TA Session 6: Heckman Selection and Multinomial Probit

Shuowen Chen[1]

EC708: PhD Econometrics I (Spring 2020)

[1]Parts of the materials are borrowed from Newey and McFadden (1994), Vella (1998), Puhani (2000), textbooks by Miranda and Fackler (2002), Train (2009) and Wooldridge (2010), and lecture slides by Gianluca Violante. I'm grateful to Yinuo Zhang for kindly sharing Violante's slides. I thank Iván Fernández-Val, Jean–Jacques Forneron and Hiroaki Kaido for helpful feedbacks.

Outline

- Sample Selection
  - Motivations and Clarifications
  - Heckman Selection Model
  - Two–Step Estimators as GMM
- Probit Estimation
  - Comparison with Logit
  - Newton–Cotes method
  - Gaussian Quadrature
  - Accept–Reject Algorithm
  - GHK (Geweke–Hajivassiliou–Keane) Estimator
  - Importance Sampling Interpretation of GHK Estimator

Returns of Schooling on Wage Rate

- Research question: How does years of schooling affect wages?
- Why OLS of wages on yrs of schooling might be problematic? For some individuals the decision to work is not random
  1. Sample (working people) is unrepresentative of population of interest (all people who received schooling)[2]
  2. Some component of work decision relevant to wage determination
  3. If component is fully controlled by observable characteristics, can still run OLS
  4. Component might be unobservable: e.g., people who don't work were offered rates below their reservation wages
  5. Reservation wage is arguably correlated with ability, which determines the wage
  6. Failure to account for this correlation brings in endogeneity problem and leads to incorrect estimation

[2] Unless assume sample of working people is chosen randomly from the population

When Does Sample Selection Matter?

Sample Selection Bias arises whenever one examines a subsample and the unobservable factors determining inclusion in the subsample are correlated with the unobservables influencing the variable of primary interest (Vella, 1998)

Heckman Selection Model

Let N and n denote whole sample and subsample with observed dependent variable

$$
\begin{aligned}
y_i^* &= x_i'\beta + \varepsilon_i \quad i = 1, \dots, N && (1)\\
d_i^* &= z_i'\gamma + \nu_i \quad i = 1, \dots, N && (2)\\
d_i &= \begin{cases} 1 & \text{if } d_i^* > 0 \\ 0 & \text{otherwise} \end{cases} && (3)\\
y_i &= y_i^* \times d_i && (4)
\end{aligned}
$$

- Eq (1) is of primary interest
- Eq (2): reduced form for latent variable capturing sample selection
- Eq (3): whether the dependent variable is observed
- Eq (4): observed outcomes (data)
- (ε, ν) independent of z with zero mean
- ε and ν are correlated

MLE for Heckman Selection

Assumption: ε_i and ν_i are i.i.d distributed $\mathcal{N}(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\nu} \\ \sigma_{\varepsilon\nu} & \sigma_\nu^2 \end{pmatrix}$$

and (ε_i, ν_i) are independent of z_i.

Average log likelihood function:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \Big\{ d_i \times \ln \Big[\int_{-z_i'\gamma}^{\infty} \phi_{\varepsilon\nu}(y_i - x_i'\beta, \nu) d\nu \Big] + (1 - d_i) \times \Big[\ln \int_{-z_i'\gamma}^{\infty} \int_{-\infty}^{\infty} \phi_{\varepsilon\nu}(\varepsilon, \nu) d\varepsilon d\nu \Big] \Big\}$$

- $\phi_{\varepsilon\nu}$: pdf of bivariate normal distribution
- Also known as Tobit type two model
- Fully efficient, but subject to misspecification
- Heckman (1976) proposes a simpler way for estimation

Correction Term of Heckman Selection

Consider the version that assumes **bivariate normality**[3]. Note:

$$\mathbb{E}[y_i|z_i, d_i = 1] = x_i'\beta + \mathbb{E}[\varepsilon_i|z_i, d_i = 1], \quad i = 1, \ldots, n$$

- $\mathbb{E}[\varepsilon_i|z_i, d_i = 1] = \frac{\sigma_{\varepsilon\nu}}{\sigma_\nu^2}\left\{\frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}\right\}$
- Blue term: **inverse Mills ratio** (λ_i)

Remarks:

- If $\sigma_{\varepsilon\nu} = 0$, no correction term, sample selection is not a problem
- Estimate γ by Probit over the entire sample N by MLE
- Plug in $\widehat{\gamma}$ for estimation of β (two–step estimation)
- Correction term plays the role of control function

[3] Other versions relax distribution assumptions or incorporate variable selections.

