

M-Estimation: Asymptotic Normality

Consider a method of moments estimator $\hat{\theta}$ that solves

$$0 = \frac{1}{n} \sum_{i=1}^n s(w_i, \hat{\theta}) \quad (1)$$

This may be motivated by the analogy principle based on economic model

$$E[s(w_i, \theta_0)] = 0$$

We assume that consistency of $\hat{\theta}$ is already established, and focus on asymptotic distribution. Expanding (1) around θ_0 , we obtain

$$0 = \frac{1}{n} \sum_{i=1}^n s(w_i, \theta_0) + \left(\frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \tilde{\theta})}{\partial \theta'} \right) (\hat{\theta} - \theta_0)$$

for some $\tilde{\theta}$ in between θ_0 and $\hat{\theta}$. We expect

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \tilde{\theta})}{\partial \theta'} = \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \theta_0)}{\partial \theta'} + o_p(1) \quad (2)$$

Because

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \theta_0)}{\partial \theta'} = E \left[\frac{\partial s(w_i, \theta_0)}{\partial \theta'} \right] + o_p(1)$$

by LLN, we expect

$$\sqrt{n} (\hat{\theta} - \theta_0) = - \left(E \left[\frac{\partial s(w_i, \theta_0)}{\partial \theta'} \right] \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n s(w_i, \theta_0) \right) + o_p(1)$$

Because $E[s(w_i, \theta_0)] = 0$, CLT applies there:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n s(w_i, \theta_0) \xrightarrow{d} N(0, E[s(w_i, \theta_0) s(w_i, \theta_0)'])$$

We therefore expect

Proposition 1

$$\sqrt{n} \left(\hat{\theta} - \theta_0 \right) \xrightarrow{d} N(0, \Omega)$$

where

$$\begin{aligned} \Omega &= A^{-1} B A'^{-1} \\ A &= E \left[\frac{\partial s(w_i, \theta_0)}{\partial \theta'} \right] \\ B &= E \left[s(w_i, \theta_0) s(w_i, \theta_0)' \right] \end{aligned}$$

M-Estimation: Estimation of Asymptotic Variance

Because the asymptotic variance of $\sqrt{n} \left(\hat{\theta} - \theta_0 \right)$ is equal to $A^{-1} B A'^{-1}$, we expect

$$\hat{A}^{-1} \hat{B} \hat{A}'^{-1} = A^{-1} B A'^{-1} + o_p(1)$$

where

$$\begin{aligned} \hat{A} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \hat{\theta})}{\partial \theta'} \\ \hat{B} &= \frac{1}{n} \sum_{i=1}^n s(w_i, \hat{\theta}) s(w_i, \hat{\theta})' \end{aligned}$$

Example: Nonlinear LS

If

$$E[y_i | x_i] = g(x_i, \theta_0)$$

then

$$\theta_0 = \underset{\theta}{\operatorname{argmin}} E \left[(y_i - g(x_i, \theta))^2 \right]$$

so it seems reasonable to estimate it by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - g(x_i, \theta))^2$$

Assuming that consistency is taken care of, we look at asymptotic variance. For simplicity, assume that $\dim(\theta) = 1$. Letting $u_i = y_i - g(x_i, \theta_0)$, we obtain

$$\begin{aligned} h(w_i, \theta_0) &= (y_i - g(x_i, \theta_0))^2 \\ s(w_i, \theta_0) &= -2(y_i - g(x_i, \theta_0)) \frac{\partial g(x_i, \theta_0)}{\partial \theta} = -2u_i \frac{\partial g(x_i, \theta_0)}{\partial \theta} \\ \frac{\partial s(w_i, \theta_0)}{\partial \theta} &= 2 \left(\frac{\partial g(x_i, \theta_0)}{\partial \theta} \right)^2 - 2(y_i - g(x_i, \theta_0)) \frac{\partial^2 g(x_i, \theta_0)}{\partial \theta^2} = 2 \left(\frac{\partial g(x_i, \theta_0)}{\partial \theta} \right)^2 - 2u_i \frac{\partial^2 g(x_i, \theta_0)}{\partial \theta^2} \end{aligned}$$

