Econometric Methods I 2016-09-08
Lecture 1 Junyong Kim

1. Linear Models

a. Regression

Think of this as the empirical analysis of the relationship between y (outcome variable) and **x** (vector of regressors). → convenient to think of regression as directed to the prediction of y given **x**.

Define

$$\begin{aligned} \hat{y} &\equiv \textbf{predictor of y as a function of } \mathbf{x} \\ \hat{u} &\equiv y - \hat{y} \\ &\equiv \textbf{prediction error} \\ L(\hat{u}) &= L(y - \hat{y}) \\ &\equiv \text{loss associated with } \hat{u} \\ &\quad (\text{differences in } \textbf{loss functions } L(\cdot) \text{ imply different regression models}) \end{aligned}$$

Assume

$$\begin{aligned} &L(0) = 0 \\ &L(u) \geq 0, \qquad \forall u \in \mathbb{R} \\ &L(u) \text{ is increasing in } |u| \end{aligned}$$

Assume $(y,\hat{y})$ is a random vector; goal is to minimize the (conditional) expected loss.

$$E[L(\hat{u})|\mathbf{x}]$$

In Economics, often no obvious choice of $L(\cdot)$, so usually specify quadratic loss, i.e.

$$L(u) = u^2 \Rightarrow E[\hat{u}^2|\mathbf{x}]$$

i.e. mean squared prediction error (MSPE).

In this case, the optimal predictor of y is the conditional mean $E[y|\mathbf{x}]$. One usually imposes a parametric model on $E[y|\mathbf{x}]$, i.e. one assumes that $\exists$ a function $g(\mathbf{x},\boldsymbol{\beta})$ and a vector $\boldsymbol{\beta}$ such that

$$P[E[y|\mathbf{x}] = g(\mathbf{x}, \boldsymbol{\beta})] = 1$$

Suppose we have observations

$$(\mathbf{x}_i^\top, y_i), \qquad i = 1, \cdots, n$$

Then the optimal predictor is

$$\hat{y} = g(\mathbf{x}, \widehat{\boldsymbol{\beta}}), \qquad \text{where } \widehat{\boldsymbol{\beta}} \text{ minimizes}$$

$$\sum_{i=1}^{n} L(\hat{u}_i) = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(y_i - g(\mathbf{x}_i, \boldsymbol{\beta})\right)^2$$

i.e. the sum of squared residuals, i.e. $\widehat{\boldsymbol{\beta}}$ is computed by nonlinear least squares (NLLS). If $g(\mathbf{x},\boldsymbol{\beta})$ is actually linear in $\mathbf{x}$, $\boldsymbol{\beta}$, i.e. if $g(\mathbf{x},\boldsymbol{\beta})=\mathbf{x}^\top\boldsymbol{\beta}$, then the optimal predictor is

$$\hat{y} = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}$$

Where $\widehat{\boldsymbol{\beta}}$ is computed by ordinary least squares (OLS) (Legendre, 1805).

By contrast, under absolute error loss (i.e. L(u)=|u|), the optimal predictor is the conditional median of y given **x**, i.e.

$$F_{y|\mathbf{x}}^{-1}(0.5)$$

If one additionally assumes linearity in the conditional median, i.e. if $F_{y|\mathbf{x}}^{-1}(0.5) = \mathbf{x}^\top\boldsymbol{\beta}$, then the optimal predictor is $\hat{y} = \mathbf{x}^\top\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ minimizes

$$\sum_{i=1}^{n} L(u_i) = \sum_{i=1}^{n} \left|y_i - \mathbf{x}_i^\top\boldsymbol{\beta}\right|$$

i.e. the sum of absolute residuals, i.e. $\widehat{\boldsymbol{\beta}}$ is the least absolute deviations (LAD) estimator (Boscovich, 1757).

Go back to prediction under quadratic error loss. If $E[y|\mathbf{x}]=\mathbf{x}^\top\boldsymbol{\beta}$ (w.p. 1), then $\boldsymbol{\beta}$ is of intrinsic interest. It has a causal interpretation (think a one unit increase in a component of **x** implies a change in the optimal prediction of y equal to the corresponding component of $\boldsymbol{\beta}$). So, OLS estimates an intrinsically important quantity when $E[y|\mathbf{x}]$ is linear.

If $E[y|\mathbf{x}]$ is nonlinear, might not be clear what OLS actually estimates.

Differentiate w.r.t. $\boldsymbol{\beta}$ the expected loss

$$E[(y - \mathbf{x}^\top\boldsymbol{\beta})^2]$$

and set the derivative to $\mathbf{0}$.

$$\begin{aligned} -2E[\mathbf{x}(y - \mathbf{x}^\top\boldsymbol{\beta})] &= \mathbf{0} \\ \Leftrightarrow \boldsymbol{\beta} &= (E[\mathbf{x}\mathbf{x}^\top])^{-1}E[\mathbf{x}y] \end{aligned}$$

which is consistently estimated by OLS.

So, OLS estimates the best linear predictor of y under quadratic loss. Now suppose

$$\mathbf{x} = [1 \quad \tilde{\mathbf{x}}^\top]^\top, \qquad \boldsymbol{\beta} = [\beta_1 \quad \tilde{\boldsymbol{\beta}}^\top]^\top$$

Then, $\mathbf{x}^\top\boldsymbol{\beta}$ "has an intercept" since $\mathbf{x}^\top\boldsymbol{\beta} = \beta_1 + \tilde{\mathbf{x}}^\top\tilde{\boldsymbol{\beta}}$. Then the first-order conditions for minimization of expected loss can be written as

$$\begin{aligned} -2E\left[y - \beta_1 - \tilde{\mathbf{x}}^\top\tilde{\boldsymbol{\beta}}\right] &= 0 \\ -2E\left[\tilde{\mathbf{x}}\left(y - \beta_1 - \tilde{\mathbf{x}}^\top\tilde{\boldsymbol{\beta}}\right)\right] &= \mathbf{0} \end{aligned}$$

This implies the existence of an unrestricted linear model for y since we can write

$$y = \beta_1 + \tilde{\mathbf{x}}^\top\tilde{\boldsymbol{\beta}} + u, \qquad \text{where } E[u] = 0 \text{ and } Cov[\tilde{\mathbf{x}}, u] = \mathbf{0}$$

Solve the above FOCs to get

$$\begin{aligned} \beta_1 &= E[y] - E[\tilde{\mathbf{x}}^\top]\tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\beta}} &= (Var[\tilde{\mathbf{x}}])^{-1}Cov[\tilde{\mathbf{x}}, y] \end{aligned}$$

which are consistently estimable by OLS.

Also, for any combination of regressors **x** and regressand y, one can (always) specify the linear projection model

$$y = \beta_1 + \tilde{\mathbf{x}}^\top\tilde{\boldsymbol{\beta}} + u, \qquad \text{where } E[u] = 0 \text{ and } Cov[\tilde{\mathbf{x}}, u] = \mathbf{0}$$

and where β_1, $\widetilde{\boldsymbol{\beta}}$ are given in the solutions of FOCs.

Note: Neither β_1 nor $\widetilde{\boldsymbol{\beta}}$ has a causal interpretation unless one imposes the additional restriction that $E[u|\mathbf{x}]=0$ (in which case we have a linear regression model).

b. OLS for linear regression

Observe

$$\left(\mathbf{x}_i^\top, y_i\right), \qquad i = 1, \cdots, n$$

where each $\mathbf{x}_i$ is d-variate. Assume

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i, \qquad i = 1, \cdots, n$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is unknown and each u_i is random and unobserved. Write

$$\mathbf{y} \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} \equiv \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}, \qquad \mathbf{u} \equiv \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}, \qquad \boldsymbol{\beta} \equiv \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$$

Then, $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$. The OLS estimator $\widehat{\boldsymbol{\beta}}$ is defined as

$$\widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^d}{\operatorname{argmax}} (\mathbf{y} - \mathbf{X}\mathbf{b})^\top (\mathbf{y} - \mathbf{X}\mathbf{b})$$

i.e. the minimizer of the sum of squared residuals. Differentiate the objective function w.r.t. **b**, solve the FOCs to get

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

(this assumes $\mathbf{X}^\top\mathbf{X}$ is invertible; otherwise replace $(\mathbf{X}^\top\mathbf{X})^{-1}$ with a generalized inverse)

Proposition (asymptotic distribution of OLS) Assume

(A1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$

(A2) $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$

$E[\mathbf{u}\mathbf{u}^\top|\mathbf{X}] = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega} = \operatorname{diag}\left[\sigma_i^2\right] = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$ for constants $\sigma_1, \cdots, \sigma_n > 0$

(A3) $\mathbf{X}$ has full column rank (so $\mathbf{X}^\top\mathbf{X}$ is nonsinular)

(A4) $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \operatorname{plim}_{n\to\infty} \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ exists and is finite and nonsingular

(A5) $\frac{1}{\sqrt{n}}\mathbf{X}^\top\mathbf{u} \xrightarrow{d} N_d(\mathbf{0}, \mathbf{S})$, where $\mathbf{S} = \operatorname{plim}_{n\to\infty} \frac{1}{n}\mathbf{X}^\top\mathbf{u}\mathbf{u}^\top\mathbf{X}$

Then, the OLS estimator $\widehat{\boldsymbol{\beta}}$ is consistent and also satisfies

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &\xrightarrow{p} \boldsymbol{\beta} \\ \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &\xrightarrow{d} N_d\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\right) \end{aligned}$$

Note

1.

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top, \qquad \text{so under A4,} \qquad \boldsymbol{\Sigma}_{\mathbf{xx}} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\mathbf{x}_i\mathbf{x}_i^\top\right]$$

2.

$$\frac{1}{n}\mathbf{X}^\top\mathbf{u}\mathbf{u}^\top\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}u_i^2\mathbf{x}_i\mathbf{x}_i^\top, \qquad \text{so under A5,} \qquad \mathbf{S} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[u_i^2\mathbf{x}_i\mathbf{x}_i^\top\right]$$

3. A1 implies that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{u} = \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{u}$$

So consistency is straightforward via A4, A5 and the continuous mapping theorem.

4. A2 implies that

$$\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$$

which implies that $\boldsymbol{\beta}$ has a causal interpretation; A2 also permits (conditional) heteroscedasticity.

5. A3 rules out perfect multicollinearity, so $\mathbf{X}^\top\mathbf{X}$ is invertible.

6. A4, A5 jointly imply that $\widehat{\boldsymbol{\beta}}$ is root-n-consistent, i.e. that $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is stochastically bounded as $n\to\infty$, i.e. $O_p(1)$.

