Lecture 11: Binary choice and maximum likelihood

Economics 527: Econometric Methods

Vadim Marmer, UBC

Binary choice

Binary choice model

- Data: observations (Y_i, X_i), $i = 1, \ldots, n$.
- Binary Y_i indicates the choice:

$$Y_i \in \{0, 1\}$$

- Examples:
  - in the labour force or not
  - went for a postgraduate degree or not
  - firm entered market i or not
- Random (latent) utility from choice 1:

$$V_i = X_i^\top \beta + \varepsilon_i$$

- Observed and unobserved:

$$\begin{aligned} V_i &: \text{ unobserved} \\ X_i &: \text{ observed characteristics} \\ \varepsilon_i &: \text{ unobserved} \end{aligned}$$

- Choice 1 when the utility is positive:

$$\begin{aligned} Y_i &= \mathbf{1}(V_i > 0) \\ &= \mathbf{1}(X_i^\top \beta + \varepsilon_i > 0) \end{aligned}$$

- Y_i generated from a nonlinear model.

Response probability

- **Assume:** F, the CDF of ε_i given X_i, known:

$$F(t) = \mathrm{P}(\varepsilon_i \leq t \mid X_i = x) \quad \text{for all } x, t$$

- Substitute the model:

$$\begin{aligned} \mathrm{P}(Y_i = 1 \mid X_i = x) &= \mathrm{P}(x^\top \beta + \varepsilon_i > 0 \mid X_i = x) \\ &= \mathrm{P}(\varepsilon_i > -x^\top \beta \mid X_i = x) \\ &= 1 - F(-x^\top \beta) \\ \implies \quad \mathrm{P}(Y_i = 1 \mid X_i) &= 1 - F(-X_i^\top \beta) \end{aligned}$$

- **Assume:** F symmetric around zero:

$$1 - F(-t) = F(t)$$

- Response probability with a symmetric F:

$$\boxed{\mathrm{P}(Y_i = 1 \mid X_i) = F(X_i^\top \beta)}$$

Linear probability model

Uniform errors

- Suppose:

$$\begin{aligned} & \varepsilon_i \mid X_i \sim \text{Uniform}(0,1) \\ \implies \quad & F(t) = t, \quad t \in (0,1) \end{aligned}$$

- Uniform CDF at $-X_i^\top \beta$:

$$\begin{aligned} \mathrm{P}(Y_i = 1 \mid X_i) &= 1 - F(-X_i^\top \beta) \\ &= 1 + X_i^\top \beta \\ &= \underbrace{(\beta_1 + 1)}_{\text{intercept}} + X_{i,2}^\top \beta_2 \end{aligned}$$

- **Assume:**

$$0 < -X_i^\top \beta < 1$$

- X_i includes an intercept:

$$X_i = \begin{pmatrix} 1 \\ X_{i,2} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

- Constant 1: absorbed by the intercept.
- **Definition.** Linear probability model (LPM):

$$\boxed{\mathrm{P}(Y_i = 1 \mid X_i) = X_i^\top \beta}$$

Connection of LPM to regression

- Conditional mean of binary Y_i:

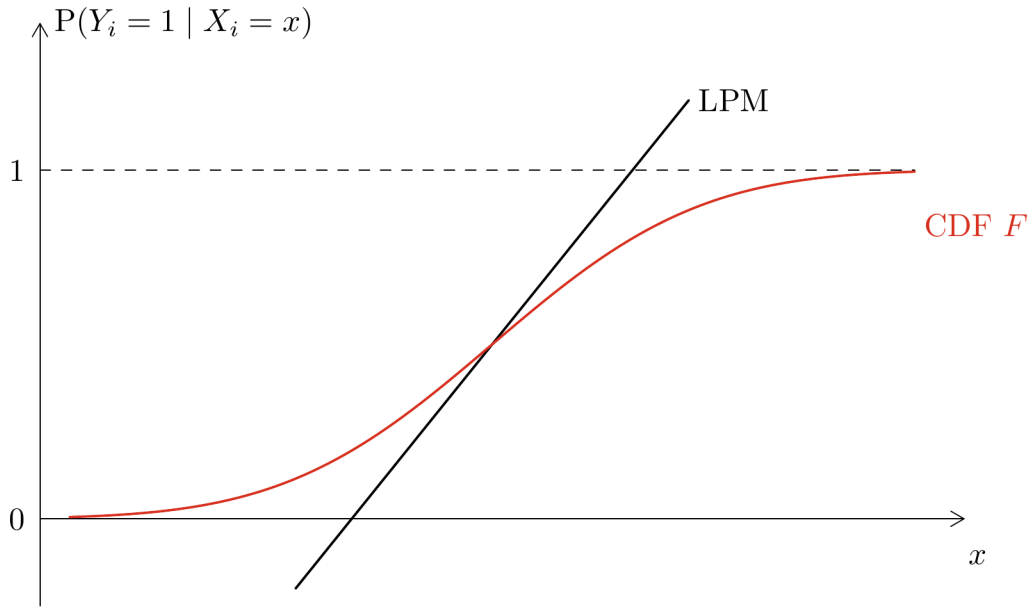
$$\begin{aligned} \mathrm{E}\left[Y_i \mid X_i\right] &= 1 \cdot \mathrm{P}(Y_i = 1 \mid X_i) + 0 \cdot \mathrm{P}(Y_i = 0 \mid X_i) \\ &= \mathrm{P}(Y_i = 1 \mid X_i) \end{aligned}$$

- LPM $\implies$ linear regression of Y_i on X_i:

$$\mathrm{E}\left[Y_i \mid X_i\right] = \mathrm{P}(Y_i = 1 \mid X_i) = X_i^\top \beta$$

LPM against an S-shaped CDF

- Scalar x: LPM, then CDF F:



Probit and logit

Probit

- $\varepsilon_i \mid X_i \sim N(0,1)$, CDF Φ:

$$\mathrm{P}(Y_i = 1 \mid X_i) = \boxed{\Phi(X_i^\top \beta)} \qquad (\Phi \text{ symmetric around zero})$$

Logit

- $\varepsilon_i \mid X_i$ logistic, CDF Λ:

$$\begin{aligned}\Lambda(t) &= \frac{e^t}{1+e^t} \\ &= \frac{1}{e^{-t}+1} \\ &= \frac{1}{1+\frac{1}{e^t}}\end{aligned}$$