Because

$$E \left[\frac{\partial s(w_i, \theta_0)}{\partial \theta} \right] = 2E \left[\left(\frac{\partial g(x_i, \theta_0)}{\partial \theta} \right)^2 \right]$$

$$E [s(w_i, \theta_0)^2] = 4E \left[u_i^2 \left(\frac{\partial g(x_i, \theta_0)}{\partial \theta} \right)^2 \right]$$

the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is equal to

$$\frac{E [u_i^2 r_i^2]}{(E [r_i^2])^2}$$

for

$$r_i = \frac{\partial g(x_i, \theta_0)}{\partial \theta}$$

Example: OLS

For the linear model

$$y_i = x_i' \theta_0 + u_i \quad i = 1, 2, \dots$$

satisfying the restriction $E[x_i u_i] = 0$, we can set

$$s(w_i, \theta) = x_i \cdot (y_i - x_i' \theta)$$

and derive OLS as the corresponding M-estimator.

Asymptotic variance now follows from the general formula. Note that

$$A = E \left[\frac{\partial s(w_i, \theta_0)}{\partial \theta'} \right]$$

$$= E [-x_i x_i']$$

$$= -E [x_i x_i']$$

$$B = E [s(w_i, \theta_0) s(w_i, \theta_0)']$$

$$= E [(x_i (y_i - x_i' \theta_0)) (x_i (y_i - x_i' \theta_0))']$$

$$= E [u_i^2 x_i x_i']$$

and

$$\Omega = (-E [x_i x_i'])^{-1} E [u_i^2 x_i x_i'] ((-E [x_i x_i'])')^{-1}$$

$$= (E [x_i x_i'])^{-1} E [u_i^2 x_i x_i'] (E [x_i x_i'])^{-1}$$

Note that

$$\begin{aligned}\hat{A} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \hat{\theta})}{\partial \theta'} \\ &= -\frac{1}{n} \sum_{i=1}^n x_i x_i'\end{aligned}$$

and

$$\begin{aligned}\hat{B} &= \frac{1}{n} \sum_{i=1}^n s(w_i, \hat{\theta}) s(w_i, \hat{\theta})' \\ &= \frac{1}{n} \sum_{i=1}^n (x_i (y_i - x_i' \hat{\theta})) (x_i (y_i - x_i' \hat{\theta}))' \\ &= \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 x_i x_i'\end{aligned}$$

Example: IV

Suppose that we have a linear model

$$y_i = x_{i,1}\theta_{0,1} + x_{i,2}\theta_{0,2} + \cdots + x_{i,K}\theta_{0,K} + u_i = x_i'\theta_0 + u_i \quad i = 1, 2, \dots$$

We assume that

$$E[x_{i,k}u_i] = 0 \quad k = 1, \dots, K-1$$

We further assume that there exists a random variable $q_{i,1}$ such that (i) $E[q_{i,1}u_i] = 0$; and (ii) $E[z_i x_i']$ is nonsingular, where $z_i = (x_{i,1}, \dots, x_{i,K-1}, q_{i,1})'$. Because

$$0 = E[z_i u_i] = E[z_i (y_i - x_i' \theta_0)] = E[z_i y_i] - E[z_i x_i'] \theta_0$$

or

$$\theta_0 = (E[z_i x_i'])^{-1} E[z_i y_i]$$

we can motivate the IV estimator by the analogy principle:

$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^n z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i y_i \right) = \left(\sum_{i=1}^n z_i x_i' \right)^{-1} \left(\sum_{i=1}^n z_i y_i \right) = (Z'X)^{-1} Z'y$$

Using

$$s(w_i, \theta) = z_i (y_i - x_i' \theta)$$

we can also understand the IV estimator as an M-estimator.

Asymptotic variance now follows from the general formula. Note that

$$\begin{aligned}
A &= E \left[\frac{\partial s(w_i, \theta_0)}{\partial \theta'} \right] \\
&= E [-z_i x_i'] \\
&= -E [z_i x_i'] \\
B &= E [s(w_i, \theta_0) s(w_i, \theta_0)'] \\
&= E [(z_i (y_i - x_i' \theta_0)) (z_i (y_i - x_i' \theta_0))'] \\
&= E [u_i^2 z_i z_i']
\end{aligned}$$

and

$$\begin{aligned}
\Omega &= (-E [z_i x_i'])^{-1} E [u_i^2 z_i z_i'] ((-E [z_i x_i'])')^{-1} \\
&= (E [z_i x_i'])^{-1} E [u_i^2 z_i z_i'] (E [x_i z_i'])^{-1}
\end{aligned}$$