To see this; because we assume

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}$$

We have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \sqrt{n}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{u} = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\frac{1}{\sqrt{n}}\mathbf{X}^\top\mathbf{u}$$

which is asymptotically normal mean zero (i.e. not just stochastically bounded) by A4, A5 and Slutsky's theorem.

7. We've seen that $\widehat{\boldsymbol{\beta}}$ has an approximate $N_d(\boldsymbol{\beta}, V[\widehat{\boldsymbol{\beta}}])$-distribution, where

$$V\left[\widehat{\boldsymbol{\beta}}\right] = \frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}, \qquad \text{when n is large}$$

Call $V[\widehat{\boldsymbol{\beta}}]$ the asymptotic variance of $\widehat{\boldsymbol{\beta}}$.

Observe $\{(\mathbf{x}_i^\top, y_i): i=1,\cdots,n\}$.

$$\mathbf{X}=\begin{bmatrix}\mathbf{x}_1^\top\\ \vdots\\ \mathbf{x}_n^\top\end{bmatrix},\qquad \mathbf{y}=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix},\qquad \text{and } \mathbf{u}=\begin{bmatrix}u_1\\ \vdots\\ u_n\end{bmatrix}$$

Assume $y_i=\mathbf{x}_i^\top\boldsymbol{\beta}+u_i$ $(i=1,\cdots,n)$.

$$\text{OLS } \widehat{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$$

We gave assumptions that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}N_d\left(\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\right)$$

i.e. $\widehat{\boldsymbol{\beta}}$ has an approximate $N_d(\boldsymbol{\beta}, n^{-1}\boldsymbol{\Sigma}_{\mathbf{xx}}{}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}{}^{-1})$-distribution when n is large.

Call $\mathbf{V}\left[\widehat{\boldsymbol{\beta}}\right]\equiv\frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$ the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. The formulation of tests/confidence regions regarding $\boldsymbol{\beta}$ requires the (consistent estimation) of $\mathbf{V}\left[\widehat{\boldsymbol{\beta}}\right]$.

Note that

- $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is estimable via $\mathbf{S}_{\mathbf{xx}}\equiv n^{-1}\mathbf{X}^\top\mathbf{X}$ (in view of Assumption 4, hereafter A4).
- White (1980) proposed to estimate S via $\widehat{\mathbf{S}}\equiv\frac{1}{n}\mathbf{X}^\top\widehat{\boldsymbol{\Omega}}\mathbf{X}$ where $\widehat{\boldsymbol{\Omega}}=\text{diag}\left[\hat{u}_i^2: i=1,\cdots,n\right]$ and where $\hat{u}_i=y_i-\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}$.
- Combining $\mathbf{S}_{\mathbf{xx}}$, $\widehat{\mathbf{S}}$ appropriately gives an estimate of $\mathbf{V}\left[\widehat{\boldsymbol{\beta}}\right]$.

$$\widehat{\mathbf{V}}\left[\widehat{\boldsymbol{\beta}}\right]\equiv\frac{1}{n}\mathbf{S}_{\mathbf{xx}}^{-1}\widehat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Omega}}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$$

- If $u_1,\cdots,u_n$ are homoskedastic, so $\boldsymbol{\Omega}\equiv E[\mathbf{u}\mathbf{u}^\top|\mathbf{X}]=\sigma^2\mathbf{I}_n$, then $\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X}=\sigma^2\mathbf{X}^\top\mathbf{X}$ and $\mathbf{S}=\sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}}$, which implies $\mathbf{V}\left[\widehat{\boldsymbol{\beta}}\right]=\frac{1}{n}\sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$. This leads to the <u>default</u> (or "homoskedasticity").

$$\text{OLS Variance Estimator } \widetilde{\mathbf{V}}\left[\widehat{\boldsymbol{\beta}}\right]\equiv\frac{1}{n}s^2\mathbf{S}_{\mathbf{xx}}^{-1}=s^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

$$\text{where } s^2=\frac{1}{n-d}\sum_{i=1}^{n}\hat{u}_i^2$$

(i.e. the square of the <u>regression standard error</u>)

($\widetilde{\mathbf{V}}\left[\widehat{\boldsymbol{\beta}}\right]$ is inconsistent under heteroskedasticity)

How realistic are the assumptions in the previous proposition? Consider six "premitive" assumptions (White, 1980) that imply A1−A5 in the previous proposition.

Suppose $\mathbf{x}_i=[x_{i1}\ x_{2i}\cdots x_{id}]^\top$ for each i.

(A1–A5 are used for the assumptions of Lecture 1. Though the professor used A1–A6 in the class, AA1–AA6 are used for the assumptions of Lecture 2 in order to distinguish them.)

AA1 The observations $[\mathbf{x}_i^\top \quad y_i]$ (i=1,⋯,n) are independent.

- This allows for the data to be generate by stratified random sampling (i.e. split the population into strata, make random draws within each stratum. Typically see oversampling of certain strata).
- AA1 (obviously) rules out most time-series settings, as well as the "spatial" clustering of observations.

AA2 $y_i=\mathbf{x}_i^\top\boldsymbol{\beta}+u_i$ for each i.

- This assumes correctness of the linear model specification. Failure of AA2 implies that OLS's consistent only for the best linear predictor.

AA3 Each vector xi of regressors is possibly random with $E[\mathbf{x}_i\mathbf{x}_i^\top]$ finite. In addition, $E[|x_{ij}x_{ik}|^{1+\delta}]<\infty$ for each $j,k\in\{1,\cdots,d\}$ and some $\delta>0$. Also, $\boldsymbol{\Sigma}_{\mathbf{xx}}\equiv\text{plim}\mathbf{X}^\top\mathbf{X}/n$ exists, is finite and positive-definite with rank d. Finally, $\text{rank}[\mathbf{X}]\equiv d$.

- AA3 allows regressors to be random (almost always the case with observational data).
- If $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are iid then $\boldsymbol{\Sigma}_{\mathbf{xx}}=E[\mathbf{x}_1\mathbf{x}_1^\top]$.
- If $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are inid then AA3 guarantees the existence of $\boldsymbol{\Sigma}_{\mathbf{xx}} \equiv \text{plim}\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ via the Markov Law of Large Numbers.

AA4 $E[u_i|\mathbf{x}_i]=0$ w.p. 1 for each i.

- AA2, AA4 imply that $E[\mathbf{y}|\mathbf{X}]=\mathbf{X}\boldsymbol{\beta}$ so $\boldsymbol{\beta}$ is intrinsically interesting.
- AA4 implies that $\text{Cov}[\mathbf{x}_i,u_i]=\mathbf{0}$ for each i, which suffices for OLS to be consistent.
- AA4 is a weak exogeneity assumption, i.e. u_i represents the total effect on y_i of all determinants of y_i that are uncorrelated with $\mathbf{x}_i$ and on average u_i has no effect on y_i.
- Failure of AA4 is usually referred to as the presence of endogenous regressors. In this case, may not be consistent.
- AA4 does not imply that u_i, $\mathbf{x}_i$ are independent.
- Strong exogeneity is the assumption that $E[u|X]=0$, i.e. $E[ui|xj]=0\ \forall$ i, $j\in\{1, \cdots, n\}$. Under strong exogeneity, OLS is unbiased for $\boldsymbol{\beta}$ (refer to Hayashi Chapter 1).

AA5 $E[u_i^2|\mathbf{x}_i]=\sigma_i^2$ for each $i\in\{1, \cdots, n\}$.

$$\text{Also,} \qquad \boldsymbol{\Omega} \equiv E[\mathbf{u}\mathbf{u}^\top|\mathbf{X}] = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \text{ is positive definite,} \qquad \text{i.e. } \sigma_i > 0\ \forall\ i$$

$$\text{Finally,} \qquad E\left[\left|u_i^2\right|^{1+\delta}\right] < \infty \text{ for some } \delta > 0 \text{ and for all i}$$

- AA5 clearly allows the errors to exhibit conditional heteroskedasticity.

AA6 The matrix $\mathbf{S} \equiv \text{plim}\frac{1}{n}\sum_{i=1}^n u_i^2\mathbf{x}_i\mathbf{x}_i^\top$ exists, is finite, positive definite and has rank d.

$$\text{In addition,} \qquad \exists\ \delta > 0 \text{ s.t. } E\left[\left|u_i^2 x_{ij}x_{ik}\right|^{1+\delta}\right] < \infty \text{ for each i and each } j,k \in \{1,\cdots,d\}$$

- AA6 guarantees that the asymptotic distribution of $\frac{1}{\sqrt{n}}\mathbf{X}^\top\mathbf{u}$ has a finite variance.

Technical aside:
a) Law of Large Numbers: We've seen the following.

Theorem (Kolmogorov LLN) Let $\{\mathbf{x}_i\}$ be iid. Let $\bar{\mathbf{x}}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$. If $E[\|\mathbf{x}_1\|] < \infty$, then

$$\bar{\mathbf{x}}_n \xrightarrow{a.s.} E[\mathbf{x}_n] = E[\mathbf{x}_1]$$

Theorem (Ergodic theorem) Let $\{\mathbf{x}_i\}$ be strictly stationary with an α-mixing base such that $E[\|\mathbf{x}_1\|] < \infty$, then

$$\bar{\mathbf{x}}_n \xrightarrow{a.s.} E[\mathbf{x}_n] = E[\mathbf{x}_1]$$

LLN for independent, but perhaps non-identically distributed processes

Theorem (Markov LLN) Let $\{x_i\}$ be independent random variables such that $Var[x_i]<\infty$ for each i. Let $\mu_i \equiv E[x_i]$, $\sigma_i^2 \equiv Var[x_i]$.

$$\text{If } \sum_{i=1}^{\infty}\frac{E\left[|x_i-\mu_i|^{1+\delta}\right]}{i^{1+\delta}} < \infty \text{ for some } \delta > 0, \qquad \text{then } |\bar{x}_n - E[\bar{x}_n]| \xrightarrow{a.s.} 0$$

b) Central Limit Theorems: We've seen the following.

Theorem (Lindberg–Lévy CLT) Let $\{x_i\}$ be iid random variables with $Var[x_1]<\infty$, then

$$\frac{\bar{x}_n - E[\bar{x}_n]}{\sqrt{Var[\bar{x}_n]}} \xrightarrow{d} N(0,1)$$

Theorem (CLT for Martingale Difference Process) Let $\{x_i\}$ be strictly stationary, scalar valued MDP with α-mixing base. Suppose $E[x_1^2]<\infty$. Let $\sigma^2 \equiv E[x_1^2]$, then

$$\frac{\bar{x}_n}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Now present a CLT for independent but not necessarily identically distributed processes.