- Probability of each choice:

$$\begin{aligned}\mathrm{P}(Y_i = 1 \mid X_i) &= \frac{e^{X_i^\top \beta}}{1+e^{X_i^\top \beta}} \\ \mathrm{P}(Y_i = 0 \mid X_i) &= 1 - \frac{e^{X_i^\top \beta}}{1+e^{X_i^\top \beta}} \\ &= \frac{1}{1+e^{X_i^\top \beta}}\end{aligned}$$

- Odds ratio:

$$\frac{\mathrm{P}(Y_i = 1 \mid X_i)}{\mathrm{P}(Y_i = 0 \mid X_i)} = e^{X_i^\top \beta}$$

- Log odds ratio:

$$\boxed{\log \frac{\mathrm{P}(Y_i = 1 \mid X_i)}{\mathrm{P}(Y_i = 0 \mid X_i)} = X_i^\top \beta}$$

Identification

Scale of β and ε_i

- Suppose $\varepsilon_i \mid X_i \sim N(0, \sigma^2)$.
- Rescaled coefficients and errors:

$$\beta^* = \frac{\beta}{\sigma}, \qquad \varepsilon_i^* = \frac{\varepsilon_i}{\sigma}, \qquad \varepsilon_i^* \mid X_i \sim N(0,1)$$

- Divide by $\sigma > 0$ inside the indicator:

$$\begin{aligned} Y_i &= \mathbf{1}(X_i^\top \beta + \varepsilon_i > 0) \\ &= \mathbf{1}\Big(X_i^\top \frac{\beta}{\sigma} + \frac{\varepsilon_i}{\sigma} > 0\Big) \\ &= \mathbf{1}(X_i^\top \beta^* + \varepsilon_i^* > 0) \end{aligned}$$

- Example with scalar $X_{i,2}$, same distribution of Y_i given X_i:

$$\begin{aligned} &1) \quad Y_i = \mathbf{1}(1 + X_{i,2} + \varepsilon_i > 0), \qquad \varepsilon_i \mid X_i \sim N(0,4) \\ &2) \quad Y_i = \mathbf{1}(\tfrac{1}{2} + \tfrac{1}{2} X_{i,2} + \varepsilon_i > 0), \qquad \varepsilon_i \mid X_i \sim N(0,1) \end{aligned}$$

- Only β/σ can be identified.

Location of ε_i and the intercept

- Suppose $\varepsilon_i \mid X_i \sim N(\mu, \sigma^2)$:

$$\begin{aligned} \mathrm{P}(Y_i = 1 \mid X_i) &= \mathrm{P}(X_i^\top \beta + \varepsilon_i > 0 \mid X_i) \\ &= \mathrm{P}(\varepsilon_i > -X_i^\top \beta \mid X_i) \\ &= \mathrm{P}\Big(\frac{\varepsilon_i - \mu}{\sigma} > \frac{-X_i^\top \beta - \mu}{\sigma} \;\Big|\; X_i\Big) \\ &= \Phi\Big(\frac{X_i^\top \beta + \mu}{\sigma}\Big) \qquad \Big(\frac{\varepsilon_i - \mu}{\sigma} \;\Big|\; X_i \sim N(0,1)\Big) \end{aligned}$$

- Intercept, $X_i = (1, X_{i,2}^\top)^\top$, $\beta = (\beta_1, \beta_2^\top)^\top$:

$$\frac{X_i^\top \beta + \mu}{\sigma} = \underbrace{\frac{\beta_1 + \mu}{\sigma}}_{\text{intercept}} + X_{i,2}^\top \frac{\beta_2}{\sigma}$$

- μ: absorbed by the intercept.

Identification of probit

- $\mathrm{P}(Y_i = 1 \mid X_i = x)$: can be estimated from the data.
- Invert Φ:

$$\begin{aligned} &\mathrm{P}(Y_i = 1 \mid X_i = x) = \Phi(x^\top \beta) \\ \iff \quad &\Phi^{-1}(\mathrm{P}(Y_i = 1 \mid X_i = x)) = x^\top \beta, \qquad \beta \in \mathbb{R}^k \end{aligned}$$

- Quantile of $N(0,1)$, with $\Phi^{-1}(\tau) = z_\tau$:

$$z(x) = \Phi^{-1}\big(\mathrm{P}(Y_i = 1 \mid X_i = x)\big) = z_{\mathrm{P}(Y_i=1|X_i=x)}$$

- Pick values $x_1, \ldots, x_k$ of X_i: $k \times 1$, linearly independent:

$$\underbrace{\begin{pmatrix} z(x_1) \\ \vdots \\ z(x_k) \end{pmatrix}}_{k\times 1} = \underbrace{\begin{pmatrix} {}^{1\times k} \\ x_1^\top \\ \vdots \\ x_k^\top \end{pmatrix}}_{k\times k} \beta$$

$$\implies \quad \beta = \begin{pmatrix} x_1^\top \\ \vdots \\ x_k^\top \end{pmatrix}^{-1} \begin{pmatrix} z(x_1) \\ \vdots \\ z(x_k) \end{pmatrix}$$

- β: identified.

Identification of logit

- Log odds ratio at $X_i = x$, identified from the data:

$$r(x) = \log \frac{\mathrm{P}(Y_i = 1 \mid X_i = x)}{\mathrm{P}(Y_i = 0 \mid X_i = x)} = x^\top \beta$$

- Same steps, $r(x)$ in place of $z(x)$:

$$\beta = \begin{pmatrix} x_1^\top \\ \vdots \\ x_k^\top \end{pmatrix}^{-1} \begin{pmatrix} r(x_1) \\ \vdots \\ r(x_k) \end{pmatrix}$$

Example: Walmart and town population

- $\mathrm{Walmart}_i = 1$: a Walmart in town i.
- Logit given the population of town i:

$$\mathrm{P}(\mathrm{Walmart}_i = 1 \mid \mathrm{Population}_i = x) = \frac{e^{\beta_1+\beta_2 x}}{1 + e^{\beta_1+\beta_2 x}}$$