Note that

$$\begin{aligned}
\hat{A} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \hat{\theta})}{\partial \theta'} \\
&= -\frac{1}{n} \sum_{i=1}^n z_i x_i'
\end{aligned}$$

and

$$\begin{aligned}
\hat{B} &= \frac{1}{n} \sum_{i=1}^n s(w_i, \hat{\theta}) s(w_i, \hat{\theta})' \\
&= \frac{1}{n} \sum_{i=1}^n (z_i (y_i - x_i' \hat{\theta})) (z_i (y_i - x_i' \hat{\theta}))' \\
&= \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 z_i z_i'
\end{aligned}$$

Conditional MLE

Suppose that y_i given x_i is known to have the conditional PDF $f(\cdot | x_i, \theta_0)$ for some $\theta_0 \in \Theta$. The (conditional) maximum likelihood estimator solves

$$\max_{\theta \in \Theta} \sum_{i=1}^n \log f(y_i | x_i, \theta)$$

This is motivated by the following theorem:

Theorem 1

$$E[\log f(y_i | x_i, \theta_0) | x_i] \geq E[\log f(y_i | x_i, \theta) | x_i] \quad \theta \in \Theta$$

Here, the conditional expectation is taken with respect to the true conditional PDF of y_i , i.e., $f(\cdot | x_i, \theta_0)$.

Proof.

$$\begin{aligned}
E[\log f(y_i|x_i, \theta)|x_i] - E[\log f(y_i|x_i, \theta_0)|x_i] &= E\left[\log \frac{f(y_i|x_i, \theta)}{f(y_i|x_i, \theta_0)} \middle| x_i\right] \\
&\leq \log E\left[\frac{f(y_i|x_i, \theta)}{f(y_i|x_i, \theta_0)} \middle| x_i\right] && \text{(Jensen)} \\
&= \log\left(\int \frac{f(y|x_i, \theta)}{f(y|x_i, \theta_0)} f(y|x_i, \theta_0) dy\right) \\
&= \log\left(\int f(y|x_i, \theta) dy\right) \\
&= \log(1) \\
&= 0
\end{aligned}$$

■

The (conditional) maximum likelihood estimator solves

$$\max_{\theta \in \Theta} \sum_{i=1}^n \log f(y_i|x_i, \theta)$$

or

$$0 = \sum_{i=1}^n \frac{\partial \log f(y_i|x_i, \hat{\theta})}{\partial \theta}$$

Because MLE is a special case of M-estimator, it inherits all the asymptotic properties of M-estimation. Asymptotic variance of (conditional) MLE can be simplified because of information equality, though.

Lemma 1

$$E\left[\frac{\partial^2 \log f(y_i|x_i, \theta)}{\partial \theta^2}\right] = -E\left[\frac{\partial \log f(y_i|x_i, \theta)}{\partial \theta} \frac{\partial \log f(y_i|x_i, \theta)}{\partial \theta'}\right]$$

Because

$$\begin{aligned}
s(w_i; \theta) &= \frac{\partial \log f(y_i|x_i, \theta)}{\partial \theta} \\
\frac{\partial s(w_i, \theta)}{\partial \theta'} &= \frac{\partial^2 \log f(y_i|x_i, \theta)}{\partial \theta^2}
\end{aligned}$$

we have

$$\begin{aligned}
A &= E\left[\frac{\partial^2 \log f(y_i|x_i, \theta)}{\partial \theta^2}\right] \\
&= -E\left[\frac{\partial \log f(y_i|x_i, \theta)}{\partial \theta} \frac{\partial \log f(y_i|x_i, \theta)}{\partial \theta'}\right] \\
&= -E[s(w_i, \theta_0) s(w_i, \theta_0)'] \\
&= -B
\end{aligned}$$