Theorem (Lyapunov CLT) Let $\{x_i\}$ be independent random variables each with $Var[x_i]<\infty$. Let $\mu_i \equiv E[x_i]$, $\sigma^2 \equiv Var[x_i]$. Then

$$\text{if } \exists\ \delta > 0 \text{ such that } \lim_{n\to\infty}\frac{\sum_{i=1}^{n}E\left[|x_i-\mu_i|^{2+\delta}\right]}{\left(\sum_{i=1}^{n}\sigma_i^2\right)^{\frac{2+\delta}{2}}} = 0, \qquad \text{then } \frac{\bar{x}_n - E[\bar{x}_n]}{\sqrt{Var[\bar{x}_n]}} \xrightarrow{d} N(0,1)$$

c) Weighted Least Squares (WLS)

Rule of thumb: The existence of robust standard errors (e.g. heteroscedasticity-robust standard errors for OLS) usually implies that the existence of a more efficient estimator that avoids the need for robust standard errors. In the case of linear regression under AA1–AA6 generalized least squares (GLS) is more efficient than just OLS.

c) Weighted Least Squares (WLS) methodology

Suppose that the error variance $E[\mathbf{u}\mathbf{u}^\top|\mathbf{X}]=\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}\neq\sigma^2\mathbf{I}_n$ and $\boldsymbol{\Omega}$ is nonsingular.

$$\text{Then } \boldsymbol{\Omega}^{-1/2} \text{ exists,} \qquad \text{and } \mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u} \Leftrightarrow \boldsymbol{\Omega}^{-1/2}\mathbf{y}=\boldsymbol{\Omega}^{-1/2}\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\Omega}^{-1/2}\mathbf{u}, \qquad \cdots(*)$$

Where

$$\mathrm{Var}\left[\boldsymbol{\Omega}^{-1/2}\mathbf{u}\middle|\mathbf{X}\right]=\mathrm{E}\left[\boldsymbol{\Omega}^{-1/2}\mathbf{u}\mathbf{u}^\top\boldsymbol{\Omega}^{-1/2}\middle|\mathbf{X}\right]=\mathbf{I}_n$$

i.e. the errors in (*) are homoskedastic, and satisfy $E[\boldsymbol{\Omega}^{-1/2}\mathbf{u}|\mathbf{X}]=\mathbf{0}$ if $E[\mathbf{u}|\mathbf{X}]=\mathbf{0}$ is assumed. So by Gauss–Markov theorem, OLS applied to (*) is BLUE. The OLS estimator of $\boldsymbol{\beta}$ in (*) is the generalized least squares (GLS) estimator of $\boldsymbol{\beta}$. In particular,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}=(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{y}$$

Naturally, $\boldsymbol{\Omega}$ is unknown, so $\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is infeasible. Instead, specify a parametric model for $\boldsymbol{\Omega}$, i.e. $\boldsymbol{\Omega}=\boldsymbol{\Omega}(\boldsymbol{\gamma})$, where $\boldsymbol{\Omega}(\cdot)$ is a known function and $\boldsymbol{\gamma}$ is a finite-dimensional vector. If the model $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ is correctly specified and $\hat{\boldsymbol{\gamma}}$ is consistent for $\boldsymbol{\gamma}$, then $\widehat{\boldsymbol{\Omega}}=\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$ is consistent for $\boldsymbol{\Omega}$.

e.g. suppose <u>multiplicative heteroskedasticity</u>, i.e.

$$\mathrm{E}\left[\mathrm{u}_i^2\middle|\mathbf{x}_i\right]=\exp\left(\mathbf{z}_i^\top\boldsymbol{\gamma}\right)$$

Where $\mathbf{z}_i$ is a subvector of $\mathbf{x}_i$. Then $\boldsymbol{\gamma}$ is consistently estimated by OLS applied to a regression of $\log\hat{\mathrm{u}}_i^2$ (where $\hat{\mathrm{u}}_i^2$ is the i-th OLS residual) on $\mathbf{z}_i$: or by nonlinear least squares applied to a regression of $\hat{\mathrm{u}}_i^2$ on $\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})$.

Given an estimate $\widehat{\boldsymbol{\Omega}}=\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$ of $\boldsymbol{\Omega}$ the feasible GLS (FGLS) estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}}=\left(\mathbf{X}^\top\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{y}$$

Under assumptions <u>AA1</u>–<u>AA6</u>, if the model $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ is correctly specified and $\hat{\boldsymbol{\gamma}}$ is consistent, then

$$\sqrt{\mathrm{n}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}}-\boldsymbol{\beta}\right)\xrightarrow{\mathrm{d}}\mathrm{N}_{\mathrm{d}}\left(0,\left(\operatorname*{plim}_{\mathrm{n}\to\infty}\frac{1}{\mathrm{n}}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\right), \qquad \cdots(**)$$

i.e. the same asymptotic variance as that of $\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}$, i.e. for large n, $\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$ is as efficient as $\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}$.
Note that the FGLS estimated asymptotic variance matrix is

$$\left(X^\top\widehat{\Omega}^{-1}X\right)^{-1}$$

Compare to

$$(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\widetilde{\boldsymbol{\Omega}}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}, \qquad \text{where } \widetilde{\boldsymbol{\Omega}}=\mathrm{diag}\left[\hat{\mathrm{u}}_i^2\right]$$

i.e. the estimated heteroscedasticity-robust asymptotic variance matrix for OLS (the FGLS estimated asymptotic variance implies standard errors that are no greater the corresponding OLS standard errors.). Naturally, (**) depends on $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ being correctly specified. If $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ is incorrectly specified, then $\widehat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$ is still consistent, although $\left(\mathbf{X}^\top\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}$ will be inconsistent for the asymptotic variance.

A solution to possible misspecification of $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ involves a working variance matrix. Let $\boldsymbol{\Sigma}\equiv\boldsymbol{\Sigma}(\boldsymbol{\gamma})$ be a working variance matrix that may not equal the true variance matrix $\boldsymbol{\Omega}\equiv E[\mathbf{u}\mathbf{u}^\top|\mathbf{X}]$. Let $\widehat{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma}(\boldsymbol{\gamma})$ where $\hat{\boldsymbol{\gamma}}$ estimates $\boldsymbol{\gamma}$. Then use WLS to $\boldsymbol{\beta}$ (with weighting matrix $\widehat{\boldsymbol{\Sigma}}^{-1}$), i.e.

$$\widehat{\boldsymbol{\beta}}_{\text{WLS}} = \left(\mathbf{X}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

The estimated asymptotic variance matrix in this case is the heteroskedasticity-robust estimated asymptotic variance for OLS applied to

$$\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{y} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{u}$$

i.e.

$$\begin{aligned}\text{Var}\left[\widehat{\boldsymbol{\beta}}_{\text{WLS}}\right] &= \left(\mathbf{X}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\left(\mathbf{X}^\top\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\\ \text{where } \widetilde{\boldsymbol{\Omega}} &= \text{diag}[(\hat{u}_i^*)^2]\\ \hat{u}_i^* &= y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_{\text{WLS}}\end{aligned}$$

In practice, working variance matrices typically involve only simple models for heteroskedasticity.

e.g. Numerical Example

$$\begin{aligned}\text{Suppose } y &= 1 + x + u\\ \text{where } u &= xv\\ x &\sim N(0,5^2)\\ v &\sim N(0,2^2), \qquad \text{is independent of } x\end{aligned}$$

Then $\text{Var}[u|x]=x^2\text{Var}[v]=4x^2$, i.e. heteroskedastic. So OLS is inefficient, GLS is relatively efficient, while the efficiency of a working variance matrix approach is typically somewhere in between.

Consider n=100 simulated (iid) observations from this model.

- OLS generates a slope estimate of $\hat{\beta}_2 = 1.0495$. The heteroskedasticity-robust standard error is .3207 (correct). Compare with the homoscedasticity-only standard error .2016 (incorrect).

  In fact, in this model, the heteroskedasticity-robust standard error for the slope is $\sqrt{12/n}$ when n is large; while the homoscedasticity only standard error for large n is $\sqrt{4/n}$.
- GLS for this model involves OLS applied to the transformed regression in which both sides of the original model are divided by |x|. Then the errors in the transformed model are homoskedastic; so the heteroskedasticity-robust and homoscedasticity-only standard errors should be close.

  In fact, we see .208 (heteroskedasticity-robust) and .209 (homoscedasticity-only).
- Examples of the working variance matrix approach. Divided both sides of the original model by $\sqrt{|x|}$ (not the correct weights). The corresponding WLS estimator has heteroskedasticity-robust standard errors of .232.
- Summary: Compare standard errors for slope estimates in this model of .3207 (OLS), .232 (WLS based on a working variance matrix), .209 (GLS).
- Most applied work in practice does not involve weighted least squares techniques, perhaps for reasons of convenience.

d) Misspecification of AA1–AA6

Suppose at least one of AA1–AA6 is false.

(1) Then OLS might be inconsistent.

Recall: If $\mathbf{y}=\mathbf{X\beta}+\mathbf{u}$ and if $\frac{1}{n}\mathbf{X}^\top\mathbf{u} \xrightarrow{p} \mathbf{0}$ then $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is consistent.

In fact, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{u}$ is consistent only if both $y=X\beta+u$ and $\operatorname{plim}\frac{1}{n}\mathbf{X}^\top\mathbf{u} = \mathbf{0}$ are true. So $\widehat{\boldsymbol{\beta}}$ is inconsistent if $\mathbf{y}=\mathbf{X\beta}+\mathbf{u}$ is false or if the regressors are asymptotically correlated with the error.

(2) If $y=X\beta+u$ is false, then OLS might not be useful as a predictor of y, because in this case

$$\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X\beta}$$

is false.

e.g. Suppose $y_i=g(\mathbf{x}_i)+v_i$ for some nonlinear function $g(\cdot)$ and where $\mathrm{E}[v_i|\mathbf{x}_i]=0$. Does OLS have a meaningful interpretation in this case? White (1980) showed that $\mathrm{E}[y_i|\mathbf{x}_i]=g(\mathbf{x}_i)$ for some nonlinear function $g(\cdot)$, then OLS estimates $\boldsymbol{\beta}\in\mathbb{R}^d$ such that

$$\boldsymbol{\beta} = \underset{\mathbf{b}\in\mathbb{R}^d}{\operatorname{argmin}}\,\mathrm{E}[(g(\mathbf{x}) - \mathbf{x}^\top\mathbf{b})^2]$$

i.e. OLS generates a best linear predictor of y with respect to mean-squared prediction error. (Note: OLS cannot have a causal interpretation in this case.)