- Log odds ratio, identified from the data:

$$r(x) = \log \frac{\mathrm{P}(\mathrm{Walmart}_i = 1 \mid \mathrm{Population}_i = x)}{\mathrm{P}(\mathrm{Walmart}_i = 0 \mid \mathrm{Population}_i = x)} = \beta_1 + \beta_2 x$$

- Two town populations, $x = 10{,}000$ and $x = 25{,}000$:

$$\begin{aligned} & r(10{,}000) = \beta_1 + \beta_2 \cdot 10{,}000 \\ & r(25{,}000) = \beta_1 + \beta_2 \cdot 25{,}000 \\ \iff \quad & \begin{bmatrix} r(10{,}000) \\ r(25{,}000) \end{bmatrix} = \begin{bmatrix} 1 & 10{,}000 \\ 1 & 25{,}000 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \\ \implies \quad & \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 10{,}000 \\ 1 & 25{,}000 \end{bmatrix}^{-1} \begin{bmatrix} r(10{,}000) \\ r(25{,}000) \end{bmatrix} \end{aligned}$$

Marginal effects

Marginal effects in the linear model, probit and logit

- LPM:

$$\begin{aligned} & \mathrm{E}\left[Y_i \mid X_i = x\right] = \mathrm{P}(Y_i = 1 \mid X_i = x) = x^\top \beta \\ \implies \quad & \beta_j = \frac{\partial \, \mathrm{P}(Y_i = 1 \mid X_i = x)}{\partial x_j} \end{aligned}$$

- β_j: the same at every x.
- Probit and logit, f: PDF of F:

$$\begin{aligned} & \mathrm{P}(Y_i = 1 \mid X_i = x) = F(x^\top \beta) \\ \implies \quad & \frac{\partial \, \mathrm{P}(Y_i = 1 \mid X_i = x)}{\partial x_j} = f(x^\top \beta) \, \beta_j \end{aligned}$$

- $f(x^\top \beta) \, \beta_j$: changes with x.

Average marginal effect (AME)

- **Definition.** Marginal effect averaged over X_i:

$$\begin{aligned} \mathrm{AME}_j &= \mathrm{E}\left[f(X_i^\top \beta) \, \beta_j\right] \\ &= \mathrm{E}\left[f(X_i^\top \beta)\right] \, \beta_j \end{aligned}$$

Maximum likelihood estimator (MLE)

Bernoulli trials

- Consider a version with only Y_i: no varying regressors X_i.
- Data $Y_1, \ldots, Y_n \in \{0, 1\}$, iid:

$$\begin{aligned} & \mathrm{P}(Y_i = 1) = \theta_0 \in (0, 1) \\ & \mathrm{P}(Y_i = 0) = 1 - \theta_0 \end{aligned}$$

- θ_0: true value of θ.
- PMF, $y \in \{0, 1\}$:

$$\mathrm{P}(Y_i = y) = \theta_0^y (1 - \theta_0)^{1-y}$$

- A possible sample, one of 2^n:

$$y_1, \ldots, y_n, \qquad y_i \in \{0, 1\}$$

- Probability of the observed sample:

$$\begin{aligned} \mathrm{P}(Y_1 = y_1, \ldots, Y_n = y_n) &= \prod_{i=1}^n \mathrm{P}(Y_i = y_i) \qquad \text{(independence)} \\ &= \prod_{i=1}^n \theta_0^{y_i} (1 - \theta_0)^{1-y_i} \\ &= \theta_0^{\sum_i y_i} (1 - \theta_0)^{n - \sum_i y_i} \end{aligned}$$

Likelihood function and the MLE

- Observed sample probability as a likelihood function of θ:

$$L_n(\theta) = \theta^{\sum_i Y_i}(1-\theta)^{n-\sum_i Y_i}$$

- **Definition.** MLE:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \; L_n(\theta) \\ &= \arg\max_{\theta} \; \frac{1}{n}\log L_n(\theta)\end{aligned}$$

- Average log likelihood, with $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$:

$$\begin{aligned}\frac{1}{n}\log L_n(\theta) &= \frac{1}{n}\sum_{i=1}^n Y_i \log\theta + \frac{1}{n}\Big(n - \sum_{i=1}^n Y_i\Big)\log(1-\theta) \\ &= \bar{Y}\log\theta + (1-\bar{Y})\log(1-\theta)\end{aligned}$$

MLE in Bernoulli trials

- First-order condition:

$$\frac{\partial}{\partial\theta}\Big(\frac{1}{n}\log L_n(\theta)\Big) = \frac{\bar{Y}}{\theta} - \frac{1-\bar{Y}}{1-\theta}$$

- $\hat{\theta}$ solves:

$$\begin{aligned}&\frac{\bar{Y}}{\hat{\theta}} - \frac{1-\bar{Y}}{1-\hat{\theta}} = 0 \\ \implies \quad &\boxed{\hat{\theta} = \bar{Y}}\end{aligned}$$

- Weak LLN for $\bar{Y}$:

$$\begin{aligned}\hat{\theta} = \bar{Y} &\to_p \mathrm{E}\,[Y_i] \\ &= 1\cdot\theta_0 + 0\cdot(1-\theta_0) \\ &= \theta_0\end{aligned}$$

Normal regression model

- Normal regression with iid observations (Y_i, X_i):

$$\begin{aligned}Y_i &= X_i^\top\beta_0 + U_i \\ U_i \mid X_i &\sim N(0,\sigma_0^2)\end{aligned}$$

- PDF of U_i given X_i, at u:

$$\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\Big(-\frac{u^2}{2\sigma_0^2}\Big)$$

- Distribution of Y_i given $X_i = x$:

$$Y_i \mid X_i = x \sim N(x^\top\beta_0,\ \sigma_0^2)$$

- PDF of Y_i at y given $X_i = x$, with $u = y - x^\top\beta_0$:

$$\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\Big(-\frac{(y-x^\top\beta_0)^2}{2\sigma_0^2}\Big)$$

MLE in the normal regression

- Conditional likelihood:

$$\begin{aligned} L_n(\beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(Y_i - X_i^\top \beta)^2}{2\sigma^2} \Big) \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}} \Big)^n \exp\Big(- \frac{\sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2}{2\sigma^2} \Big) \end{aligned}$$