Writing

$$\mathcal{I} = E [s(w_i, \theta_0) s(w_i, \theta_0)'] = -E \left[\frac{\partial^2 \log f(y_i | x_i, \theta)}{\partial \theta^2} \right]$$

we have

$$\sqrt{n} (\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1})$$

Note that we can estimate the asymptotic variance \mathcal{I} by

$$\mathcal{I}_1 = \frac{1}{n} \sum_{i=1}^n \frac{\partial \log f(y_i | x_i, \hat{\theta}_{MLE})}{\partial \theta} \frac{\partial \log f(y_i | x_i, \hat{\theta}_{MLE})}{\partial \theta'}$$

or

$$\mathcal{I}_2 = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(y_i | x_i, \hat{\theta}_{MLE})}{\partial \theta \partial \theta'}$$

Two Step Estimation: 2SLS

Consider a very simple model

$$\begin{aligned} y_i &= \beta x_i + \varepsilon_i \\ x_i &= z_i' \pi + v_i \end{aligned}$$

where $\dim(z_i) = K$, and

$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \Big| z_i \sim N(0, \Sigma) = N\left(0, \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 \end{bmatrix}\right)$$

Note that the first stage OLS solves

$$\frac{1}{n} \sum_{i=1}^n \psi_1(w_i, \hat{\pi}) = \frac{1}{n} \sum_{i=1}^n z_i (x_i - z_i' \hat{\pi}) = 0$$

Also note that the second stage regression solves

$$\frac{1}{n} \sum_{i=1}^n \psi_2(w_i, \hat{\pi}, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^n (z_i' \hat{\pi}) (y_i - \hat{\beta} (z_i' \hat{\pi})) = 0$$

These estimators can be understood as a component of one single M-estimator for $(\beta, \pi)'$: Let

$$\psi(w_i, \pi, \beta) = \begin{pmatrix} \psi_1(w_i, \pi) \\ \psi_2(w_i, \pi, \beta) \end{pmatrix}$$

and note that the solution to

$$\frac{1}{n} \sum_{i=1}^n \psi(w_i, \pi, \beta) = 0$$

is given by $(\hat{\pi}, \hat{\beta})$. We can therefore use Proposition 1 to obtain the asymptotic variance of $(\hat{\pi}, \hat{\beta})$.

Binary Response Model

Two Choices:

$$\begin{aligned} y_i = 1 & \text{ Choice 1 is made} \\ y_i = 0 & \text{ Choice 0 is made} \end{aligned}$$

Assume that

$$y_i = 1 \Leftrightarrow U_i = x_i' \beta - \varepsilon_i \geq 0$$

Let

$$G(t) \equiv \Pr[\varepsilon_i \leq t].$$

Then,

$$\Pr[y_i = 1] = \Pr[\varepsilon_i \leq x_i' \beta] = G(x_i' \beta).$$

Example 1 $G(t) = \Phi(t)$: *Probit Model*

Example 2 $G(t) = \frac{e^t}{e^t + 1} = \Lambda(t)$: *Logit Model*

Note that individual likelihood equals

$$G(x_i' \beta)^{y_i} [1 - G(x_i' \beta)]^{1-y_i}$$

It follows that the joint log likelihood equals

$$\sum_i y_i \cdot \log G(x_i' \beta) + (1 - y_i) \cdot \log [1 - G(x_i' \beta)]$$

MLE from FOC:

$$\sum_i \frac{y_i - G(x_i' \hat{b})}{G(x_i' \hat{b}) [1 - G(x_i' \hat{b})]} \cdot g(x_i' \hat{b}) \cdot x_i = 0.$$

Proposition 2 *The log likelihood of the Probit or Logit model is globally concave.*

Binary Response Model: Asymptotic Distribution of MLE

Proposition 3

$$\sqrt{n} (\hat{b} - \beta) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\beta)).$$

Proposition 4 *The Fisher Information $I(\beta)$ from the individual observation equals*

$$E \left[\frac{g(x_i' \beta)^2}{G(x_i' \beta) [1 - G(x_i' \beta)]} \cdot x_i x_i' \right].$$

Proof. Obvious from

$$\frac{\partial \log f(z_i, \beta)}{\partial \beta} = \frac{y_i - G(x_i' \beta)}{G(x_i' \beta) [1 - G(x_i' \beta)]} \cdot g(x_i' \beta) \cdot x_i$$

and

$$I(\beta) = E \left[\frac{\partial \log f}{\partial \beta} \frac{\partial \log f}{\partial \beta'} \right].$$