(3) If $\operatorname{plim}\frac{1}{n}\mathbf{X}^\top\mathbf{u} = \mathbf{c} \neq 0$, i.e. the error and regressors are asymptotically correlated, then say that at least one element of $\mathbf{x}$ is endogenous.

In this case, $\mathrm{E}[u_i|\mathbf{x}_i]$ cannot be 0 and OLS is inconsistent.

Corresponding SAS Code

```
resetline;
ods html close;
ods graphics off;
ods listing;

data _01;
   do M=1 to 5;
   do N=1 to 100;
      X=5*rannor(1);
      V=2*rannor(2);
      U=X*V;
      Y=1+X+U;
      TX1=1/abs(X);
      TX2=X/abs(X);
      TY=Y/abs(X);
      UX1=1/sqrt(abs(X));
      UX2=X/sqrt(abs(X));
      UY=Y/sqrt(abs(X));
      output;
   end;
   end;
run;

proc reg;
   model Y=X/white;
   model TY=TX1 TX2/noint white;
   model UY=UX1 UX2/noint white;
   by M;
run;

quit;
```

(3) Suppose at least one of the chose regressors is endogenous.

Endogeneity of a regressor is usually defined to exist iff that regressor is not correlated with the error term.

This will be the case if AA4 fails, i.e. if $E[u_i|\mathbf{x}_i]=0$ w.p. 1 is not true.

If a regressor is correlated with the error term (for whatever reason), then OLS is inconsistent.

(4) A classic cause of regressor endogeneity is due to one or more omitted variables.

e.g. Suppose $y_i=\mathbf{x}_i^\top\boldsymbol{\beta}+z_i\alpha+v_i$ where $[\mathbf{x}_i^\top\ y_i\ z_i]^\top$ are observed for all i and v_i is unobserved with $E[v_i|\mathbf{x}_i,z_i]=0$ w.p. 1. Then OLS estimator of $[\boldsymbol{\beta}^\top\ \alpha]^\top$ is consistent.

Now suppose y_i is regressed on $\mathbf{x}_i$ alone (so, z_i is omitted). In this case we are really dealing with the model

$$y_i = \mathbf{x}_i^\top\boldsymbol{\beta} + u_i, \qquad \text{where } u_i = z_i\alpha + v_i \text{ is unobserved}$$

If $Cov[\mathbf{x}_i,z_i]\neq 0$, then the OLS estimator of $\boldsymbol{\beta}$ is inconsistent, with the inconsistency in this case called omitted variable bias (although it would be more accurate to call it omitted variable inconsistency).

e.g. Suppose

$$\begin{aligned} y_i &= \text{individual log average hourly earnings} \\ x_i &= \text{vector of individual attributes (including years of schooling)} \\ z_i &= \text{unobservable individual ability (i.e. IQ)} \end{aligned}$$

for a sample of n workers.

Generally assume z_i and years of schooling to be positively correlated; z_i and y_i to be positively correlated. If true, this assumption would make the OLS estimate of the coefficient vector attached to $\mathbf{x}_i$ in a regression of $\mathbf{y}_i$ on $\mathbf{x}_i$ inconsistent.

One can sometimes pin down the direction of inconsistency due to an omitted variable.

Let $\mathbf{z}\equiv[z_1\ \cdots\ z_n]^\top$ denote the omitted observations on z_i. Let $\mathbf{v}\equiv[v_1\ \cdots\ v_n]^\top$. Then the model for $\mathbf{y}$ is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\alpha + \mathbf{v}$$

and the OLS "estimate" of $\boldsymbol{\beta}$ in a regression of $\mathbf{y}$ on $\mathbf{X}$ alone is

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \mathbf{z}\alpha + \mathbf{v}) \\ &= \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{z}\alpha + \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{v}, \qquad \text{where } \underbrace{\frac{1}{n}\mathbf{X}^\top\mathbf{v} \xrightarrow{p} \mathbf{0}}_{\text{since } Cov[\mathbf{x}_i,v_i]=\mathbf{0} \text{ is assumed}} \end{aligned}$$

If $\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{z} \xrightarrow{p} \boldsymbol{\delta} \neq \mathbf{0}$, then $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} + \boldsymbol{\delta}\alpha$, which shows that the direction of inconsistency in $\widehat{\boldsymbol{\beta}}$ depends on the sign of each component of $\boldsymbol{\delta}\alpha$.

e.g. Suppose for a sample of n workers

$$y_i \equiv \text{individual log average hourly earnings}$$
$$x_i \equiv \text{individual years of schooling}$$
$$z_i \equiv \text{unobserved individual ability}$$

Then, if $E[v_i|x_i,z_i]=0$, then

$$\beta = \text{(causal) return to an additional year of schooling}$$
$$\alpha = \text{(causal) return to ability}$$

In this example, expect $\delta_j\alpha$ to be positive, so OLS applied to a regression of y_i on x_i will overestimate β in the limit.

(5) Suppose we have parameter heterogeneity. Suppose that instead of

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i$$

We have instead

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_i + u_i, \qquad \cdots (*)$$

when $E[u_i|\mathbf{x}_i]=0$.

In this case, the minimum MSPE (mean squared prediction error) predictor of y_i is

$$E[y_i|\mathbf{x}_i] = \mathbf{x}_i^\top \boldsymbol{\beta}_i$$

which varies across i.

A random coefficients model is one where $\boldsymbol{\beta}_i$ are iid random vectors independent of $\mathbf{x}_i$.

Let $\boldsymbol{\beta}=E[\boldsymbol{\beta}_i]$. Then rewrite (*) as

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \left[u_i + \mathbf{x}_i^\top (\boldsymbol{\beta}_i - \boldsymbol{\beta})\right], \qquad \cdots (**)$$

If conditions AA1, AA3–AA6 hold, then OLS will be consistent for $\boldsymbol{\beta}$.

Special case of (**): A random effects model is one where $\mathbf{x}_i=[1\ \tilde{\mathbf{x}}_i^\top]^\top$ and only the first component of $\boldsymbol{\beta}_i$ (i.e. the intercept) is random iid and independent of $\tilde{\mathbf{x}}_i$.

Also, OLS applied to (**) is inconsistent for $\boldsymbol{\beta}$ if $\boldsymbol{\beta}_i$ is correlated with $\mathbf{x}_i$.

A leading special case of this is the fixed effects model, i.e. $\mathbf{x}_i=[1\ \tilde{\mathbf{x}}_i^\top]^\top$ and only the first component (i.e. the intercept) of $\boldsymbol{\beta}_i$ is random, but correlated with $\tilde{\mathbf{x}}_i$ (need panel data to handle this).

e) Instrumental Variables

Suppose $y_i=\mathbf{x}_i^\top\boldsymbol{\beta}+u_i$ where (unfortunately)

$$\mathrm{Cov}[\mathbf{x}_i, u_i] \neq \mathbf{0}$$

i.e. $\mathbf{x}_i$ contains an endogenous component.

In this case, OLS is inconsistent. Can estimate $\boldsymbol{\beta}$ consistently if there exists a random m-vector of instruments $\mathbf{z}_i$ where m≥d where

(1) $E[\mathbf{z}_i u_i]=\mathbf{0}$	(otherwise $\mathbf{z}_i$ is invalid)
(2) $E[\mathbf{z}_i \mathbf{x}_i^\top]\neq\mathbf{0}$	(otherwise $\mathbf{z}_i$ is irrelevant)
(3) $\|E[\mathbf{z}_i \mathbf{x}_i^\top]\|$ is large (rather than small)	(otherwise $\mathbf{z}_i$ is weak)

Properties (1), (2) (as discussed in Hayashi Ch. 3) are necessary for consistency. Property (3) has to do with controlling small-sample bias (more later).

The condition m≥d is the order condition for identification of $\boldsymbol{\beta}$.

Say that $\boldsymbol{\beta}$ is identified by $\mathbf{z}_i$ if m≥d

overidentified	m>d (GMM, 2SLS)
just identified	m=d (IV regression)
not identified	m<d

$$\text{Let } \mathbf{Z} \equiv [\mathbf{z}_1 \quad \cdots \quad \mathbf{z}_n]^\top, \qquad \text{suppose } m = d$$

Then, the instrumental variable (IV) estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{IV} = (\mathbf{Z}^\top\mathbf{X})^{-1}\mathbf{Z}^\top\mathbf{y}$$

This is consistent because

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{IV} &= (\mathbf{Z}^\top\mathbf{X})^{-1}\mathbf{Z}^\top\mathbf{y}\\ &= (\mathbf{Z}^\top\mathbf{X})^{-1}\mathbf{Z}^\top(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})\\ &= \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{Z}^\top\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{Z}^\top\mathbf{u}\end{aligned}$$

Where

$$\frac{\mathbf{Z}^\top\mathbf{X}}{n} \xrightarrow{p} E[\mathbf{z}_1\mathbf{x}_1^\top] = \boldsymbol{\Sigma}_{\mathbf{zx}} \neq 0, \qquad \text{if } \mathbf{z}_1 \text{ is relevant}$$

$$\text{and } \frac{\mathbf{Z}^\top\mathbf{u}}{n} \xrightarrow{p} E[\mathbf{z}_1 u_1] = \mathbf{0}, \qquad \text{if } \mathbf{z}_1 \text{ is valid}$$

Can show that (under AA5, where $\boldsymbol{\Omega}\equiv E[\mathbf{u}\mathbf{u}^\top|\mathbf{X}]=\mathrm{diag}[\sigma_i^2: i=1,\cdots,n]$ and AA6) $\widehat{\boldsymbol{\beta}}_{IV}$ is $\sqrt{n}$-consistent and asymptotically Normal with asymptotic variance consistency estimated as

$$\widehat{\mathbf{V}}\left[\widehat{\boldsymbol{\beta}}_{IV}\right] \equiv (\mathbf{Z}^\top\mathbf{X})^{-1}\mathbf{Z}^\top\widehat{\boldsymbol{\Omega}}\mathbf{Z}(\mathbf{Z}^\top\mathbf{X})^{-1}$$

Where $\widehat{\boldsymbol{\Omega}} = \mathrm{diag}\left[\hat{u}_i^2 : i = 1, \cdots, n\right]$, $\hat{u}_i = y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_{IV}$.