Two–Step Estimation Procedures & Remarks

1. Estimate γ using the entire sample via MLE, compute $\widehat{\lambda}$
2. Use the subsample with observed dependent variable to run OLS on the following specification:

$$y_i = x_i'\beta + \mu\widehat{\lambda}_i + \eta_i,$$

where $\mu := \sigma_{\varepsilon\nu}/\sigma_\nu^2$.

Remarks:

- Need to adjust for standard errors (Greene, 1981)
- Can formulate as GMM and get standard errors easily
- Identification concern: the function mapping the single index $z_i'\gamma$ into inverse Mills ratio is **linear** for certain ranges of the index
  - Needs exclusion restriction: at least one variable in z_i is not in x_i
  - Otherwise identification of β relies on the **nonlinear part** of the inverse Mills ratio
  - Leads to weak identification and inflated second–step SE

▸ Variance of η ▸ Graph of Inverse Mills Ratio

Two–Step Estimator as GMM

- Depends on some preliminary “first–step” estimator
- First–step estimation affects asymptotic variance of the second if consistency of the former affects that of the latter

Consider the following sample moments

$$\frac{1}{N}\sum_{i=1}^{N} g(w_i, \theta, \widehat{\gamma}) = 0,$$

where θ is parameter of interest and $\widehat{\gamma}$ is an estimator of γ, which satisfies the following sample moment condition

$$\frac{1}{N}\sum_{i=1}^{N} m(z_i, \gamma) = 0$$

Form $\widetilde{g}(w, \theta, \gamma) = \left[m(z, \theta)', g(w, \theta, \gamma)'\right]'$ and consider

$$\frac{1}{N}\sum_{i=1}^{N} \widetilde{g}(w_i, \widehat{\theta}, \widehat{\gamma}) = 0$$

Heckman Selection Model as Two–Step Estimator

Denote $\theta = (\beta', \mu)'$,

$$g(w, \theta, \gamma) = d \begin{bmatrix} x \\ \lambda(z'\gamma) \end{bmatrix} (y - x'\beta - \mu\lambda(z'\gamma)), \tag{5}$$

$$m(z, \gamma) = \lambda(z'\gamma)\Phi^{-1}(-z'\gamma)z(d - \Phi(z'\gamma)) \tag{6}$$

FOCs for OLS on the selected sample and probit estimation

Theorem 6.1 in Newey and McFadden (1994), can show that $\widehat{\theta}$ and $\widehat{\gamma}$ are asymptotically normal and $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$, where

$$V = G_\theta^{-1}\mathbb{E}[\{g(w, \theta_0, \gamma_0) + G_\gamma\psi(z)\}\{g(w, \theta_0, \gamma_0) + G_\gamma\psi(z)\}']G_\theta^{-1\prime}$$

- $G_\theta = \mathbb{E}[\nabla_\theta g(w, \theta_0, \gamma_0)]$
- $G_\gamma = \mathbb{E}[\nabla_\gamma g(w, \theta_0, \gamma_0)]$
- $\psi(z) = -\mathbb{E}[\nabla_\gamma m(z, \gamma_0)]^{-1} m(z, \gamma_0)$

How First–Step Estimation Affects Second–Step SE

- Suppose we only work with the following moment condition:

$$\frac{1}{N}\sum_{i=1}^{N} g(w_i, \theta, \gamma_0) = 0,$$

Then we have $Avar(\widehat{\theta}) = G_\theta^{-1}\mathbb{E}[g(w,\theta_0,\gamma_0)g(w,\theta_0,\gamma_0)']G_\theta^{-1\prime}$

- Simplified because we don't use first step $\widehat{\gamma}$
- Implication: unless $G_\gamma = 0$, first–step estimation $\widehat{\gamma}$ affects standard errors of $\widehat{\theta}$
- When does $G_\gamma \neq 0$? If inconsistency of $\widehat{\gamma}$ leads to inconsistency of $\widehat{\theta}$ (Newey & McFadden, 1994, Theorem 6.2)

First–Step Estimation in Heckman Selection

Denote $\lambda_\nu(\nu) = \frac{d\lambda(\nu)}{d\nu}$, G_γ in Heckman selection model is

$$G_\gamma = -\mu\mathbb{E}\left[d\begin{pmatrix} x \\ \lambda(z'\gamma_0)\end{pmatrix}\lambda_\nu(z'\gamma_0)z'\right],$$

