta_1\, Income + \ldots)}{\delta x_k} \qquad \text{(derivative)}$$

$$\text{Marginal effect} = \frac{\text{E}[y_n|x]}{\delta x_k}$$

$$\text{Elasticity} = \frac{\delta log P(\alpha + \beta_1\, Income + \ldots)}{\delta \log(x_k)}$$

$$= \text{Partial effect} * \frac{x_k}{P(\alpha + \beta_1\, Income + \ldots)}$$

• These effects vary with level of $\boldsymbol{x}$: larger near the center of the distribution, smaller in the tail.

• Use delta method to calculate standard errors for these effects.

DCM: Partial Effects – Delta Method

• We know the distribution of b_n, with mean θ and variance σ^2/n, but we are interested in the distribution of $g(b_n)$, where $g(b_n)$ is a continuous differentiable function, independent of n.)

• After some work ("inversion"), we obtain:

$$g(b_n) \xrightarrow{a} \text{N}(g(\theta), [g'(\theta)]^2\, \sigma^2/n)$$

When b_n is a vector, $g(\boldsymbol{b}_n) \xrightarrow{a} \text{N}(g(\boldsymbol{\theta}), [G(\boldsymbol{\theta})]'\, \text{Var}[\boldsymbol{b}_n]\, [G(\boldsymbol{\theta})])$,

where $[G(\boldsymbol{\theta})]$ is the Jacobian of $g(.)$.

• In the DCM case, $g(b_n) = F(x_n{}'\boldsymbol{\beta})$

• Note: A bootstrap can also be used.

DCM: Partial Effects – Sample Means or Average

• The partial and marginal effects will vary with the values of $\boldsymbol{x}$.

• It is common to calculate these values at, say, the sample means of the $\boldsymbol{x}$. For example:

Estimated Partial effect = $f(\alpha + \beta_1\, Income + \ldots)$ -$f(.)$ =pdf

• The marginal effects can also be computed as the average of the marginal effects at every observation.

• In principle, different models will have different effects.

• Practical Question: Does it make a difference the P(.) used?

DCM: Goodness of Fit

• Q: How well does a DCM fit?

In the regression framework, we used R^2. But, with DCM, there are no residuals or RSS. The model is not computed to optimize the fit of the model: $\Rightarrow$ There is no R^2.

• "Fit measures" computed from log L:

- Let Log L(β_0) only with constant term. Then, define Pseudo R^2:

  Pseudo R^2 = 1 – Log L(β)/Log L(β_0) ("*likelihood ratio index*")

  (This McFadden's Pseudo R^2 . There are many others.)

- LR-test : LR = -2(Log L(β_0) – Log L(β)) ~ χ^2_k

- Information Criterion: AIC, BIC

  $\Rightarrow$ sometimes conflicting results

DCM: Goodness of Fit

• "Fit measures" computed from accuracy of predictions

Direct assessment of the effectiveness of the model at predicting the outcome –i.e., a 1 or a 0.

- Computation
  - Use the model to compute predicted probabilities
  - Use the model and a rule to compute predicted y = 0 or 1

    Rule: Predict y = 1 if estimated F is "large", say 0.5 or greater

    More general, use $\hat{y}$ = 1 if estimated F is greater than P*

Example: Cramer Fit measure

$\hat{F}$ = Predicted Probability

$$\hat{\lambda} = \frac{\Sigma_{i=1}^{N} y_i \hat{F}}{N1} - \frac{\Sigma_{i=1}^{N} (1-y_i)\hat{F}}{N0}$$

$\hat{\lambda}$ = Mean $\hat{F}$ | when $y = 1$ - Mean $\hat{F}$ | when $y = 0$

= reward for correct predictions minus penalty for incorrect predictions

Cross Tabulation of Hits and Misses

Let

$$\hat{y}_i = \begin{cases} 1 & \hat{F}_i \geq 0.5 \\ 0 & \hat{F}_i < 0.5 \end{cases}$$

		Predicted	
		$\hat{F}_i \geq 0.5$	$\hat{F}_i < 0.5$
Actual	$y_i = 1$		
	$y_i = 0$		

• The prediction rule is arbitrary.

– No weight to the costs of individual errors made. It may be more costly to make an error to classify a "yes" as a "no" than viceversa.

– In this case, some loss functions would be more helpful than others.

• There is no way to judge departures from a diagonal table.

DCM: Model Selection

• Model selection based on nested models:

  • Use the Likelihood:

  - LR-test

    $LR = -2(\text{Log } L(\beta_r) – \text{Log } L(\beta_u))$

    r=restricted model; u=unrestricted (full) model

    $LR \sim \chi^2_k$ (k = difference in # of parameters)

• Model selection based for non-nested models:

  • AIC, CAIC, BIC ⇒ lowest value

DCM: Testing

• Given the ML estimation setup, the trilogy of tests (LR, W, and LM) is used:

- LR Test: Based on unrestricted and restricted estimates.
- Distance Measures - Wald test: Based on unrestricted estimates.
- LM tests: Based on restricted estimates.