In practice: $\widehat{\boldsymbol{\beta}}_{IV}$ is often associated with large standard errors.

Now suppose m≥d (this allows for m>d). Consider the two-stage least squares (2SLS) estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{2SLS} &= [\mathbf{X}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{X}]^{-1}\mathbf{X}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y} \\ &= (\mathbf{X}^\top\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{P}_\mathbf{Z}\mathbf{y}\end{aligned}$$

Where $\mathbf{P}_\mathbf{Z}=\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$ is the projection matrix onto the column space of $\mathbf{Z}$. $\mathbf{P}_\mathbf{Z}$ is symmetric and idempotent, so

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\widehat{\mathbf{X}}^\top\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}\mathbf{y}, \qquad \text{where } \widehat{\mathbf{X}} = \mathbf{P}_\mathbf{Z}\mathbf{X}$$

Can show that $\widehat{\boldsymbol{\beta}}_{2SLS}$ is $\sqrt{n}$-consistent and asymptotically Normal with asymptotic variance consistently estimated by

$$\widehat{\mathbf{V}}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right] = n(\mathbf{X}^\top\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\hat{\mathbf{S}}_n(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top(\mathbf{X}^\top\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}$$

$$\text{where } \hat{\mathbf{S}}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\mathbf{z}_i\mathbf{z}_i^\top$$

$$\hat{u}_i = y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_{2SLS}, \qquad (\text{note: need conditions AA5 and AA6})$$

e.g. Suppose

$$\begin{aligned} &y = .5x + u, \qquad \text{where } x = z + v \\ &z \sim N(2,1^2) \\ &\begin{bmatrix}u\\v\end{bmatrix} \sim N_2\left(\mathbf{0}, \begin{bmatrix}1^2 & .8\\ .8 & 1^2\end{bmatrix}\right) \\ &z \text{ is independent of } [u \quad v]^\top\end{aligned}$$

We have a problem because

$$\text{Cov}[x,u] = \text{Cov}[z,u] + \text{Cov}[v,u] = \text{Cov}[v,u] = .8$$

So, OLS applied to a regression of y on x is inconsistent for .5.

On the other hand, IV estimation with instrument z is consistent because z is valid (E[zu]=Cov[z,u]+E[z]E[u]=0+0=0) and relevant (E[zx]=Cov[z,x]+E[z]E[x]=1+4=5≠0).

In addition, certain functions of z (e.g. z^3) can be used as instruments. Consider n=10,000 simulated samples from this model.

The OLS estimate (robust standard error) is .902 (.006).

The IV estimate using z is .510 (.010).

The 2SLS estimate using z is .510 (.014).

The IV estimate using z^3 is .509 (.012).

Obviously the IV and 2SLS standard errors are larger than the OLS standard error; in this case this is compensated by much smaller bias.

Corresponding SAS Code

```
resetline;
ods html close;
ods graphics off;
ods listing;

/*
proc iml;
    VAR={1 .8,.8 1};
    HALF=root(VAR);
    print HALF;
quit;
*/

data _01;
    do N=1 to 10000;
        Z=2+rannor(1);
        Z3=Z**3;
        U=rannor(2);
        V=.8*U+.6*rannor(3);
        X=Z+V;
        Y=.5*X+U;
        output;
    end;
run;

proc reg;
    model Y=X/white;
run;

proc iml;
    use _01;
    read all var{Y} into Y;
    read all var{X} into X;
    X=j(nrow(X),1,1)||X;
    read all var{Z} into Z;
    Z=j(nrow(Z),1,1)||Z;
    IV=inv(Z`*X)*Z`*Y;
    E1=Y-X*IV;
    VIV=inv(Z`*X)*Z`*diag(E1##2)*Z*inv(Z`*X);
    SEIV=sqrt(vecdiag(VIV));
    PZ=Z*inv(Z`*Z)*Z`;
    TSLS=inv(X`*PZ*X)*X`*PZ*Y;
    E2=Y-X*TSLS;
    S=0;
    do N=1 to 10000;
        S=S+E2[N]**2*Z[N,]`*Z[N,];
    end;
    VTSLS=inv(X`*PZ*X)*X`*Z*inv(Z`*Z)*S*inv(Z`*Z)*Z`*X*inv(X`*PZ*X);
    SETSLS=sqrt(vecdiag(VTSLS));
    read all var{Z3} into Z3;
    Z3=j(nrow(Z3),1,1)||Z3;
    IV3=inv(Z3`*X)*Z3`*Y;
    E3=Y-X*IV3;
    VIV3=inv(Z3`*X)*Z3`*diag(E1##2)*Z3*inv(Z3`*X);
    SEIV3=sqrt(vecdiag(VIV3));
    print IV[format=8.4]
        SEIV[format=8.4],
        TSLS[format=8.4]
        SETSLS[format=8.4],
        IV3[format=8.4]
        SEIV3[format=8.4];
quit;

proc model;
    Y=B0+B1*X;
    instruments Z;
    fit Y/2sls hccme=0;
run;

quit;
```

Recall: IV/2SLS tends to have smaller bias than OLS, but at the cost of larger standard errors.

Today: If the chosen instruments are weak, then IV/2SLS can be as biased (or even more biased) than OLS while containing to be relatively inefficient.

Weak instruments are associated with

$$\left\|\mathrm{E}\left[\mathbf{z}_i\mathbf{x}_i^\top\right]\right\|$$

being "small."
(Instrument weakness causes finite sample bias)
e.g. Consider

$$y = \beta_1 x_1 + \mathbf{x}_2^\top \boldsymbol{\beta}_2 + u, \qquad \cdots (*)$$
$$\text{where } \mathrm{Cov}[x_1, u] \neq 0$$
$$\text{and } \mathrm{Cov}[\mathbf{x}_2, u] = \mathbf{0}$$

Assume that the instrument $\mathbf{z}$ includes $\mathbf{x}_2$ and at least one other exogenous variable.
Intuition: $\mathbf{z}$ is weak if the R^2 from a regression of x_1 on $\mathbf{z}$ is "low," i.e. $\mathbf{z}$ are not strongly correlated.
But: Really only need x_1 to be strongly correlated with the elements of $\mathbf{z}$ not in $\mathbf{x}_2$.

Bound, Jaeger and Baker (1995, JASA): Proposal for a partial R^2, i.e. the R^2 from the regression

$$x_1 - \hat{x}_1 = (\mathbf{z} - \hat{\mathbf{z}})^\top \boldsymbol{\gamma} + v_1, \qquad \cdots (**)$$

Where $\hat{x}_1$ is the LS predicted value from a regression of x_1 on $\mathbf{x}_2$ and $\hat{\mathbf{z}}$ is the vector of LS predicted values from regressions of each element of $\mathbf{z}$ on $\mathbf{x}_2$.

Case with more than one endogenous regressor, i.e. the case where $\mathbf{x}_2$ in (*) contains an endogenous regressor.

In this case, can diagnose instrument weakness via the proposal of Shea (1997, RES): Use the R^2 from the following regression

$$x_1 - \hat{x}_1 = \left(\tilde{x}_1 - \tilde{\tilde{x}}_1\right)\eta + v_2$$

Where $\hat{x}_1$ is the LS predicted value from a regression of x_1 on $\mathbf{x}_2$ and $\tilde{x}_1$ is the LS predicted value from a regression of x_1 on $\mathbf{z}$; $\tilde{\tilde{x}}_1$ is the LS predicted value from a regression of $\tilde{x}_1$ on $\tilde{\mathbf{x}}_2$; where $\tilde{\mathbf{x}}_2$ is the LS predicted value from a regression of $\mathbf{x}_2$ on $\mathbf{z}$.

Another method of diagnosing instrument weakness uses the F-statistic for the hypothesis that all coefficients are zero in a regression of an endogenous regressor on instruments.
e.g. Go back to (*). Suppose $\mathbf{z}=[\mathbf{z}_1^\top\ \mathbf{z}_2^\top]^\top$ where $\mathbf{z}_1$ are exogenous variables not in $\mathbf{x}_2$. Then can use the F-statistic for H_0: $\boldsymbol{\pi}_1=\mathbf{0}$ in the "reduced-form" regression

$$x_1 = \mathbf{z}_1^\top \boldsymbol{\pi}_1 + \mathbf{z}_2^\top \boldsymbol{\pi}_2 + v_3$$

Rule of thumb: $\mathbf{z}$ is "very weak" if this F-statistic is ≤ 5; $\mathbf{z}$ is "potentially weak" if the F-statistic is (5,10).

Next: IV/2SLS is inconsistent if the chosen instrument is invalid, i.e. if **z** is correlated with u.

Deciding on questions of instrument validity is usually subjective, but note that instrument validity is closely tied to a model classification.

e.g.

$$\begin{aligned} y &= x_1\beta_1 + x_2\beta_2 + u \\ \text{where } \mathrm{Cov}[x_1, u] &= 0 \\ \text{but } \mathrm{Cov}[x_2, u] &\neq 0 \\ \text{and } \mathrm{E}[x_1] = \mathrm{E}[x_2] &= 0 \end{aligned}$$

Then OLS is inconsistent. "Plausible" that x_1^2 can serve as an instrument. This may not work, however, because it assumes that $E[y|x_1,x_2]$ is linear but not quadratic in x_1.

Better not to tie instrument validity to a restriction on the function form of $E[y|x_1,x_2]$.

Next: The bias due to very slight instrument endogeneity is much worse if the instrument is also weak.

e.g.

$$y = x\beta + u$$

Suppose z is a candidate instrument. Assume E[x]=E[z]=E[u]=0 and Var[x]=Var[z]=Var[u]=1^2. Suppose we observe (x_i,y_i,z_i) $(i=1,\cdots,n)$ iid. We have

$$\begin{aligned} \hat{\beta}_{IV} &= \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i x_i}, \qquad \hat{\beta}_{OLS} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \\ &\xrightarrow{p} \frac{\mathrm{E}[z_1 y_1]}{\mathrm{E}[z_1 x_1]} = \beta + \frac{\mathrm{E}[z_1 u_1]}{\mathrm{E}[z_1 x_1]} \end{aligned}$$

and

$$\hat{\beta}_{OLS} \xrightarrow{p} \frac{\mathrm{E}[x_1 y_1]}{\mathrm{E}[x_1^2]} = \mathrm{E}[x_1 y_1] = \beta + \mathrm{E}[x_1 u_1]$$

So

$$\frac{\operatorname{plim} \hat{\beta}_{IV} - \beta}{\operatorname{plim} \hat{\beta}_{OLS} - \beta} = \frac{\mathrm{E}[z_1 u_1]}{\mathrm{E}[z_1 x_1]} \frac{1}{\mathrm{E}[x_1 u_1]} = \frac{\mathrm{Corr}[z_1, u_1]}{\mathrm{Corr}[x_1, u_1]} \frac{1}{\mathrm{Corr}[z_1, x_1]}$$

i.e. if z is invalid and weak, then $\hat{\beta}_{IV}$ will be even more inconsistent for β than $\hat{\beta}_{OLS}$. Then

$$\begin{aligned} &\left|\operatorname{plim} \hat{\beta}_{IV} - \beta\right| > \left|\operatorname{plim} \hat{\beta}_{OLS} - \beta\right| \\ &\text{if } |\mathrm{Corr}[z_1, u_1]| > .10|\mathrm{Corr}[x_1, u_1]| \end{aligned}$$

This shows that only a slight amount of instrument endogeneity suffices to make a weak instrument generate a large degree of inconsistency in the corresponding IV estimate.