- Generally nonzero unless $\mu = 0$, which implies $\sigma_{\varepsilon\nu} = 0$
- When does this happen? Sample selection doesn't matter

More on Asymptotic Variance

The correct asymptotic variance can be either larger or smaller than the one that ignores the first–step estimation

- Condition that correct asymptotic variance is larger:

$$\mathbb{E}[g(w,\theta_0,\gamma_0)m(z,\gamma_0)'] = 0$$

In this case $\mathbb{E}[g(w,\theta_0,\gamma_0)\psi(z)'] = 0$, and the correct variance is

$$G_\theta^{-1}\mathbb{E}[g(w,\theta_0,\gamma_0)g(w,\theta_0,\gamma_0)']G_\theta^{-1\prime} + G_\theta^{-1}G_\gamma\mathbb{E}[\psi(z)\psi(z)']G_\gamma' G_\theta^{-1\prime}$$

Heckman selection satisfies this condition since

$$\mathbb{E}[y - x'\beta_0 - \mu_0\lambda(z'\gamma_0)|x, d = 1, z] = 0$$

- Condition that correct asymptotic variance is smaller:

$$m(z,\gamma_0) = \nabla_\gamma \ln f(z|\theta_0,\gamma_0),$$

where f is likelihood of z. Somewhat rare to satisfy in practice.

Consistent Asymptotic Variance Estimation

- The advantage of writing two–step estimation as a GMM is to get consistent asymptotic variance when $G_\gamma \neq 0$
- Recall asymptotic variance of $\widehat{\theta}$:

$$V = G_\theta^{-1}\mathbb{E}[\{g(w,\theta_0,\gamma_0) + G_\gamma\psi(z)\}\{g(w,\theta_0,\gamma_0) + G_\gamma\psi(z)\}']G_\theta^{-1\prime}$$

- Plugged–in approach:

$$\widehat{G}_\theta = \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta g(w_i,\widehat{\theta},\widehat{\gamma}), \quad \widehat{G}_\gamma = \frac{1}{N}\sum_{i=1}^{N}\nabla_\gamma g(w_i,\widehat{\theta},\widehat{\gamma})$$

$$\widehat{g}_i = g(w_i,\widehat{\theta},\widehat{\gamma}), \quad \widehat{m}_i = m(z_i,\widehat{\gamma}) \quad \widehat{\psi}_i = -\left[\frac{1}{N}\sum_{i=1}^{N}\nabla_\gamma m(z_i,\widehat{\gamma})\right]^{-1}\widehat{m}_i$$

Consistent Asymptotic Variance Estimation

The estimator of asymptotic variance of $\widehat{\theta}$ is

$$\widehat{V} = \widehat{G}_\theta^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_i + \widehat{G}_\gamma\widehat{\psi}_i\right)\left(\widehat{g}_i + \widehat{G}_\gamma\widehat{\psi}_i\right)'\right]\widehat{G}_\theta^{-1\prime}$$

If moment functions are uncorrelated (recall this means first–step estimation increases second–step variance):

$$\mathbb{E}[g(w, \theta_0, \gamma_0)m(z, \gamma_0)'] = 0$$

Denote $\widehat{V}_\gamma = \frac{1}{N}\sum_{i=1}^{N}\widehat{\psi}_i\widehat{\psi}_i'$,

$$\widehat{V} = \widehat{G}_\theta^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_i\widehat{g}_i'\right]\widehat{G}_\theta^{-1\prime} + \widehat{G}_\theta^{-1}\widehat{G}_\gamma\widehat{V}_\gamma\widehat{G}_\gamma'\widehat{G}_\theta^{-1\prime}$$

Can be applied to obtain variance estimator of Heckman selection

▸ Form

Comparison Between Logit and Probit

- A standard logistic distribution has PDF:

$$f(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

  and variance $\pi^2/3$. Looks very similar to normal distribution.
- (Multinomial) logit is widely used in empirical IO due to analytical forms and fast computations, but also limited:
  1. Can't represent random taste variation
  2. Restrictive substitution patterns due to IIA[4]
  3. Can't be used in panel if unobserved factors correlated over time for each individual
- Probit probabilities don't have closed–form and requires numerical approximations.
- We explain why this is the case and cover some methods.