• Chow Tests that check the constancy of parameters can be easily constructed.

- Fit an unrestricted model, based on model for the different categories (say, female and male) or subsamples (regimes), and compare it to the restricted model (pooled model) ⇒ LR test.

DCM: Testing

• Issues:
- Linear or nonlinear functions of the parameters
- Constancy of parameters (Chow Test)
- Correct specification of distribution
- Heteroscedasticity

• Remember, there are no residuals. There is no F statistic.

DCM: Heteroscedasticity

• In the RUM, with binary data agent n selects

$$y_n = 1 \text{ iff } \qquad U_n = \boldsymbol{x}_n{}'\boldsymbol{\beta} + \varepsilon_n > 0,$$

where the unobserved ε_n has E[ε_n] = 0, and Var[ε_n] = 1

• Given that the data do not provide information on σ, we assume Var[ε_n] = 1, an identification assumption. But, implicitly we are assuming homoscedasticity across individuals.

• Q: Is this a good assumption?

• The RUM framework resembles a regression, where in the presence of heteroscedasticity, we scale each observation by the squared root of its variance.

DCM: Heteroscedasticity

• Q: How to accommodate heterogeneity in a DCM?

Use different scaling for each individual. We need to know the model for the variance.

– Parameterize: $\text{Var}[\varepsilon_n] = \exp(\boldsymbol{z}_n{}'\,\gamma)$

– Reformulate probabilities

Binary Probit or Logit: $P_n[y_n = 1 \,|\, \boldsymbol{x}] = P(\boldsymbol{x}_n{}'\beta/\exp(\boldsymbol{z}_n{}'\,\gamma))$

• Marginal effects (derivative of $E[y_n]$ w.r.t. $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$) are now more complicated. If $\boldsymbol{x}_n = \boldsymbol{z}_n$, signs and magnitudes of marginal effects tend to be ambiguous.

DCM: Heteroscedasticity - Testing

• There is no generic, White-type test for heteroscedasticity. We do the tests in the context of the maximum likelihood estimation.

• Likelihood Ratio, Wald and Lagrange Multiplier Tests are all straightforward

• All heteroscedasticity tests require a specification of the model under H_1 (heteroscedasticity), say,

$$H_1\text{: Var}[\varepsilon_n] = \exp(\boldsymbol{z}_n{}'\,\gamma)$$

DCM: Robust Covariance Matrix (Greene)

• In the context of maximum likelihood estimation, it is common to define the Var[$\mathbf{b_M}$] = $(1/T)\ \boldsymbol{H_0}^{-1}\,\boldsymbol{V_0}\ \boldsymbol{H_0}^{-1}$, where if the model is correctly specified: -**H** = **V**. Similarly, for a DCM we can define:

"Robust" Covariance Matrix: $\mathbf{V} = \mathbf{A\,B\,A}$

$\mathbf{A}$ = negative inverse of second derivatives matrix

$$= \text{ estimated E}\left[-\frac{\partial^2 \log L}{\partial\beta\partial\beta'}\right]^{-1} = \left[-\sum_{i=1}^{N}\frac{\partial^2 \log \text{Prob}_i}{\partial\hat{\beta}\partial\hat{\beta}'}\right]^{-1}$$

$\mathbf{B}$ = matrix sum of outer products of first derivatives

$$= \text{ estimated E}\left[\frac{\partial \log L}{\partial\beta}\frac{\partial \log L}{\partial\beta'}\right] = \left[\sum_{i=1}^{N}\frac{\partial \log \text{Prob}_i}{\partial\hat{\beta}}\frac{\partial \log \text{Prob}_i}{\partial\hat{\beta}'}\right]^{-1}$$

For a logit model, $\mathbf{A} = \left[\sum_{i=1}^{N}\hat{P}_i(1-\hat{P}_i)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$

$$\mathbf{B} = \left[\sum_{i=1}^{N}(y_i-\hat{P}_i)^2\mathbf{x}_i\mathbf{x}_i'\right] = \left[\sum_{i=1}^{N}e_i^2\mathbf{x}_i\mathbf{x}_i'\right]$$

(Resembles the White estimator in the linear model case.)

DCM: Robust Covariance Matrix (Greene)

• Q: Is this matrix robust to what?

• It is not "robust" to:
  – Heteroscedasticity
  – Correlation across observations
  – Omitted heterogeneity
  – Omitted variables (even if orthogonal)
  – Wrong functional form for index function

• In all cases, the estimator is inconsistent so a "robust" covariance matrix is pointless.

• (In general, it is merely harmless.)

DCM: Endogeneity

• It is possible to have in a DCM endogenous covariates. For example, many times we include education as part of an individual's characteristics or the income/benefits generated by the choice as part of its characteristics.