Also: IV/2SLS become even more imprecise when the chosen instruments are weak.

e.g.

$$y = x\beta + u, \qquad \text{Corr}[x, u] \neq 0$$

Suppose z is a valid instrument, and we observe (x_i,y_i,z_i) (i=1,⋯,n) iid. Assume that the corresponding errors $u_1,\cdots,u_n$ are iid mean zero with variance σ^2.

In this case, can use a "homoscedasticity-only" estimator of the $\hat{\beta}_{IV}$ asymptotic variance

$$\begin{aligned}\hat{V}[\hat{\beta}_{IV}] &= s^2(\mathbf{x}^\top\mathbf{z})^{-1}\mathbf{z}^\top\mathbf{z}(\mathbf{z}^\top\mathbf{x})^{-1} \\ &= \frac{s^2}{\mathbf{x}^\top\mathbf{x}}\frac{(\mathbf{z}^\top\mathbf{z})(\mathbf{x}^\top\mathbf{x})}{(\mathbf{z}^\top\mathbf{x})^2} = \hat{V}[\hat{\beta}_{OLS}]\frac{1}{R^2_{x|z}} + o_p(1)\end{aligned}$$

where $R_{x|z}^2$ is the squared sample correlation between x and z and where $\hat{V}[\hat{\beta}_{OLS}]$ is the homoscedasticity-only OLS asymptotic variance estimator.

Therefore, the relative inefficiency of $\hat{\beta}_{IV}$ (relative to $\hat{\beta}_{OLS}$) is made worse when $R_{x|z}^2$ is small.

e.g. Suppose

$$y = x_1\beta_1 + \mathbf{x}_2^\top\boldsymbol{\beta}_2 + u$$

where $\mathbf{x}_1$ is endogenous and $\mathbf{x}_2$ is exogenous.

Suppose $\mathbf{z}$ is a valid instrument that includes $\mathbf{x}_2$ and at least one additional exogenous variable.

Let R_p^2 be the R^2 from the regression

$$x_1 - \hat{x}_1 = (\mathbf{z} - \hat{\mathbf{z}})^\top\gamma + v$$

where $\hat{x}_1$ is the LS fitted value of x_1 from a regression of x_1 on $\mathbf{x}_2$, and $\hat{\mathbf{z}}$ is the vector of LS fitted values from regressions of each component of $\mathbf{z}$ on $\mathbf{x}_2$.

Let $\hat{\beta}_{1,2SLS}$ be the 2SLS estimator of β_1 using $\mathbf{z}$.

Let $\hat{\beta}_{1,OLS}$ be the OLS estimator of β_1. Can show

$$SE[\hat{\beta}_{1,2SLS}] = \frac{SE[\hat{\beta}_{1,OLS}]}{R_p}$$

So, $SE[\hat{\beta}_{1,2SLS}]$ will generally be larger than $SE[\hat{\beta}_{1,OLS}]$.

Also, instrumental weakness affects IV/2SLS standard errors only for estimated coefficients attached to endogenous variables.

Instrument weakness inflates IV/2SLS standard errors via R_p^2 (i.e. partial R^2) and not through the R^2 on the first-stage regression.

(So, 1) compare SE_{2SLS} and SE_{OLS} and 2) if there is a significant gap between them, have to think about instrument weakness)

(This is because that correlation between an endogenous regressor and exogenous regressors is unimportant; want correlation between an endogenous regressor and instruments that are not also exogenous regressors)

2SLS standard errors on estimated coefficients attached to endogenous regressors that are much larger than the corresponding OLS standard errors indicate instrument weakness.

Relationship between instrument weakness and the finite-sample bias of IV or 2SLS estimators

Suppose $\mathbf{Z}$ is an (n×m)-matrix of valid instruments.

Suppose m=d=# of regressors.

Simulations have shown that

$$\widehat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\mathbf{Z}^{\top}\mathbf{X})^{-1}\mathbf{Z}^{\top}\mathbf{y}$$

can be highly biased in "small" samples (even for n≥100,000) despite being consistent.

In addition, the estimated asymptotic standard errors of $\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}$ can be highly biased even in "small" samples.

First consider that $\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}\right]$ is basically intractable.

$$\begin{aligned}\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}\right] &= \mathrm{E}[(\mathbf{Z}^{\top}\mathbf{X})^{-1}\mathbf{Z}^{\top}\mathbf{y}] \\ &= \boldsymbol{\beta} + \mathrm{E}[(\mathbf{Z}^{\top}\mathbf{X})^{-1}\mathbf{Z}^{\top}\mathbf{u}] \\ &= \boldsymbol{\beta} + \mathrm{E}\left[(\mathbf{Z}^{\top}\mathbf{X})^{-1}\mathbf{Z}^{\top}\mathrm{E}[\mathbf{u}|\mathbf{Z},\mathbf{X}]\right]\end{aligned}$$

Therefore, $\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}$ is unbiased if E[**u**|**Z**,**X**]=**0** w.p. 1.

But, if E[**u**|**Z**,**X**]=**0** a.s. then E[**u**|**X**]=**0** a.s., which implies that **X** is exogenous, i.e. $\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}$ is unnecessary. Therefore, the assumption that

$$\mathrm{E}[\mathbf{u}|\mathbf{Z},\mathbf{X}] = \mathbf{0}, \qquad \text{w.p. 1}$$

is much too strong in this case. Therefore, we expect that

$$\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}\right] \neq \boldsymbol{\beta}$$

in general.

Research into the value of

$$\left|\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}\right] - \boldsymbol{\beta}\right|, \qquad \text{or } \left|\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{2SLS}}\right] - \boldsymbol{\beta}\right|$$

Take the case d=1, i.e. just one regressor, which is endogenous. Consider the reduced-form regression

$$\mathrm{x} = \mathbf{z}^{\top}\boldsymbol{\pi} + \mathrm{v}, \qquad \text{where } \mathrm{E}[\mathrm{v}] = 0 \text{ and } \mathrm{Var}[\mathrm{v}] = \sigma_{\mathrm{v}}^2$$

Define the concentration parameter

$$\tau^2 = \frac{\boldsymbol{\pi}^{\top}\mathbf{Z}^{\top}\mathbf{Z}\boldsymbol{\pi}}{\sigma_{\mathrm{v}}^2}$$

i.e. the signal-to-noise ratio in the reduced form regression.

Can show that $\left|\mathrm{E}\left[\hat{\beta}_{\mathrm{2SLS}}\right] - \beta\right|$ is decreasing in τ^2, so having weak instruments tends to increase bias.

Note: (τ^2/m) is approximately the F-statistic (in fact the F-statistic minus one) for a test of H_0: $\boldsymbol{\pi}$=**0**.

So: A test for finite-sample bias/instrument weakness can be based on (τ^2/m).

In fact: For the case m=d=1, then 1/F is approximately unbiased for

$$\frac{\left|\mathrm{E}\left[\hat{\beta}_{\mathrm{IV}}\right]-\beta\right|}{\left|\mathrm{E}\left[\hat{\beta}_{\mathrm{2SLS}}\right]-\beta\right|}$$

(Straiger and Stock, 1997, EMA)

In general, the extent to which $\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{2SLS}}$ is biased relative to $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ depends on

- The # of endogenous regressors
- M

Straiger and Stock's simulations proposed the rule of thumb.

- "F>10" to keep the bias of $\widehat{\boldsymbol{\beta}}_{\mathrm{IV}}$ or $\widehat{\boldsymbol{\beta}}_{\mathrm{2SLS}}$ to no more than .10 of $\left\|\mathrm{E}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right]-\boldsymbol{\beta}\right\|$

Empirical example

Kling (2001, JBES) considered the causal effect of schooling on earnings. Kling used college proximity (to college) as an instrument for schooling.

The model is

$$\log \mathrm{W_i} = \beta_0 + \beta_1 \mathrm{S_i} + \beta_2 \mathrm{E_i} + \beta_3 \mathrm{E_i^2} + \mathbf{x}_{2\mathrm{i}}^{\top}\boldsymbol{\gamma} + \mathrm{u_i}, \qquad \mathrm{i} = 1, \cdots, 3010$$

where $\mathrm{W_i} \equiv$ average hourly earnings

$\mathrm{S_i} \equiv$ years of schooling

$\mathrm{E_i} \equiv \mathrm{Age_i} - \mathrm{S_i} - 6$

$\mathbf{x}_{2\mathrm{i}} \equiv$ vector of 26 additional variables

(Data were from the NLSY and consisted of observations on men aged 24–34 in 1976)

Assume

- S_i, E_i, E_i^2 are endogenous because of unobservable ability
- Because of this, require at least 3 instruments not present in $\mathbf{x}_{2i}$
- Kling used three excluded instruments

$$\mathrm{Col4_i} \equiv \text{indicator for proximity to a 4 year college}$$
$$\mathrm{Age_i}, \mathrm{Age_i^2} \text{ in years and years squared}$$

- The overall instrument vector was completed with a constant and $\mathbf{x}_2$, i.e. m=d=30
- Kling tested for homoscedasticity and did not reject, so used homoscedasticity-only standard errors
- The OLS coefficient on S_i is .073 (.004)

Interpretation: An additional year of schooling causes wages to rise on average by 7.6%=exp(.073)−1

- The IV coefficient on S_i is .132 (.049)
- Shea's partial R^2 on schooling is R_p^2=.0064, which implies that the IV standard error on schooling should be 1/.08=12.5 times the OLS standard error, which is a close to what we observe
- Now consider the reduced-form regression of S_i on all instruments

Get an F-statistic of 8.07, which suggests that the finite-sample bias of the IV estimator is between 10 and 20% of that of OLS.