▸ Graph

[4] Independence of irrelevant alternatives

Multinomial Probit: Discrete Choice Utility

Consider individual n's additive utility of choosing one product among $\mathcal{J}$ options

$$U_{n,j} = V_{n,j} + \varepsilon_{n,j}$$

Assume $\varepsilon_n = (\varepsilon_{n,1}, ..., \varepsilon_{n,\mathcal{J}})$ is normally distributed with mean vector of zeros and var-cov matrix Ω. The probability of choose i is

$$\begin{aligned} P_{n,i} &= Pr(V_{n,i} + \varepsilon_{n,i} > V_{n,j} + \varepsilon_{n,j} \ \forall j \neq i) \\ &= \int \mathbb{1}(V_{n,i} + \varepsilon_{n,i} > V_{n,j} + \varepsilon_{n,j} \ \forall j \neq i)\phi(\varepsilon_n)d\varepsilon_n, \end{aligned}$$

where $\phi(\varepsilon_n) = \frac{1}{(2\pi)^{\mathcal{J}}|\Omega|^{1/2}} \exp\left(- \frac{1}{2}\varepsilon_n'\Omega^{-1}\varepsilon_n\right)$

The integration is over a vector, evaluated numerically. Essentially a numerical integration problem. Let's look at nonsimulation and simulation–based methods.

Newton–Cotes Method: Trapezoid Rule

Consider the following integration problem: $\int_a^b f(x)dx$. Approximate f with a piecewise linear polynomial $\tilde{f}$ whose integral is easy to compute:

$$\int_a^b f(x)dx \approx \int_a^b \tilde{f}(x)dx$$

1. Partition $[a, b]$ into n subintervals of equal length $h = \frac{b-a}{n}$ and endpoint nodes $x_k = a + kh$
2. For each node k, compute $y_k = f(x_k)$
3. Form a piecewise linear approximation of the function between successive points (x_k, x_{k+1}):

$$f(x) \approx f(x_k) + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k}(x - x_k)$$

Simple and robust, increasing nodes from N to $M \times N$ reduces error by factors of M^2 ▸ Graph

Newton–Cotes Method: Simpson Rule

Instead of using piecewise linear polynomials, use quadratic approximation

- Form a piecewise quadratic approximation $\tilde{f}$ that interpolates f at successive triplets (x_{k-1}, x_k, x_{k+1}) with quadratic functions
- Works better than trapezoid rule if f is smooth
- Works worse if f is non–differentiable at some points

▸ Graph

Gaussian Quadrature

Conisder a slightly more general integration problem: $\int_a^b f(x)w(x)dx$, where $w(\cdot)$ is a weight function (like *pdf*)

- Choose n nodes $x_1, \ldots, x_n$ and weights[5] $w_1, \ldots, w_n$ to match the $2n$ conditions: $\int_a^b x^k w(x)dx = \sum_{i=1}^n w_i x_i^k, \quad k = 0, \ldots, 2n-1$
- Integral approximation is thus $\int_a^b f(x)w(x)dx \approx \sum_{i=1}^n w_i f(x_i)$
- When $w(x) = 1$, called Gauss–Legendre quadrature
- When $w(x)$ is pdf, $\sum_{i=1}^n w_i x_i^k = EX^k$: discretize continuous r.v. with mass x_i and prob w_i and match moments
- For smooth integrands, converges exponentially fast as $n \uparrow$
- For more details, refer to *Numerical Recipes* by Press et al. (2007)

[5] Some common weight functions: uniform, normal, gamma, beta, etc.

Accept–Reject Algorithm

$$P_{n,i} = \int \mathbb{1}(V_{n,i} + \varepsilon_{n,i} > V_{n,j} + \varepsilon_{n,j} \; \forall j \neq i) \phi(\varepsilon_n) d\varepsilon_n$$

1. Draw $\varepsilon_n = (\varepsilon_{n,1}, \dots, \varepsilon_{n,\mathcal{J}})$ from a normal density with zero mean and var-cov Ω
2. Calculate $U_{n,j} = V_{n,j} + \varepsilon_{n,j} \; \forall j$
3. Determine if $U_{n,i} > U_{n,j} \; \forall j \neq i$. If yes, count as accept, otherwise count as reject
4. Repeat steps 1–3 many times (S)
5. Simulated probability is the proportion of draws that are accepted

Remarks:

- How to draw from a joint normal density? ▸ Procedure
- Simulated prob might be zero for any finite number of draws, especially if true prob is low
- Simulated prob not smooth, hinders numerical optimization

Smoothed Accept–Reject Simulators

Accept–Reject algorithm is not smooth because it uses 0–1 indicator, replace with a smooth and strictly positive function. Following McFadden (1989), we use logit function.