■

Proposition 5

$$\begin{aligned} \text{plim} \frac{1}{n} \sum_i \frac{g(x_i' \hat{\beta})^2}{G(x_i' \hat{\beta}) [1 - G(x_i' \hat{\beta})]} \cdot x_i x_i' &= I(\beta), \\ \text{plim} \frac{1}{n} \sum_i \frac{[y_i - G(x_i' \hat{b})]^2}{G(x_i' \hat{b})^2 [1 - G(x_i' \hat{b})]^2} \cdot g(x_i' \hat{b})^2 \cdot x_i x_i' &= I(\beta). \end{aligned}$$

Binary Response Model: Average of Marginal Effects

Recall that

$$\Pr[y_i = 1 | x_i] = G(x_i' \beta).$$

Therefore, we have

$$\frac{\partial \Pr[y_i = 1 | x_i]}{\partial x_i} = g(x_i' \beta) \beta$$

It follows that

$$\phi \equiv E \left[\frac{\partial \Pr[y_i = 1 | x_i]}{\partial x_i} \right] = E[g(x_i' \beta) \beta]$$

How can we estimate ϕ ? For this purpose, it is useful to write

$$\begin{aligned} \psi_1(w_i, \beta) &\equiv \frac{y_i - G(x_i' \beta)}{G(x_i' \beta) [1 - G(x_i' \beta)]} \cdot g(x_i' \beta) \cdot x_i \\ \psi_2(w_i, \phi, \beta) &\equiv \phi - g(x_i' \beta) \beta \end{aligned}$$

Note that we obtain the MLE by solving

$$0 = \frac{1}{n} \sum_{i=1}^n \psi_1(w_i, \hat{\beta})$$

and we obtain the estimate of ϕ by solving

$$\begin{aligned} 0 &= \psi_2(w_i, \hat{\phi}, \hat{\beta}) \\ &= \hat{\phi} - \frac{1}{n} \sum_{i=1}^n g(x_i' \hat{\beta}) \hat{\beta} \end{aligned}$$

These estimators can be understood as a component of one single M-estimator for $(\phi, \beta)'$: Let

$$\psi(w_i, \phi, \beta) = \begin{pmatrix} \psi_1(w_i, \beta) \\ \psi_2(w_i, \phi, \beta) \end{pmatrix}$$

Using the usual asymptotic variance formula, we can then obtain the asymptotic variance for ϕ as well.

Censoring

Model

$$y_i^* = x_i' \beta + u_i, \quad u_i | x_i \quad i.i.d. \quad N(0, \sigma^2)$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Individual Likelihood:

$$\Phi\left(-\frac{x_i' \beta}{\sigma}\right)^{1-d_i} \left\{ \frac{1}{\sigma} \phi\left(\frac{y_i - x_i' \beta}{\sigma}\right) \right\}^{d_i}$$

Censoring: Least Squares Bias

Theorem 2

$$E[y_i | x_i, d_i = 1] = x_i' \beta + \sigma \lambda\left(\frac{x_i' \beta}{\sigma}\right)$$

where

$$\lambda(t) \equiv \frac{\phi(t)}{\Phi(t)}$$

Proof. We have

$$\begin{aligned} E[y_i | x_i, d_i = 1] &= \beta' x_i + E[u_i | x_i, u_i \geq -x_i' \beta] \\ &= \beta' x_i + \sigma E\left[\frac{u_i}{\sigma} \mid x_i, \frac{u_i}{\sigma} \geq -\frac{x_i' \beta}{\sigma}\right] \end{aligned}$$

By the Lemma below, we have

$$\begin{aligned} E \left[\frac{u_i}{\sigma} \mid x_i, \frac{u_i}{\sigma} \geq -\frac{\beta' x_i}{\sigma} \right] &= \frac{\phi \left(-\frac{x_i' \beta}{\sigma} \right)}{1 - \Phi \left(-\frac{x_i' \beta}{\sigma} \right)} \\ &= \frac{\phi \left(\frac{x_i' \beta}{\sigma} \right)}{\Phi \left(\frac{x_i' \beta}{\sigma} \right)} = \lambda \left(\frac{x_i' \beta}{\sigma} \right) \end{aligned}$$

■

Lemma 2 *If $e \sim N(0, 1)$, then*

$$E[e | e \geq t] = \frac{\phi(t)}{1 - \Phi(t)}.$$

Proof.