Summary: Working with instruments is inherently difficult. Ideally, one should avoid using instruments by exploiting "better" data from randomized controlled experiments or from a natural experiment

Nonlinear estimators

Suppose we observe

$$\{[\mathbf{x}_i^\top \quad y_i]^\top : i = 1, \cdots, n\}$$

A linear estimator is one that is linear in y_i.

e.g. OLS

$$\widehat{\boldsymbol{\beta}}_{OLS} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$$

i.e. weighted sum of the y_is.

e.g. 2SLS

$$\widehat{\boldsymbol{\beta}}_{2SLS} = (\mathbf{X}^\top \mathbf{P}_\mathbf{Z} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_\mathbf{Z} \mathbf{y}$$

i.e. weighted sum of the y_is.

A nonlinear estimator is one that is not linear in the y_is. These are often associated with

a) Estimates of conditional mean function parameters in conditional means that are not linear in parameters
b) Estimators of parameters defined implicitly as solutions to optimization problems
c) "Limited" dependent variable models, e.g. models in which the y_is are generated via censoring or trunction

We will discuss

1) Extremum (M-) estimators

(i.e. estimators defined to solve an optimization problem)

2) Estimating-equations (Z-) estimators

(i.e. estimators defined to solve a system of estimating equations)

3) Maximum-likelihood estimators (MLEs) and quasi-maximum-likelihood estimators (QMLEs)
4) Nonlinear least squares estimators

Note: These four categories overlap and include linear estimators

e.g. OLS is an extremum estimator since

$$\widehat{\boldsymbol{\beta}}_{OLS} = \underset{\mathbf{b} \in \mathbb{R}^d}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\mathbf{b})^\top (\mathbf{y} - \mathbf{X}\mathbf{b})$$

$\widehat{\boldsymbol{\beta}}_{OLS}$ is also an estimating equations estimator because it is a solution to the system

$$\mathbf{X}^\top (\mathbf{X}\mathbf{b} - \mathbf{y}) = \mathbf{0}$$

$\widehat{\boldsymbol{\beta}}_{OLS}$ is an MLE of $\boldsymbol{\beta}$ in the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad \text{where } \mathbf{u}|\mathbf{X} \sim N(0, \sigma^2 \mathbf{I}_n)$$

e.g. Poisson regression

Observe $(\mathbf{x}_i^\top y_i)^\top$ (i=1,⋯,n).

Where for each i the conditional distribution of y_i given xi is Poisson($\exp(\mathbf{x}_i^\top\boldsymbol{\beta})$), i.e. the conditional density of $\mathbf{y}_i$ given $\mathbf{x}_i$ is

$$f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = \begin{cases} \dfrac{\exp\left(-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right)\left(\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right)^{y_i}}{y_i!}, & y_i = 0,1,2,\cdots \\ 0, & \text{otherwise} \end{cases}$$

Assuming iid observations, the MLE of $\boldsymbol{\beta}$ maximizes

$$\prod_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\beta}), \qquad \text{as a function of } \boldsymbol{\beta}$$

i.e. the MLE of $\boldsymbol{\beta}$ maximizes

$$\begin{aligned} \widehat{Q}_n(\boldsymbol{\beta}) &\equiv \frac{1}{n}\log\prod_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(-\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right) + y_i\mathbf{x}_i^\top\boldsymbol{\beta} - \log y_i!\right) \end{aligned}$$

i.e. the MLE of $\boldsymbol{\beta}$ solves the FOCs

$$\frac{\partial}{\partial\boldsymbol{\beta}^\top}\widehat{Q}_n(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_n} = \mathbf{0} \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \exp\left(\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_n\right)\right)\mathbf{x}_i = \mathbf{0}$$

(No closed form expression for $\widehat{\boldsymbol{\beta}}_n$)

Extremum estimators: In general, an extremum estimator/M-estimator $\widehat{\boldsymbol{\theta}}_n$ of a parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ maximizes an objective function $\widehat{Q}_n(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i, y_i, \boldsymbol{\theta})$ for a scalar-valued contrast function $g(\cdot,\cdot,\cdot)$.

Different extremum estimators correspond to different contrast functions (e.g. MLE $g(\cdot,\cdot,\cdot)$ is a log density) (e.g. OLS $g(\cdot,\cdot,\cdot)$ is the squared residual).

Consider an extremum estimator $\widehat{\boldsymbol{\theta}}_n$ of a parameter $\boldsymbol{\theta}_0$, i.e.

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i, y_i, \boldsymbol{\theta})$$

e.g. MLE of $\boldsymbol{\beta}$ in a Poisson regression

$$y|\mathbf{x} \sim \text{Poisson}\left(e^{\mathbf{x}^\top\boldsymbol{\beta}}\right)$$
$$\Rightarrow g(\mathbf{x}_i, y_i, \boldsymbol{\beta}) = -e^{\mathbf{x}_i^\top\boldsymbol{\beta}} + y_i\mathbf{x}_i^\top\boldsymbol{\beta} - \log y_i!$$

e.g. OLS estimator of $\boldsymbol{\beta}$ in a linear model with

$$E[y|\mathbf{x}] = \mathbf{x}^\top\boldsymbol{\beta}$$
$$\Rightarrow g(\mathbf{x}_i, y_i, \boldsymbol{\beta}) = \left(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}\right)^2$$

We usually work with M-estimators $\widehat{\boldsymbol{\theta}}_n$ that solve systems of first-order conditions

$$\frac{\partial}{\partial\boldsymbol{\theta}}\widehat{Q}_n(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} = \mathbf{0}$$

i.e. M-estimators that are also Z-estimators

An M-estimator is worthwhile if it is both consistent ($\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$) and if it has a "tractable" asymptotic distribution, i.e. if there is an increasing sequence of constants $\{a_n\}$ s.t. $a_n\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathbf{L}(\boldsymbol{\theta}_0)$, where $\mathbf{L}(\boldsymbol{\theta}_0)$ is a random vector whose distribution is a known function of $\boldsymbol{\theta}_0$. This implies that if n is large, then $\mathbf{L}(\boldsymbol{\theta}_0)$ has a distribution approximately equal to that of $\mathbf{L}\left(\widehat{\boldsymbol{\theta}}_n\right)$; i.e. tests/confidence regions regarding $\boldsymbol{\theta}_0$ can be based off the (known) distribution of $\mathbf{L}\left(\widehat{\boldsymbol{\theta}}_n\right)$.

Usual case in this course

$$a_n = \sqrt{n}, \qquad \mathbf{L}(\boldsymbol{\theta}_0) \sim N_d\left(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}_0)\right)$$

In particular, if $\widehat{\boldsymbol{\theta}}_n$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}g(\mathbf{x}_i, y_i, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} = \mathbf{0}$$

then a Taylor expansion (assuming that it is feasible) shows that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) = -\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\top}g(\mathbf{x}_i, y_i, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}_n}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}g(\mathbf{x}_i, y_i, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

where $\bar{\boldsymbol{\theta}}_n$ is a point on the line segment connecting $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$.

Therefore

$$\text{if } \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}g(\mathbf{x}_i,y_i,\boldsymbol{\theta}_0)\xrightarrow{d}N_d(\mathbf{0},\mathbf{B}_0)$$

where $\mathbf{B}_0$ is finite and non-singular and

$$\text{if } \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\top}g(\mathbf{x}_i,y_i,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}_n}\xrightarrow{p}\mathbf{A}_0$$

where $\mathbf{A}_0$ is finite and non-singular, then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\xrightarrow{d}N_d(\mathbf{0},\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1})$$

by Slutsky's theorem.

Suppose $\hat{\mathbf{A}}_n\xrightarrow{p}\mathbf{A}_0$, $\hat{\mathbf{B}}_n\xrightarrow{p}\mathbf{B}_0$, i.e. $\mathbf{A}_0$, $\mathbf{B}_0$ can be consistently estimated. In this case, the asymptotic variance $\frac{1}{n}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$ of $\hat{\boldsymbol{\theta}}_n$ can be consistently estimated as

$$\hat{\mathbf{V}}_n\left[\hat{\boldsymbol{\theta}}_n\right]\equiv\frac{1}{n}\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n\hat{\mathbf{A}}_n^{-1}$$

e.g. Suppose we observe $[\mathbf{x}_i^\top \quad y_i]^\top$ (i=1,⋯,n) iid with $y_i|\mathbf{x}_i\sim\text{Poisson}\left(e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right)$.

Then the Poisson MLE $\hat{\boldsymbol{\beta}}_n$ is given by

$$\hat{\boldsymbol{\beta}}_n=\underset{\boldsymbol{\beta}}{\text{argmax}}\frac{1}{n}\sum_{i=1}^{n}g(\mathbf{x}_i,y_i,\boldsymbol{\beta})$$

where

$$g(\mathbf{x}_i,y_i,\boldsymbol{\beta})=-e^{\mathbf{x}_i^\top\boldsymbol{\beta}}+y_i\mathbf{x}_i^\top\boldsymbol{\beta}-\log y_i!$$

So

$$\frac{\partial}{\partial\boldsymbol{\beta}}g(\mathbf{x}_i,y_i,\boldsymbol{\beta})=\left(-e^{\mathbf{x}_i^\top\boldsymbol{\beta}}+y_i\right)\mathbf{x}_i$$

and

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top}g(\mathbf{x}_i,y_i,\boldsymbol{\beta})=-e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\mathbf{x}_i\mathbf{x}_i^\top$$

Therefore

$$\mathbf{A}_0=-\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}e^{\mathbf{x}_i^\top\boldsymbol{\beta}_0}\mathbf{x}_i\mathbf{x}_i^\top$$

$$\mathbf{B}_0=\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right)^2\mathbf{x}_i\mathbf{x}_i^\top$$

The asymptotic variance of the Poisson MLE is consistently estimated as

$$\widehat{\mathbf{V}}_n\left[\widehat{\boldsymbol{\beta}}_n\right] = \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n\widehat{\mathbf{A}}_n^{-1}$$

where

$$\widehat{\mathbf{A}}_n = \frac{1}{n}\sum_{i=1}^{n} e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\mathbf{x}_i\mathbf{x}_i^\top$$

$$\widehat{\mathbf{B}}_n = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - e^{\mathbf{x}_i^\top\boldsymbol{\beta}}\right)^2\mathbf{x}_i\mathbf{x}_i^\top$$