- Average conditional log likelihood:

$$\begin{aligned} \frac{1}{n} \log L_n(\beta, \sigma^2) &= \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 \\ &= \log \left((2\pi\sigma^2)^{-1/2} \right) - \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 \\ &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 \end{aligned}$$

- MLE of β:

$$\begin{aligned} \hat{\beta}^{MLE} &= \arg\max_{\beta} \frac{1}{n} \log L_n(\beta, \sigma^2) \\ &= \arg\max_{\beta} \Big(- \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 \Big) \\ &= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 \\ &= \hat{\beta}^{OLS} \end{aligned}$$

Conditional likelihood for probit and logit

- iid observations (Y_i, X_i):

$$\{(Y_i, X_i) : i = 1, \dots, n\}$$

- Probit and logit response probability:

$$\mathrm{P}(Y_i = 1 \mid X_i = x) = \underbrace{F(x^\top \beta_0)}_{= \theta_0(x)}$$

- Conditional PMF, $y \in \{0, 1\}$:

$$\mathrm{P}(Y_i = y \mid X_i) = F(X_i^\top \beta_0)^y (1 - F(X_i^\top \beta_0))^{1-y}$$

- Probability of the observed outcomes given the regressors:

$$\begin{aligned} &\mathrm{P}(Y_1 = y_1, \dots, Y_n = y_n \mid X_1, \dots, X_n) \\ &= \prod_{i=1}^{n} \mathrm{P}(Y_i = y_i \mid X_i) \\ &= \prod_{i=1}^{n} F(X_i^\top \beta_0)^{y_i} (1 - F(X_i^\top \beta_0))^{1-y_i} \end{aligned}$$

Conditional log likelihood and grid search

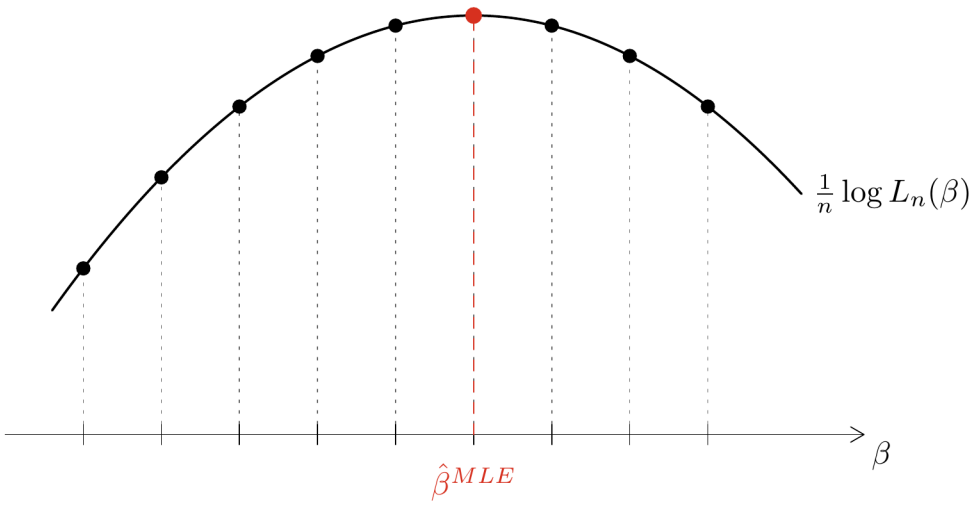
- Average conditional log likelihood:

$$\frac{1}{n}\log L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \log F(X_i^\top \beta) + (1-Y_i)\log\left(1 - F(X_i^\top \beta)\right)\right\}$$

- MLE of β:

$$\hat{\beta}^{MLE} = \arg\max_{\beta \in \mathbb{R}^k} \frac{1}{n}\log L_n(\beta)$$

- Grid search, scalar β:



Consistency of the MLE

General setup

- Data $\{W_i : i = 1, \ldots, n\}$, PDF or PMF $f(w, \theta_0)$:

$$W_i \overset{\text{iid}}{\sim} f(w, \theta_0), \qquad \theta_0 \in \mathbb{R}^k : \text{ true value}$$

- **Assume:** common support $\{w : f(w, \theta) > 0\}$ for all θ.
- **Assume:** $\mathrm{E}\left[|\log f(W_i, \theta_0)|\right] < \infty$.
- Logit: $W_i = (Y_i, X_i^\top)^\top$ and $\theta = \beta$.
- f: conditional PMF of Y_i given X_i:

$$f(w, \theta_0) = \Lambda(x^\top \theta_0)^y (1 - \Lambda(x^\top \theta_0))^{1-y}, \qquad w = (y, x^\top)^\top, \quad y \in \{0, 1\}$$

- MLE, maximizer of the average log likelihood:

$$\hat{\theta} = \arg\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \log f(W_i, \theta)$$

- To show (consistency):

$$\hat{\theta} \to_p \theta_0$$

Consistency argument

- θ^*: unique maximizer of $\mathrm{E}\left[\log f(W_i, \theta)\right]$.
- Weak LLN for each θ, then maximize both sides, under regularity conditions:

$$\begin{array}{ccc} \frac{1}{n}\sum_{i=1}^{n} \log f(W_i, \theta) & \to_p & \mathrm{E}\left[\log f(W_i, \theta)\right] \\ \downarrow \arg\max & & \downarrow \arg\max \\ \hat{\theta} & \to_p & \theta^* \end{array}$$

- Consistency if the population maximizer is θ_0:

$$\boxed{\theta^* = \theta_0 \quad \Longrightarrow \quad \hat{\theta} \to_p \theta_0}$$

Jensen's inequality and the population log likelihood

- We will use Jensen's inequality, h concave:

$$\mathrm{E}\left[h(Z)\right] \leq h(\mathrm{E}\left[Z\right])$$

- **Claim.** For all θ:

$$\mathrm{E}\left[\log f(W_i, \theta)\right] \leq \mathrm{E}\left[\log f(W_i, \theta_0)\right]$$