1. Draw $\varepsilon_n = (\varepsilon_{n,1}, \ldots, \varepsilon_{n,\mathcal{J}})$ from a normal density with zero mean and var-cov Ω
2. Calculate $U_{n,j} = V_{n,j} + \varepsilon_{n,j}\ \forall j$
3. Calculate $M = \frac{\exp(U_{n,i})/\lambda}{\sum_j \exp(U_{n,j})/\lambda}$
4. Repeat steps 1–3 many times (S)
5. Simulated probability is the average of Ms: $\widehat{P}_{n,i} = \frac{1}{S}\sum_{s=1}^{S} M_s$

Remark:

- λ: degree of smoothing, approaches Accept–Reject as $\lambda \to 0$

Geweke–Hajivassiliou–Keane (GHK) Simulator

- Based on sampling from recursive truncated normals after a Cholesky transformation ▸ Cholesky Decomposition
- First suggested by Geweke (1989), independently developed by Hajivassiliou (1992) and Keane (1994)
- Operates on utility differences: if we want to simulate $P_{n,i}$, need to subtract $U_{n,i}$ from the other utilities

Consider a three–alternative case and simulate $P_{n,1}$:

$$\widetilde{U}_{n,j,1} \equiv U_{n,j} - U_{n,1} = (V_{n,j} - V_{n,1}) + (\varepsilon_{n,j} - \varepsilon_{n,1}) \equiv \widetilde{V}_{n,j,1} + \widetilde{\varepsilon}_{n,j,1}$$

$\varepsilon_n = (\varepsilon_{n,1}, \varepsilon_{n,2}, \varepsilon_{n,3})'$ has distribution $\mathcal{N}(0, \Omega)$; $\widetilde{\varepsilon}_{n,1} = (\widetilde{\varepsilon}_{n,2,1}, \widetilde{\varepsilon}_{n,3,1})'$ has distribution $\mathcal{N}(0, \widetilde{\Omega}_1)$: $\widetilde{\Omega}_1 = A_1 \Omega A_1'$, where

$$A_1 = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

GHK Simulator: Preparation

- Consider the lower–triangular Cholesky factor of $\widetilde{\Omega}_1$,

$$L = \begin{pmatrix} c_{aa} & 0 \\ c_{ab} & c_{bb} \end{pmatrix}$$

- Rewrite $(\widetilde{\varepsilon}_{n,2,1}, \widetilde{\varepsilon}_{n,3,1})$ as

$$\widetilde{\varepsilon}_{n,2,1} = c_{aa}\eta_1, \quad \widetilde{\varepsilon}_{n,3,1} = c_{ab}\eta_1 + c_{bb}\eta_2$$

  where η_1 and η_2 are i.i.d. standard normal
- Hence we have

$$\widetilde{U}_{n,2,1} = \widetilde{V}_{n,2,1} + c_{aa}\eta_1, \quad \widetilde{U}_{n,3,1} = \widetilde{V}_{n,3,1} + c_{ab}\eta_1 + c_{bb}\eta_2$$

- Prob that option 1 is chosen:

$$P_{n,1} = Pr(\widetilde{U}_{n,2,1} < 0 \ \& \ \widetilde{U}_{n,3,1} < 0)$$

- Why need Cholesky decomposition? $(\widetilde{\varepsilon}_{n,2,1}, \widetilde{\varepsilon}_{n,3,1})$ are correlated, hard to evaluate numerically, but η_1 and η_2 are easy to draw