$$\begin{aligned} \Pr(e \leq s | e \geq t) &= \frac{\int_t^s \phi(u) du}{\int_t^\infty \phi(u) du} = \frac{\int_t^s \phi(u) du}{1 - \Phi(t)} \\ &\Rightarrow f(s | e \geq t) \equiv \frac{\partial \Pr(e \leq s | e \geq t)}{\partial s} = \frac{\phi(s)}{1 - \Phi(t)}. \end{aligned}$$

Thus,

$$E[e | e \geq t] = \int_t^\infty s f(s | e \geq t) ds = \frac{\int_t^\infty s \phi(s) ds}{1 - \Phi(t)}.$$

But because

$$\frac{d\phi(s)}{ds} = \frac{1}{\sqrt{2\pi}} \frac{d \exp(-s^2/2)}{ds} = -s \frac{1}{\sqrt{2\pi}} \exp(-s^2/2) = -s\phi(s),$$

we have

$$\int_t^\infty s \phi(s) ds = -\phi(s) \Big|_t^\infty = -\phi(\infty) + \phi(t) = \phi(t).$$

Thus,

$$E[e | e \geq t] = \frac{\phi(t)}{1 - \Phi(t)}.$$

■

Censoring: Some Details

Note that We have

$$\begin{aligned} b_{OLS} &= \left(\sum_{d_i=1} x_i x_i' \right)^{-1} \left(\sum_{d_i=1} x_i y_i \right) = \left(\sum_{i=1}^n d_i x_i x_i' \right)^{-1} \left(\sum_{i=1}^n d_i x_i y_i \right) \\ &= (E [d_i x_i x_i'])^{-1} (E [d_i x_i y_i]) + o_p(1) \end{aligned}$$

Write

$$E [d_i x_i x_i'] = E [E [d_i | x_i] x_i x_i'] = E [\pi (x_i) x_i x_i']$$

where

$$\pi (x_i) \equiv E [d_i | x_i] = \Pr [d_i = 1 | x_i]$$

Likewise, we have

$$\begin{aligned} E [d_i x_i y_i] &= E [x_i E [d_i y_i | x_i]] \\ &= E [\pi (x_i) x_i E [y_i | x_i, d_i = 1]] \\ &= E \left[\pi (x_i) x_i \left(x_i' \beta + \sigma \lambda \left(\frac{x_i' \beta}{\sigma} \right) \right) \right] \end{aligned}$$

It follows that

$$(E [d_i x_i x_i'])^{-1} (E [d_i x_i y_i]) = \beta + E [\pi (x_i) x_i x_i'] E \left[\sigma \pi (x_i) \lambda \left(\frac{x_i' \beta}{\sigma} \right) x_i \right] \neq 0$$

Sample Selection Bias

$$\begin{aligned} y_{i1} &= x_{i1}' \beta + u_i, \\ y_{i2}^* &= x_{i2}' \gamma + v_i, \\ d_i &= 1 \quad \text{if } y_{i2}^* \geq 0, \\ &= 0 \quad \text{if } y_{i2}^* < 0. \end{aligned}$$

y_{i1} is observed only if $d_i = 1$; we always observe d_i , x_{i1} , and x_{i2} . We assume that

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \Big| x_i \quad i.i.d. \sim N(0, \Sigma).$$

Least Squares Bias: Write

$$u_i = \frac{\rho \sigma_u \sigma_v}{\sigma_v^2} v_i + w_i$$

and note that w_i is independent of v_i . We then obtain

$$\begin{aligned}
E[u_i | v_i \geq t] &= E\left[\frac{\rho\sigma_u\sigma_v}{\sigma_v^2}v_i + w_i \mid v_i \geq t\right] \\
&= \frac{\rho\sigma_u}{\sigma_v}E[v_i | v_i \geq t] + E[w_i | v_i \geq t] \\
&= \frac{\rho\sigma_u}{\sigma_v}\sigma_v E\left[\frac{v_i}{\sigma_v} \mid \frac{v_i}{\sigma_v} \geq \frac{t}{\sigma_v}\right] + E[w_i] \\
&= \rho\sigma_u \frac{\phi\left(\frac{t}{\sigma_v}\right)}{1 - \Phi\left(\frac{t}{\sigma_v}\right)}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
E[y_i | x_i, d_i = 1] &= \beta'x_{i1} + E[u_i | x_i, v_i \geq -\gamma'x_{i2}] \\
&= \beta'x_{i1} + \rho\sigma_u \frac{\phi\left(-x'_{i2}\frac{\gamma}{\sigma_v}\right)}{1 - \Phi\left(-x'_{i2}\frac{\gamma}{\sigma_v}\right)} \\
&= \beta'x_{i1} + \rho\sigma_u\lambda\left(x'_{i2}\frac{\gamma}{\sigma_v}\right)
\end{aligned}$$