- Proof, $f(w, \theta_0)$: the true PDF of W_i:

$$\begin{aligned} &\mathrm{E}\left[\log f(W_i, \theta)\right] - \mathrm{E}\left[\log f(W_i, \theta_0)\right] \\ &= \mathrm{E}\left[\log f(W_i, \theta) - \log f(W_i, \theta_0)\right] \\ &= \mathrm{E}\left[\log \frac{f(W_i, \theta)}{f(W_i, \theta_0)}\right] \\ &\leq \log \mathrm{E}\left[\frac{f(W_i, \theta)}{f(W_i, \theta_0)}\right] \qquad \text{(Jensen's inequality, log concave)} \\ &= \log \int \frac{f(w, \theta)}{f(w, \theta_0)} f(w, \theta_0)\, dw \\ &= \log \underbrace{\int f(w, \theta)\, dw}_{=1} \\ &= 0 \end{aligned}$$

Strict inequality and identification

- Suppose $\theta \neq \theta_0$, log strictly concave:

$$\mathrm{E}\left[\log \frac{f(W_i, \theta)}{f(W_i, \theta_0)}\right] < \log \mathrm{E}\left[\frac{f(W_i, \theta)}{f(W_i, \theta_0)}\right]$$

- Unless the ratio is constant with probability one:

$$\begin{aligned} &\frac{f(W_i, \theta)}{f(W_i, \theta_0)} = \text{const} \\ \iff \quad &f(W_i, \theta) = \underbrace{\text{const}}_{=1} \cdot f(W_i, \theta_0) \end{aligned}$$

- **Assume** (identification), for all $\theta \neq \theta_0$:

$$\mathrm{P}(f(W_i, \theta) \neq f(W_i, \theta_0)) > 0$$

- Hence:

$$\begin{aligned} & \mathrm{E}\left[\log f(W_i,\theta)\right] < \mathrm{E}\left[\log f(W_i,\theta_0)\right] \quad \text{for all } \theta \neq \theta_0 \\ \implies \quad & \theta^* = \theta_0 \\ \implies \quad & \hat{\theta} \to_p \theta_0 \end{aligned}$$

Identification in the logit

- Logit:

$$f(W_i,\theta) = \Lambda(X_i^\top\theta)^{Y_i}(1-\Lambda(X_i^\top\theta))^{1-Y_i}$$

- Λ strictly increasing:

$$\Lambda(u_1) = \Lambda(u_2) \quad \iff \quad u_1 = u_2$$

- Suppose $\theta_1 \neq \theta_0$:

$$\begin{aligned} f(W_i,\theta_1) = f(W_i,\theta_0) \quad & \iff \quad X_i^\top\theta_1 = X_i^\top\theta_0 \\ & \iff \quad X_i^\top \underbrace{(\theta_1-\theta_0)}_{=a} = 0 \end{aligned}$$

- To rule out $X_i^\top a = 0$ with probability one, **assume:** $\mathrm{E}\left[X_iX_i^\top\right]$ p.d.:

$$\begin{aligned} a \neq 0 \quad & \implies \quad \mathrm{P}(X_i^\top a \neq 0) > 0 \\ & \implies \quad \mathrm{P}(f(W_i,\theta_1) \neq f(W_i,\theta_0)) > 0 \end{aligned}$$

Asymptotic normality and standard errors

Gradient and Hessian

- $h : \mathbb{R}^k \to \mathbb{R}$, gradient:

$$\underset{k\times 1}{\frac{\partial h}{\partial \theta}} = \begin{bmatrix} \frac{\partial h}{\partial \theta_1} \\ \vdots \\ \frac{\partial h}{\partial \theta_k} \end{bmatrix}$$

- Hessian:

$$\underset{k\times k}{\underset{k\times 1}{\frac{\partial}{\partial\theta}}\underset{1\times k}{\frac{\partial}{\partial\theta^\top}}h} = \begin{bmatrix} \frac{\partial^2 h}{\partial\theta_1\partial\theta_1} & \cdots & \frac{\partial^2 h}{\partial\theta_1\partial\theta_k} \\ \vdots & & \vdots \\ \frac{\partial^2 h}{\partial\theta_k\partial\theta_1} & \cdots & \frac{\partial^2 h}{\partial\theta_k\partial\theta_k} \end{bmatrix}$$

- Score of observation i, gradient of its log likelihood:

$$\frac{\partial}{\partial\theta}\log f(W_i,\theta)$$

First-order condition and expansion

- We will use the expansion of h around θ_0:

$$h(\hat{\theta}) \approx h(\theta_0) + \frac{\partial h(\theta_0)}{\partial\theta^\top}(\hat{\theta}-\theta_0)$$

- First-order condition for $\hat{\theta}$, each element expanded around θ_0:

$$\begin{aligned}
\underset{k\times 1}{0} &= \frac{\partial}{\partial\theta}\Big[\frac{1}{n}\sum_{i=1}^{n}\log f(W_i,\hat{\theta})\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(W_i,\hat{\theta}) \\
&\approx \frac{1}{n}\sum_{i=1}^{n}\underbrace{\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)}_{k\times 1} + \underbrace{\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\Big]}_{k\times k}\underbrace{(\hat{\theta}-\theta_0)}_{k\times 1}
\end{aligned}$$

- OLS normal equations:

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n}X_i\hat{U}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i(Y_i - X_i^\top\hat{\beta}^{OLS}) \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i(U_i - X_i^\top(\hat{\beta}^{OLS}-\beta_0))
\end{aligned}$$

Solving for $\sqrt{n}(\hat{\theta}-\theta_0)$

- Solve for $\hat{\theta}-\theta_0$, then multiply by $\sqrt{n}$:

$$\begin{aligned}
&\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\Big](\hat{\theta}-\theta_0) \approx -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(W_i,\theta_0) \\
\implies\quad &\hat{\theta}-\theta_0 \approx -\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(W_i,\theta_0) \\
\implies\quad &\sqrt{n}(\hat{\theta}-\theta_0) \\
&\qquad \approx -\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\Big]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)
\end{aligned}$$

- Same form for the scaled OLS estimation error:

$$\sqrt{n}(\hat{\beta}^{OLS}-\beta_0) = \Big(\frac{1}{n}\sum_{i=1}^{n}X_iX_i^\top\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_iU_i$$

Average Hessian and scaled score

- Weak LLN for the average Hessian, limit assumed negative definite:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top} \to_p \mathrm{E}\left[\frac{\partial^2\log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\right]$$