GHK Simulator: Form to Work with

Using Bayes rule,

$$
\begin{aligned}
P_{n,1} &= Pr(\widetilde{V}_{n,2,1} + c_{aa}\eta_1 < 0 \;\&\; \widetilde{V}_{n,3,1} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0) \\
&= Pr(\widetilde{V}_{n,2,1} + c_{aa}\eta_1 < 0)Pr(\widetilde{V}_{n,3,1} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0 | \widetilde{V}_{n,2,1} + c_{aa}\eta_1 < 0) \\
&= Pr\Big(\eta_1 < -\frac{\widetilde{V}_{n,2,1}}{c_{aa}}\Big)Pr\Big(\eta_2 < -\frac{\widetilde{V}_{n,3,1} + c_{ab}\eta_1}{c_{bb}}\Big|\eta_1 < -\frac{\widetilde{V}_{n,2,1}}{c_{aa}}\Big) \\
&= \Phi\Big(-\frac{\widetilde{V}_{n,2,1}}{c_{aa}}\Big)\int_{-\infty}^{-\frac{\widetilde{V}_{n,2,1}}{c_{aa}}} \Phi\Big(-\frac{\widetilde{V}_{n,3,1} + c_{ab}\eta_1}{c_{bb}}\Big)\overline{\phi}(\eta_1)d\eta_1
\end{aligned}
$$

where

$$
\overline{\phi}(\eta_1) = \begin{cases} \frac{\phi(\eta_1)}{\Phi(-\widetilde{V}_{n,2,1}/c_{aa})} & -\infty < \eta_1 < -\widetilde{V}_{n,2,1}/c_{aa} \\ 0 & o.w \end{cases}
$$

- Blue term is easy to compute
- Red term: **GHK punchline** (approximate integral by simulations)

GHK Simulator: Algorithm

Procedures to approximate the red term:

1. Draw η_1 from truncated normal density $\overline{\phi}$
2. Compute $\Phi\left(-\frac{\widetilde{V}_{n,3,1}+c_{ab}\eta_1}{c_{bb}}\right)$
3. Repeat 1–2 many times and take average

How to draw from a truncated univariate distribution? ▸ Illustration

GHK Algorithm:

1. Compute $\Phi\left(-\frac{\widetilde{V}_{n,2,1}}{c_{aa}}\right)$
2. Conduct procedures to approximate the red term, denote the outcome as B
3. Simulated probability $\widehat{P}_{n,1} = \Phi\left(-\frac{\widetilde{V}_{n,2,1}}{c_{aa}}\right)B$

Remarks:

- Applicable to simulate other choice probabilities, but with different $\widetilde{\Omega}_i$ and hence Cholesky factors
- Can be generalized to many–alternative problems

Detour: Importance Sampling

Suppose r.v. u has a density $f(u)$ that's hard to draw from directly, but there is another density $g(u)$ that is easy to draw from. We can obtain draws from f using the following procedure:

1. Draw from g and denote it as $u^{(1)}$
2. Weight the draw by $f(u^{(1)})/g(u^{(1)})$
3. Repeat 1–2 many times
4. The set of weighted draws is equal to a set of draws from f

CDF of weighted draws from g equals that from f:

$$\int \frac{f(u)}{g(u)} \mathbb{1}\{u < m\} g(u) du = \int_{-\infty}^{m} \frac{f(u)}{g(u)} g(u) du = F(m)$$

Requires two conditions:

- Support of g covers that of f: any possible u of f drawable from g
- $\forall u, \mathbb{E}\left(\frac{f(u)}{g(u)}\right) < \infty$

Approximate Integral via Importance Sampling

Suppose we want to calculate the following:

$$\int t(\varepsilon) f(\varepsilon) d\varepsilon,$$

where $t(\varepsilon)$ is a function of ε and f is the density that is hard to draw from directly. Suppose an alternative density g is easy to draw from. Rewrite integral to be

$$\int \left[t(\varepsilon) \frac{f(\varepsilon)}{g(\varepsilon)} \right] g(\varepsilon) d\varepsilon$$

Draw ε from g and evaluate $t\frac{f}{g}$, repeat S times and take average

Remarks:

- In practice g should have a larger tail than f
- Compute the effective sample size $\frac{1}{\sum_i^S w_i^2}$, where w_i is the normalized importance weights, which can be very low (as in particle filters)

Another Look of GHK

Recall the choice 1 probability:

$$P_{n,1} = \int \mathbb{1}(\eta \in \mathcal{B}) f(\eta) d\eta,$$

where $\mathcal{B} = \{\eta : \widetilde{U}_{n,2,1} < 0 \;\&\; \widetilde{U}_{n,3,1} < 0\}$ and $f(\eta) = \phi(\eta_1)\phi(\eta_2)$