• Now, we divide the covariates in endogenous and exogenous. Suppose agent n selects $y_n = 1$ iff

$$U_n = \boldsymbol{x}_n{}'\boldsymbol{\beta} + \boldsymbol{h}_n{}'\boldsymbol{\theta} + \varepsilon_n > 0,$$

where E[$\varepsilon_n \mid h$] ≠ 0 (n is endogenous)

• There are two cases:
  – Case 1: h is continuous (complicated)
  – Case 2: h is discrete, say, binary. (Easier, a treatment effect)

DCM: Endogeneity

• Approaches
 - Maximum Likelihood (parametric approach)
 - GMM
 - Various approaches for case 2. Case 2 is the easier case: SE DCM!

• Concentrate on Case 1 (h is continuous).
The usual problems with endogenous variables are made worse in nonlinear models. In a DCM is not clear how to use IVs.

• If moments can be formulated, GMM can be used. For example, in a Probit Model: $E[(y_n - \Phi(\boldsymbol{x}_n{}'\,\boldsymbol{\beta}))(\boldsymbol{x}_n\,\boldsymbol{z})]=0$
⇒ This moment equation forms the basis of a straightforward two step GMM estimator. Since we specify Φ(.), it is parametric.

DCM: Endogeneity - ML

• ML estimation requires full specification of the model, including the assumption that underlies the endogeneity of $\boldsymbol{h}_n$. For example:

- RUM: $U_n = \boldsymbol{x}_n{'}\boldsymbol{\beta} + \boldsymbol{h}_n{'}\theta + \varepsilon_n$
- Revealed preference: $y_n = 1[U_n > 0]$
- Endogenous variable: $\boldsymbol{h}_n = \boldsymbol{z}_n' \boldsymbol{\alpha} + u_n$, with

$\text{E}[\varepsilon_n | \boldsymbol{h}] \neq 0 \Rightarrow \text{Cov}[u, \varepsilon] \neq 0$ $\quad$ ($\rho = \text{Corr}[u, \varepsilon]$)

- Additional Assumptions:

1) $\begin{bmatrix} \varepsilon_n \\ u_n \end{bmatrix} \xrightarrow{a} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma_u \\ \rho\sigma_u & \sigma_u^2 \end{bmatrix}\right)$

2) $\boldsymbol{z}$ = IV, a valid set of exogenous variables, uncorrelated with (u, ε).

• ML becomes a simultaneous equations model.

DCM: Endogeneity - ML

• ML becomes a simultaneous equations model.

- Reduced form estimation is possible:

  - Insert the second equation in the first. If we use a Probit Model, this becomes $P[y_n = 1 | \boldsymbol{x}_n, \boldsymbol{z}_n] = \Phi(\boldsymbol{x}_n{'}\boldsymbol{\beta}^* + \boldsymbol{z}_n{'}\alpha^*)$.

- FIML is probably simpler:

  - Write down the joint density: $f(y_n | \boldsymbol{x}_n, \boldsymbol{z}_n)\, f(\boldsymbol{z}_n)$
  - Assume probability model for $f(y_n | \boldsymbol{x}_n, \boldsymbol{z}_n)$, say a Probit Model.
  - Assume marginal for $f(\boldsymbol{z}_n)$, say a normal distribution.
  - Use the projection: $\varepsilon_n | u_n = [(\rho\sigma)/\sigma_u^2]\, u_n + v_n$, $\quad \sigma_v^2 = (1 - \rho^2)$.
  - Insert projection in $P(y_n)$
  - Replace $u_n = (\boldsymbol{h}_n - \boldsymbol{z}_n{'}\boldsymbol{\alpha})$ in $P(y_n)$.
  - Maximize Log L(.) w.r.t. $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \theta, \rho, \sigma_u)$

DCM: Endogeneity – ML: Probit (Greene)

Probit fit of y to **x** and *h* will not consistently estimate (β,θ) because of the correlation between h and ε induced by the correlation of u and ε. Using the bivariate normality,

$$\text{Prob}(y = 1 \mid \mathbf{x}, h) = \Phi\left[\frac{\boldsymbol{\beta}'\mathbf{x} + \theta h + (\rho / \sigma_u)u}{\sqrt{1-\rho^2}}\right]$$

Insert $u_i = (h_i - \boldsymbol{\alpha}'\mathbf{z})/\sigma_u$ and include f(h|**z**) to form logL

$$\text{logL} = \sum_{i=1}^{N} \left\{ \begin{array}{l} \log \Phi\left[(2y_i - 1)\left(\dfrac{\boldsymbol{\beta}'\mathbf{x}_i + \theta h_i + \rho\left(\dfrac{h_i - \boldsymbol{\alpha}'\mathbf{z}_i}{\sigma_u}\right)}{\sqrt{1-\rho^2}}\right)\right] + \\ \log \dfrac{1}{\sigma_u}\phi\left[\left(\dfrac{h_i - \boldsymbol{\alpha}'\mathbf{z}_i}{\sigma_u}\right)\right] \end{array} \right\}$$

DCM: Endogeneity – ML: 2-Step LIML

• Two step limited information ML (Control Function) is also possible:

- Use OLS to estimate **α**, σ_u
- Compute the residual v_n.
- Plug residuals v_n into the assumed model P(y_n)
- Fit the probability model for P(y_n).
- Transform the estimated coefficients into the structural ones.
- Use delta method to calculate standard errors.