Sample Selection: Two Step Estimator

Note that

$$\Pr[d_i = 1 | x_i] = \Pr\left[\frac{v_i}{\sigma_v} \geq -x'_{i2}\frac{\gamma}{\sigma_v}\right] = \Phi(x'_{i2}\alpha) = \Phi\left(x'_{i2}\frac{\gamma}{\sigma_v}\right)$$

for $\alpha \equiv \frac{\gamma}{\sigma_v}$.

- Step 1 - Estimate α from MLE Probit.
- Step 2 - Estimate β and $\xi \equiv \rho\sigma_u$ from the OLS of y_{i1} on x_{i1} and $\lambda(x'_{i2}\hat{\alpha})$ for the $d_i = 1$ subsample.

This two step estimator fits within the method of moments framework. Let $\theta = (\beta, \alpha, \xi)$, and

$$\psi(z_i, \theta) \equiv \begin{bmatrix} \psi_1(z_i, \theta) \\ \psi_2(z_i, \theta) \end{bmatrix}$$

where

$$\begin{aligned}
\psi_1(z_i, \theta) &\equiv \frac{d_i - \Phi(x'_{i2}\alpha)}{\Phi(x'_{i2}\alpha)(1 - \Phi(x'_{i2}\alpha))} \phi(x'_{i2}\alpha) x_{i2}, \\
\psi_2(z_i, \theta) &\equiv d_i \cdot (y_{i1} - \beta'x_{i1} - \xi\lambda(x'_{i2}\alpha)) \begin{bmatrix} x_{i1} \\ \lambda(x'_{i2}\alpha) \end{bmatrix}.
\end{aligned}$$

Then,

$$\frac{1}{n} \sum_i \psi(z_i, \theta) = 0$$

is solved by choosing α as the ML probit estimator, and β and ξ and the second step estimator. Thus, the asymptotic variance can be obtained from the method of moments formula.

Justification of (2)

Here's why we expect the first equality to be valid. Assume for simplicity that $\dim(\theta) = \dim(s) = 1$. We then have

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \tilde{\theta})}{\partial \theta} - \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \theta_0)}{\partial \theta} \right| &= \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \tilde{\theta})}{\partial \theta} - \frac{\partial s(w_i, \theta_0)}{\partial \theta} \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \left| \frac{\partial s(w_i, \tilde{\theta})}{\partial \theta} - \frac{\partial s(w_i, \theta_0)}{\partial \theta} \right| \\ &= \frac{1}{n} \sum_{i=1}^n \left| \frac{\partial^2 s(w_i, \tilde{\theta})}{\partial \theta^2} (\tilde{\theta} - \theta_0) \right| \end{aligned}$$

for some $\tilde{\theta}$ in between θ_0 and $\tilde{\theta}$. We therefore have

$$\left| \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \tilde{\theta})}{\partial \theta} - \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \theta_0)}{\partial \theta} \right| \leq \left(\frac{1}{n} \sum_{i=1}^n \left| \frac{\partial^2 s(w_i, \tilde{\theta})}{\partial \theta^2} \right| \right) \cdot |\tilde{\theta} - \theta_0|$$

If there exists a function $B(w_i)$ such that

$$\left| \frac{\partial^2 s(w_i, \theta)}{\partial \theta^2} \right| \leq B(w_i)$$

and

$$E[B(w_i)] < \infty,$$

we further have

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \tilde{\theta})}{\partial \theta} - \frac{1}{n} \sum_{i=1}^n \frac{\partial s(w_i, \theta_0)}{\partial \theta} \right| &\leq \left(\frac{1}{n} \sum_{i=1}^n B(w_i) \right) \cdot |\tilde{\theta} - \theta_0| \\ &= O_p(1) \cdot o_p(1) \end{aligned}$$