- Accept–Reject algorithm draws from f and evaluate if $\mathbb{1}(\eta \in \mathcal{B})$ is accepted given the draws
- GHK doesn't draw from f but from g, truncated normal

$$g(\eta) = \begin{cases} \frac{\phi(\eta_1)}{\Phi(-\tilde{V}_{n,2,1})/c_{aa}} \times \frac{\phi(\eta_2)}{\Phi(-(\widetilde{V}_{n,3,1}+c_{ab}\eta_1))/c_{bb}} & \eta \in \mathcal{B} \\ 0 & \eta \notin \mathcal{B} \end{cases}$$

- GHK only draws from densities that are consistent with player choosing option 1

Another Look of GHK

Recall in the GHK algorithm, for each draw η we essentially calculate the following:

$$\widehat{P}_{n,1}(\eta) = \Phi\Big(-\frac{\widetilde{V}_{n,2,1}}{c_{aa}}\Big) \times \Phi\Big(-\frac{\widetilde{V}_{n,3,1} + c_{ab}\eta_1}{c_{bb}}\Big)$$

which is the denominator of g for $\eta \in \mathcal{B}$, so rewrite g as

$$g(\eta) = \begin{cases} \frac{f(\eta)}{\widehat{P}_{n,1}(\eta)} & \eta \in \mathcal{B} \\ 0 & \eta \notin \mathcal{B} \end{cases}$$

Importance Sampling Interpretation of GHK

Hence we have

$$
\begin{aligned}
P_{n,1} &= \int \mathbb{1}(\eta \in \mathcal{B}) f(\eta) d\eta \\
&= \int \mathbb{1}(\eta \in \mathcal{B}) \frac{f(\eta)}{g(\eta)} g(\eta) d\eta \\
&= \int \mathbb{1}(\eta \in \mathcal{B}) \frac{f(\eta)}{\frac{f(\eta)}{\widehat{P}_{n,1}(\eta)}} g(\eta) d\eta \\
&= \int \mathbb{1}(\eta \in \mathcal{B}) \widehat{P}_{n,1}(\eta) g(\eta) d\eta \\
&= \int \widehat{P}_{n,1}(\eta) g(\eta) d\eta
\end{aligned}
$$

Remarks:

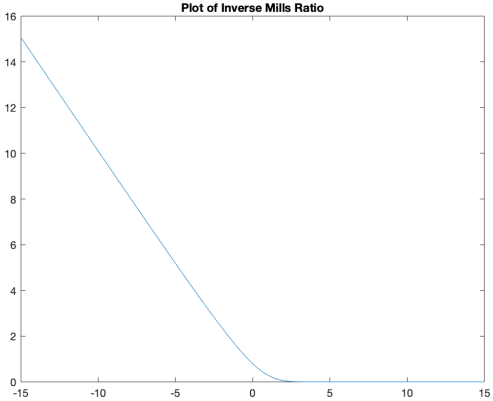
- Last equality is due to $g(\eta) > 0$ only when $\mathbb{1}(\eta \in \mathcal{B}) = 1$
- $\widehat{P}_{n,1}(\eta)$ is the weight imposed on draws from g
- GHK replaces $\mathbb{1}(\eta \in \mathcal{B})$ with smoothed $\widehat{P}_{n,1}(\eta)$

Variance of Errors in the Second Step

$$Var(\eta_i) = \sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon\nu}^2}{\sigma_\nu^2}\left[\frac{z_i'\gamma}{\sigma_\nu} \cdot \frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)} + \left(\frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}\right)^2\right]$$

◂ Back

Illustration of Inverse Mills Ratio



Back

Heckman Selection Variance Estimator

Denote $W_i = d_i[x_i', \lambda(z_i'\gamma_0)]'$ and $\widehat{W}_i = d_i[x_i', \lambda(z_i'\widehat{\gamma})]'$, then

$$G_\theta = -\mathbb{E}[d_i W_i W_i'], \quad G_\gamma = -\mu_0 \mathbb{E}[d_i \lambda_\nu(z_i'\gamma_0) W_i z_i'],$$

with sample analog

$$\widehat{G}_\theta = -\frac{1}{N}\sum_{i=1}^{N} \widehat{W}_i \widehat{W}_i', \quad \widehat{G}_\gamma = -\frac{1}{N}\widehat{\mu}\sum_{i=1}^{N} \lambda_\nu(z_i'\widehat{\gamma})\widehat{W}_i z_i'$$