- Variance of the score:

$$\begin{aligned}
\Omega_0 &= \mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)\right) \\
&= \mathrm{E}\left[\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)\,\frac{\partial}{\partial\theta^\top}\log f(W_i,\theta_0)\right]
\end{aligned}$$

- Second-moment form of Ω_0 and the CLT need:

$$\mathrm{E}\left[\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)\right] = 0$$

- CLT for the scaled score:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(W_i, \theta_0) \to_d N(0, \Omega_0)$$

Mean of the score

- **Assume:** $\frac{\partial}{\partial \theta}$ and $\int$ interchange.
- Population first-order condition, $\theta_0 = \arg\max_\theta \mathrm{E}\left[\log f(W_i, \theta)\right]$:

$$\begin{aligned} \mathrm{E}\left[\frac{\partial}{\partial \theta} \log f(W_i, \theta_0)\right] &= \frac{\partial}{\partial \theta} \mathrm{E}\left[\log f(W_i, \theta)\right] \Big|_{\theta=\theta_0} \\ &= 0 \end{aligned}$$

- A more direct argument:

$$\begin{aligned} \mathrm{E}\left[\frac{\partial}{\partial \theta} \log f(W_i, \theta_0)\right] &= \mathrm{E}\left[\frac{\partial f(W_i, \theta_0)/\partial \theta}{f(W_i, \theta_0)}\right] \\ &= \int \frac{\partial f(w, \theta_0)/\partial \theta}{f(w, \theta_0)} \cdot f(w, \theta_0)\, dw \\ &= \int \frac{\partial f(w, \theta_0)}{\partial \theta}\, dw \\ &= \frac{\partial}{\partial \theta} \underbrace{\int f(w, \theta_0)\, dw}_{=1} \\ &= 0 \end{aligned}$$

Sandwich form

- Weak LLN for the average Hessian, CLT for the scaled score:

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &\approx -\Big[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(W_i, \theta_0)}{\partial \theta\, \partial \theta^\top}\Big]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(W_i, \theta_0) \\ &\to_d -\Big(\mathrm{E}\left[\frac{\partial^2 \log f(W_i, \theta_0)}{\partial \theta\, \partial \theta^\top}\right]\Big)^{-1} N(0, \Omega_0) \end{aligned}$$

- Asymptotic variance, the Hessian symmetric:

$$\mathrm{AsyVar}(\hat{\theta}) = \Big(\mathrm{E}\left[\frac{\partial^2 \log f(W_i, \theta_0)}{\partial \theta\, \partial \theta^\top}\right]\Big)^{-1} \Omega_0 \Big(\mathrm{E}\left[\frac{\partial^2 \log f(W_i, \theta_0)}{\partial \theta\, \partial \theta^\top}\right]\Big)^{-1}$$

Information matrix equality

- **Claim.**

$$\mathrm{E}\left[\frac{\partial^2 \log f(W_i, \theta_0)}{\partial \theta\, \partial \theta^\top}\right] = -\Omega_0$$

- We will use:

$$\begin{aligned} \mathrm{E}\left[\frac{\partial^2 f(W_i, \theta_0)/\partial \theta\, \partial \theta^\top}{f(W_i, \theta_0)}\right] &= \int \frac{\partial^2 f(w, \theta_0)}{\partial \theta\, \partial \theta^\top} \frac{1}{f(w, \theta_0)} \cdot f(w, \theta_0)\, dw \\ &= \frac{\partial^2}{\partial \theta\, \partial \theta^\top} \underbrace{\int f(w, \theta_0)\, dw}_{=1} \\ &= 0 \end{aligned}$$

- Differentiate the score:

$$
\begin{aligned}
&\mathrm{E}\left[\frac{\partial^2 \log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\right] \\
&= \mathrm{E}\left[\frac{\partial}{\partial\theta^\top}\,\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)\right] \\
&= \mathrm{E}\left[\frac{\partial}{\partial\theta^\top}\,\frac{\partial f(W_i,\theta_0)/\partial\theta}{f(W_i,\theta_0)}\right] \\
&= \mathrm{E}\left[\frac{\partial^2 f(W_i,\theta_0)/\partial\theta\,\partial\theta^\top}{f(W_i,\theta_0)} - \underbrace{\frac{\partial f(W_i,\theta_0)/\partial\theta}{f(W_i,\theta_0)}}_{\frac{\partial}{\partial\theta}\log f(W_i,\theta_0)}\,\underbrace{\frac{\partial f(W_i,\theta_0)/\partial\theta^\top}{f(W_i,\theta_0)}}_{\frac{\partial}{\partial\theta^\top}\log f(W_i,\theta_0)}\right] \\
&= \underbrace{\mathrm{E}\left[\frac{\partial^2 f(W_i,\theta_0)/\partial\theta\,\partial\theta^\top}{f(W_i,\theta_0)}\right]}_{=0} - \Omega_0 \\
&= \boxed{-\Omega_0}
\end{aligned}
$$

Information matrix and asymptotic variance

- **Definition.** Information matrix:

$$I(\theta_0) = -\mathrm{E}\left[\frac{\partial^2 \log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\right]$$

- $I(\theta_0)$: symmetric, p.d.
- Substitute Ω_0 from the information matrix equality:

$$
\begin{aligned}
&\mathrm{AsyVar}(\hat\theta) \\
&= \Big(\mathrm{E}\left[\frac{\partial^2 \log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\right]\Big)^{-1} \underbrace{\Omega_0}_{=-\mathrm{E}\left[\frac{\partial^2 \log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\right]} \Big(\mathrm{E}\left[\frac{\partial^2 \log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\right]\Big)^{-1} \\
&= -\Big(\mathrm{E}\left[\frac{\partial^2 \log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\right]\Big)^{-1} \\
&= \Big(-\mathrm{E}\left[\frac{\partial^2 \log f(W_i,\theta_0)}{\partial\theta\,\partial\theta^\top}\right]\Big)^{-1} \\
&= I(\theta_0)^{-1}
\end{aligned}
$$

- Asymptotic distribution of the MLE:

$$\boxed{\sqrt{n}(\hat\theta - \theta_0) \to_d N(0,\ I(\theta_0)^{-1})}$$

Standard errors

- $\hat\theta_j$, $\theta_{0,j}$: element j of $\hat\theta$, θ_0.
- Estimator of Ω_0:

$$\hat\Omega = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(W_i,\hat\theta)}{\partial\theta}\,\frac{\partial \log f(W_i,\hat\theta)}{\partial\theta^\top}$$