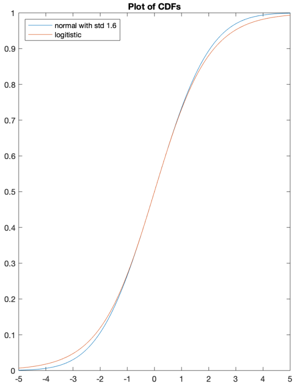
Denote $\widehat{\eta}_i = y_i - \widehat{W}_i'(\widehat{\beta}', \widehat{\mu})'$, the variance estimator is thus

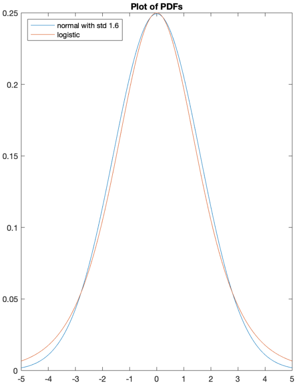
$$\widehat{V} = N\Big(\sum_{i=1}^{N} \widehat{W}_i \widehat{W}_i'\Big)^{-1} \sum_{i=1}^{N} \widehat{W}_i \widehat{W}_i \widehat{\eta}_i^2 \Big(\sum_{i=1}^{N} \widehat{W}_i \widehat{W}_i'\Big)^{-1} + \widehat{\Pi}\widehat{V}_\gamma \widehat{\Pi}'$$

- Blue: sum of White estimator for least squares
- Red: correction term for first step. $\widehat{V}_\gamma$: *variance estimator of first step probit*; $\widehat{\Pi} = \widehat{G}_\theta^{-1}\widehat{G}_\gamma$.

Back

Plots of Logit and Normal Distributions

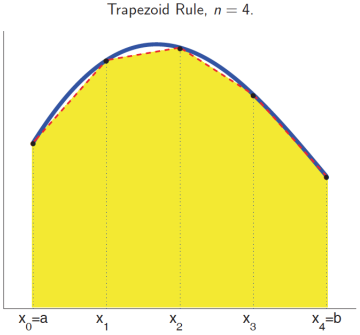




- I picked $\sigma = 1.6$ for normality to fit the two distributions.

Back

Illustration of Newton–Cotes Trapezoid

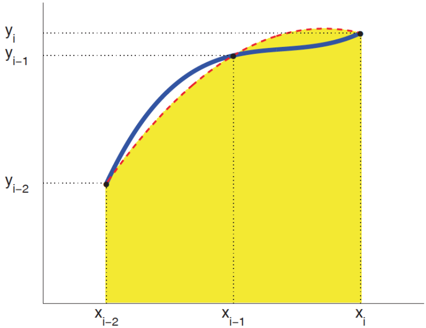


Why trapezoid? The area under the piecewise linear approximation for subinterval k is

$$\int_{x_k}^{x_{k+1}} \tilde{f}(x)dx \approx \left[\frac{f(x_{k+1}) + f(x_k)}{2}\right]h$$

Back

Illustration of Newton–Cotes Simpson



Back

Cholesky Decomposition

Consider a symmetric and positive–definite square matrix A. We can construct a lower triangular matrix L whose transpose serves as the upper triangular part:

$$L \cdot L' = A$$

Remark: Cholesky decomposition is an efficient way to check if a matrix is truly positive–definite. An alternative way is to check if the minimum eigenvalue is positive. ◂ Back

Drawing from Joint Normal Density

Consider a $\mathcal{J}$–vector joint normal distribution $\mathcal{N}(b, \Omega)$. The Cholesky decomposition is

$$L \cdot L' = \Omega$$

1. Take $\mathcal{J}$ draws from a standard normal and denote as $\nu = (\nu_1, \ldots, \nu_{\mathcal{J}})'$
2. Calculate $\varepsilon = b + L\nu$

ε is normally distributed, has mean b and covariance Ω ◂ Back

Drawing From Truncated Univariate Density

Consider a r.v ranging from a to b with density proportional to $f(\varepsilon)$ within the range. In other words:

$$k = \int_a^b f(\varepsilon) d\varepsilon = F(b) - F(a)$$

The r.v has density $\frac{1}{k}f(\varepsilon)$ for $a \leq \varepsilon \leq b$ and 0 otherwise. The following procedure draws a value from this truncated density.

1. Draw μ from standard uniform $\mathcal{U}[0, 1]$
2. Calculate weighted average $\overline{\mu} = (1 - \mu)F(a) + \mu F(b)$
3. Calculate $\varepsilon = F^{-1}(\overline{\mu})$

The draw of μ determines how far to go between a and b ◂ Back