- Standard error of $\hat\theta_j$, from the average Hessian at $\hat\theta$:

$$\text{std. err}(\hat\theta_j) = \sqrt{\frac{\Big(\Big[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(W_i,\hat\theta)}{\partial\theta\,\partial\theta^\top}\Big]^{-1}\Big)_{jj}}{n}}$$

- Or from $\hat{\Omega}$:

$$\text{std. err}(\hat{\theta}_j) = \sqrt{\frac{\left(\hat{\Omega}^{-1}\right)_{jj}}{n}}$$

- t test of $H_0 : \theta_{0,j} = c$ vs $H_1 : \theta_{0,j} \neq c$: $|\hat{\theta}_j - c|/\text{std. err}(\hat{\theta}_j) > z_{1-\alpha/2} \implies$ reject.

Estimated average marginal effect

AME estimator

- AME estimator from $\hat{\beta}$ and known f:

$$\widehat{\text{AME}}_j = \frac{1}{n}\sum_{i=1}^{n} f(X_i^\top \hat{\beta})\,\hat{\beta}_j$$

Stein's lemma and LPM slopes

Normal regressors and Stein's lemma

- $\text{E}[Y_i \mid X_i] = g(X_{i,2})$: nonlinear, unknown.
- AME, with $g'(x) = \partial g(x)/\partial x$:

$$\text{AME} = \text{E}\left[g'(X_{i,2})\right]$$

- Probit and logit:

$$\begin{aligned} g(x) &= F(\beta_1 + x^\top \beta_2) \\ g'(x) &= f(\beta_1 + x^\top \beta_2)\,\beta_2 \end{aligned}$$

- **Stein's lemma.** $\underset{1\times 1}{X_{i,2}} \sim N(\mu_X, \sigma_X^2)$, g differentiable:

$$\underbrace{\text{E}\left[g(X_{i,2})(X_{i,2} - \mu_X)\right]}_{=\text{Cov}(g(X_{i,2}),\, X_{i,2})} = \sigma_X^2\,\text{E}\left[g'(X_{i,2})\right]$$

- Proof: appendix.
- Vector $\underset{(k-1)\times 1}{X_{i,2}} \sim N(\mu_X, \underset{(k-1)\times(k-1)}{\Sigma})$, Σ p.d., g differentiable:

$$\text{E}\left[g(X_{i,2})(X_{i,2} - \mu_X)\right] = \Sigma\,\text{E}\left[g'(X_{i,2})\right]$$

LPM slope with a scalar normal regressor

- Scalar $X_{i,2} \sim N(\mu_X, \sigma_X^2)$, g differentiable.

- LPM: population OLS of Y_i on 1 and $X_{i,2}$, slope δ:

$$\begin{aligned}
\delta &= \frac{\overbrace{\mathrm{E}\left[(X_{i,2}-\mu_X)Y_i\right]}^{=\mathrm{Cov}(X_{i,2},Y_i)}}{\mathrm{E}\left[(X_{i,2}-\mu_X)^2\right]} \\
&= \frac{\mathrm{E}\left[(X_{i,2}-\mu_X)Y_i\right]}{\sigma_X^2} \\
&\overset{\text{LIE}}{=} \frac{\mathrm{E}\left[(X_{i,2}-\mu_X)\,g(X_{i,2})\right]}{\sigma_X^2} \\
&= \frac{\sigma_X^2\,\mathrm{E}\left[g'(X_{i,2})\right]}{\sigma_X^2} \qquad \text{(Stein's lemma)} \\
&= \mathrm{E}\left[g'(X_{i,2})\right] \\
&= \boxed{\text{AME}}
\end{aligned}$$

LPM slopes with vector normal regressors

- Vector $X_{i,2} \sim N(\mu_X, \Sigma)$, Σ p.d.
- Population OLS slopes δ (Y_i on 1, $X_{i,2}$):

$$\begin{aligned}
\delta &= \Sigma^{-1}\mathrm{Cov}\left(X_{i,2}, Y_i\right) \\
&= \Sigma^{-1}\mathrm{E}\left[(X_{i,2}-\mu_X)Y_i\right] \\
&\overset{\text{LIE}}{=} \Sigma^{-1}\mathrm{E}\left[g(X_{i,2})(X_{i,2}-\mu_X)\right] \\
&= \Sigma^{-1}\Sigma\,\mathrm{E}\left[g'(X_{i,2})\right] \qquad \text{(Stein's lemma)} \\
&= \mathrm{E}\left[g'(X_{i,2})\right] \\
&= \boxed{\text{AME}}
\end{aligned}$$

- Normal $X_{i,2}$: the LPM slopes equal the AME, for any differentiable g.

Summary

Binary choice and identification

- $Y_i = \mathbf{1}(X_i^\top\beta + \varepsilon_i > 0)$, $\varepsilon_i \mid X_i \sim F \implies \mathrm{P}(Y_i = 1 \mid X_i) = 1 - F(-X_i^\top\beta)$.
- F symmetric around zero $\implies \mathrm{P}(Y_i = 1 \mid X_i) = F(X_i^\top\beta)$.
- LPM: $\mathrm{E}\left[Y_i \mid X_i\right] = \mathrm{P}(Y_i = 1 \mid X_i) = X_i^\top\beta$.
- Probit: $F = \Phi$.
- Logit: $F(t) = \Lambda(t) = e^t/(1+e^t)$.
- Logit, log odds ratio: $\log\left(\mathrm{P}(Y_i = 1 \mid X_i)/\mathrm{P}(Y_i = 0 \mid X_i)\right) = X_i^\top\beta$.
- $\varepsilon_i \mid X_i \sim N(0, \sigma^2)$: only β/σ can be identified.
- $\varepsilon_i \mid X_i \sim N(\mu, \sigma^2)$: μ absorbed by the intercept.
- Probit: $\Phi^{-1}(\mathrm{P}(Y_i = 1 \mid X_i = x)) = x^\top\beta$.
- Probit and logit: k linearly independent values of $X_i \implies \beta$ identified.

Marginal effects

- LPM: $\partial\,\mathrm{P}(Y_i = 1 \mid X_i = x)/\partial x_j = \beta_j$ at every x.

- Probit and logit: $\partial \, \mathrm{P}(Y_i = 1 \mid X_i = x)/\partial x_j = f(x^\top \beta) \, \beta_j$.
- $\mathrm{AME}_j = \mathrm{E}\left[f(X_i^\top \beta)\right] \, \beta_j$.

Maximum likelihood and consistency

- $W_i \overset{\text{iid}}{\sim} f(w, \theta_0)$: $\hat{\theta} = \arg\max_\theta \frac{1}{n} \sum_{i=1}^n \log f(W_i, \theta)$.
- **Assume:** common support $\{w : f(w, \theta) > 0\}$ for all θ.
- **Assume:** $\mathrm{E}\left[|\log f(W_i, \theta_0)|\right] < \infty$.
- Bernoulli trials: $\hat{\theta} = \bar{Y}$.
- Normal regression: $\hat{\beta}^{MLE} = \hat{\beta}^{OLS}$.
- Probit and logit:

$$\frac{1}{n} \log L_n(\beta) = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i \log F(X_i^\top \beta) + (1 - Y_i) \log\left(1 - F(X_i^\top \beta)\right) \right\}$$

  - $\hat{\beta}^{MLE}$ by numerical maximization, for example grid search.
- Jensen's inequality: $\mathrm{E}\left[\log f(W_i, \theta)\right] \leq \mathrm{E}\left[\log f(W_i, \theta_0)\right]$ for all θ.
- Under regularity conditions: $\mathrm{P}(f(W_i, \theta) \neq f(W_i, \theta_0)) > 0$ for all $\theta \neq \theta_0 \implies \hat{\theta} \to_p \theta_0$.
- Logit: $\mathrm{E}\left[X_i X_i^\top\right]$ p.d. $\implies$ identification condition.

Asymptotic normality

- **Assume:** $\frac{\partial}{\partial \theta}$ and $\int$ interchange.
- $\mathrm{E}\left[\frac{\partial}{\partial \theta} \log f(W_i, \theta_0)\right] = 0$.
- $\Omega_0 = \mathrm{E}\left[\frac{\partial}{\partial \theta} \log f(W_i, \theta_0) \, \frac{\partial}{\partial \theta^\top} \log f(W_i, \theta_0)\right]$.
- Information matrix equality: $\mathrm{E}\left[\frac{\partial^2 \log f(W_i, \theta_0)}{\partial \theta \, \partial \theta^\top}\right] = -\Omega_0$.
- $I(\theta_0) = -\mathrm{E}\left[\frac{\partial^2 \log f(W_i, \theta_0)}{\partial \theta \, \partial \theta^\top}\right]$: information matrix, p.d.
- $\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, \; I(\theta_0)^{-1})$.
- $\text{std. err}(\hat{\theta}_j) = \sqrt{(\hat{\Omega}^{-1})_{jj}/n}$.
- Or minus the average Hessian at $\hat{\theta}$ in place of $\hat{\Omega}$.

AME estimator and Stein's lemma

- $\widehat{\mathrm{AME}}_j = \frac{1}{n} \sum_{i=1}^n f(X_i^\top \hat{\beta}) \, \hat{\beta}_j$.
- Stein's lemma, $X_{i,2} \sim N(\mu_X, \sigma_X^2)$, g differentiable: $\mathrm{E}\left[g(X_{i,2})(X_{i,2} - \mu_X)\right] = \sigma_X^2 \, \mathrm{E}\left[g'(X_{i,2})\right]$.
- $\mathrm{E}\left[Y_i \mid X_i\right] = g(X_{i,2})$, $X_{i,2} \sim N(\mu_X, \Sigma)$, Σ p.d. $\implies$ LPM slopes $\delta = \mathrm{E}\left[g'(X_{i,2})\right] = \mathrm{AME}$, for any differentiable g.

Appendix

Proof of Stein's lemma

- Suppose $\underset{1\times 1}{X_{i,2}} \sim N(\mu_X, \sigma_X^2)$, g differentiable.

- PDF of $X_{i,2}$:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}}\, e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

- Write the expectation with the normal PDF:

$$\begin{aligned}
\mathrm{E}\left[g(X_{i,2})(X_{i,2}-\mu_X)\right] &= \int g(x)\,(x-\mu_X)\,f_X(x)\,dx \\
&= \sigma_X^2 \int g(x)\,\frac{x-\mu_X}{\sigma_X^2}\,f_X(x)\,dx \\
&= \sigma_X^2 \int g(x)\,\frac{x-\mu_X}{\sigma_X^2}\,\frac{1}{\sqrt{2\pi\sigma_X^2}}\,e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}\,dx \\
&= -\sigma_X^2 \int g(x)\,\Big(-\frac{x-\mu_X}{\sigma_X^2}\Big)\frac{1}{\sqrt{2\pi\sigma_X^2}}\,e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}\,dx
\end{aligned}$$

- Differential of f_X:

$$\begin{aligned}
df_X(x) &= \frac{1}{\sqrt{2\pi\sigma_X^2}}\,e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \times \frac{-2(x-\mu_X)}{2\sigma_X^2}\,dx \\
&= -\frac{x-\mu_X}{\sigma_X^2}\,f_X(x)\,dx
\end{aligned}$$

- Integration by parts:

$$\int_{-\infty}^{\infty} g\,dh = g\,h\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} h\,dg$$

- Continue, integrate by parts with $h = f_X$:

$$\begin{aligned}
\mathrm{E}\left[g(X_{i,2})(X_{i,2}-\mu_X)\right] &= -\sigma_X^2 \int g(x)\,df_X(x) \\
&= \underbrace{\sigma_X^2 \lim_{a\to\infty}\Big[-g(x)\,f_X(x)\Big]_{-a}^{a}}_{=0} \\
&\qquad + \sigma_X^2 \int f_X(x)\,dg(x) \\
&= \sigma_X^2 \int f_X(x)\,g'(x)\,dx \\
&= \boxed{\sigma_X^2\,\mathrm{E}\left[g'(X_{i,2})\right]}
\end{aligned}$$