Note: If $y_i|\mathbf{x}_i \sim \text{Poisson}\left(e^{\mathbf{x}_i^\top\boldsymbol{\beta}_0}\right)$ then $E[y_i|\mathbf{x}_i] = \text{Var}[y_i|\mathbf{x}_i] = e^{\mathbf{x}_i^\top\boldsymbol{\beta}_0}$

So

$$E\left[\frac{\partial}{\partial\boldsymbol{\beta}}g(\mathbf{x},y,\boldsymbol{\beta}_0)\middle|\mathbf{x}\right] = \mathbf{0}$$

$$\text{Var}\left[\frac{\partial}{\partial\boldsymbol{\beta}}g(\mathbf{x},y,\boldsymbol{\beta}_0)\middle|\mathbf{x}\right] = \text{Var}[y|\mathbf{x}]\mathbf{x}\mathbf{x}^\top = e^{\mathbf{x}^\top\boldsymbol{\beta}}\mathbf{x}\mathbf{x}^\top$$

which implies that

$$\mathbf{B}_0 = E\left[e^{\mathbf{x}^\top\boldsymbol{\beta}_0}\mathbf{x}\mathbf{x}^\top\right] = -\mathbf{A}_0$$

In this case, the asymptotic variance of the Poisson MLE can also be estimated as

$$\widetilde{\mathbf{V}}_n\left[\widehat{\boldsymbol{\beta}}_n\right] = -\frac{1}{n}\widehat{\mathbf{A}}_n^{-1}$$

which implies smaller standard errors than those available from $\widehat{\mathbf{V}}_n\left[\widehat{\boldsymbol{\beta}}_n\right]$. More popular in this case to use $\widehat{\mathbf{V}}_n\left[\widehat{\boldsymbol{\beta}}_n\right]$ because of its robustness against deviations from the assumption $y|\mathbf{x} \sim \text{Poisson}\left(e^{\mathbf{x}^\top\boldsymbol{\beta}_0}\right)$.

Interpretation of coefficients in nonlinear regression models

Suppose $E[y|\mathbf{x}]=g(\mathbf{x},\boldsymbol{\beta}_0)$ w.p. 1. Basically care about marginal effects, i.e. the vector

$$\frac{\partial}{\partial\mathbf{x}}E[y|\mathbf{x}]$$

as a function of $\mathbf{x}$. If $E[y|\mathbf{x}]=\mathbf{x}^\top\boldsymbol{\beta}$ (i.e. linear in $\boldsymbol{\beta}$) then the vector of marginal effect is constant in $\mathbf{x}$ and in fact is $\boldsymbol{\beta}$.

In nonlinear models, $\frac{\partial}{\partial\mathbf{x}}E[y|\mathbf{x}]$ varies with $\mathbf{x}$. In this case, not usually easy to interpret $\boldsymbol{\beta}_0$.

e.g. $E[y|\mathbf{x}] = e^{\mathbf{x}^\top\boldsymbol{\beta}_0}$

$$\frac{\partial}{\partial\mathbf{x}}E[y|\mathbf{x}] = e^{\mathbf{x}^\top\boldsymbol{\beta}_0}\boldsymbol{\beta}_0$$

which varies clearly with $\mathbf{x}$.

In nonlinear models, the distribution of $\frac{\partial}{\partial \mathbf{x}} \mathrm{E}[\mathrm{y}|\mathbf{x}]$ gives a comprehensive picture of the relation between **x** and y.

But: Difficult to interpret when **x** is relatively high dimensional.

Instead, popular to focus on

(1)

$$\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\partial}{\partial \mathbf{x}_{\mathrm{i}}} \mathrm{E}[\mathrm{y}_{\mathrm{i}}|\mathbf{x}_{\mathrm{i}}]$$

i.e. vector of average marginal responses

(2)

$$\frac{\partial}{\partial \mathbf{x}} \mathrm{E}[\mathrm{y}|\mathbf{x}]\Big|_{\mathbf{x}=\bar{\mathbf{x}}}$$

i.e. marginal response for the "average individual"

(3)

$$\frac{\partial}{\partial \mathbf{x}} \mathrm{E}[\mathrm{y}|\mathbf{x}]\Big|_{\mathbf{x}=\mathbf{x}^*}$$

i.e. marginal response for an individual with characteristics $\mathbf{x}^*$

Note:

- (1)–(3) are all $\boldsymbol{\beta}_0$ when $\mathrm{E}[\mathrm{y}|\mathbf{x}]=\mathbf{x}^{\top}\boldsymbol{\beta}_0$
- (1) is probably most relevant for policy analysis, followed closely by (3)
- (2) tends to be default output in the statistical software packages

Extremum estimation in some detail

Prerequisite: Uniform weak law of large numbers (UWLLN). This relates to certain random functions.

Definition (random function) Let $(\mathrm{S},\mathcal{A},\mathrm{P})$ be a probability space. A random function $\mathrm{f}(\cdot,\boldsymbol{\theta})$ on a subset Θ of $\mathbb{R}^{\mathrm{m}}$ is a mapping

$$\mathrm{f}(\mathrm{s},\boldsymbol{\theta}): \mathrm{S} \times \Theta \to \mathbb{R}$$

such that $\forall$ Borel set B in $\mathbb{R}$ and $\forall\ \boldsymbol{\theta} \in \Theta$, the set $\{\mathrm{s} \in \mathrm{S}: \mathrm{f}(\mathrm{s},\boldsymbol{\theta}) \in \mathrm{B}\}$ is measurable $\mathcal{A}$, i.e. $\mathrm{P}[\{\mathrm{s} \in \mathrm{S}: \mathrm{f}(\mathrm{s},\boldsymbol{\theta}) \in \mathrm{B}\}]$ is defined $\mathrm{B} \in \mathcal{B}$ and $\forall\ \boldsymbol{\theta} \in \Theta$. So, random functions generally look like $\mathrm{g}(\mathbf{x},\boldsymbol{\theta})$, where **x** is a random vector and $\boldsymbol{\theta}$ is a vector of constants.

Theorem 1 (UWLLN) Let $\mathbf{x}_1,\cdots,\mathbf{x}_n$ be iid random k-vectors. Let $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$, where Θ is closed and bounded. Let $g(\mathbf{x},\boldsymbol{\theta})$ be a Borel measurable function on $\mathbb{R}^k \times \Theta$ s.t. $g(\mathbf{x},\boldsymbol{\theta})$ is continuous on $\Theta \; \forall \, \mathbf{x} \in \mathbb{R}^k$. Assume

$$E\left[\sup_{\boldsymbol{\theta}\in\Theta}|g(\mathbf{x}_1,\boldsymbol{\theta})|\right] < \infty$$

Then

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i,\boldsymbol{\theta}) - E[g(\mathbf{x}_1,\boldsymbol{\theta})]\right| \xrightarrow{p} 0$$

i.e. $\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i,\boldsymbol{\theta})$ converges in probability to $E[g(\mathbf{x}_1,\boldsymbol{\theta})]$ uniformly on Θ.

Theorem 1 implies a result on the consistency of M-estimators based on iid observations.

Suppose the parameter of interest is

$$\boldsymbol{\theta}_0 \equiv \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, E[g(\mathbf{x}_1,\boldsymbol{\theta})]$$

where $g(\mathbf{x},\boldsymbol{\theta})$ satisfies the conditions of Theorem 1.

$$\text{Let } \widehat{\boldsymbol{\theta}}_n \equiv \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i,\boldsymbol{\theta})$$

Theorem 2 (consistency of M-estimators) Assume

$$\mathbf{x}_1,\cdots,\mathbf{x}_n, \qquad g(\cdot,\cdot), \qquad \Theta$$

are all as in the statement of Theorem 1. Assume that $\boldsymbol{\theta}_0$ is unique, i.e.

$$\forall\, \varepsilon > 0, \qquad \exists\, \delta(\varepsilon) > 0, \qquad \text{s.t. } E[g(\mathbf{x}_1,\boldsymbol{\theta}_0)] - \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\varepsilon} E[g(\mathbf{x}_1,\boldsymbol{\theta})] > \delta_\varepsilon$$

Then

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$$

Note: If Θ is compact and $E[g(\mathbf{x}_1,\boldsymbol{\theta})]$ is continuous on Θ then $E[g(\mathbf{x}_1,\boldsymbol{\theta})]$ is uniquely maximized on Θ.

Proof Let

$$\widehat{Q}_n(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i, \boldsymbol{\theta})$$
$$\overline{Q}_n(\boldsymbol{\theta}) \equiv E[g(\mathbf{x}_1, \boldsymbol{\theta})]$$

Then $\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0 \in \Theta$ because $g(\mathbf{x},\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \Theta$, which is compact. By the definition of $\boldsymbol{\theta}_0$,

$$\begin{aligned} 0 &\leq \overline{Q}(\boldsymbol{\theta}_0) - \overline{Q}(\widehat{\boldsymbol{\theta}}_n) \\ &= \overline{Q}(\boldsymbol{\theta}_0) - \widehat{Q}_n(\boldsymbol{\theta}_0) + \widehat{Q}_n(\boldsymbol{\theta}_0) - \overline{Q}(\widehat{\boldsymbol{\theta}}_n) \\ &\leq \overline{Q}(\boldsymbol{\theta}_0) - \widehat{Q}_n(\boldsymbol{\theta}_0) + \widehat{Q}_n(\widehat{\boldsymbol{\theta}}_n) - \overline{Q}(\widehat{\boldsymbol{\theta}}_n) \\ &\leq 2 \sup_{\boldsymbol{\theta}\in\Theta} \left|\widehat{Q}_n(\boldsymbol{\theta}) - \overline{Q}(\mathbf{n})\right| \\ &\xrightarrow{p} 0, \qquad \text{by Theorem 1,} \qquad \cdots (*) \end{aligned}$$

Therefore,

$$\overline{Q}(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{p} \overline{Q}(\boldsymbol{\theta}_0)$$

But $\boldsymbol{\theta}_0$ is unique so $\forall\, \varepsilon > 0,\ \exists \delta(\varepsilon) > 0$ s.t. $\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\| > \varepsilon$ implies

$$\overline{Q}(\boldsymbol{\theta}_0) - \overline{Q}(\widehat{\boldsymbol{\theta}}_n) \geq \delta_\varepsilon$$

which implies

$$P\left[\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right\| > \varepsilon\right] \leq P\left[\overline{Q}(\boldsymbol{\theta}_0) - \overline{Q}(\widehat{\boldsymbol{\theta}}_n) > \delta_\varepsilon\right] \to 0, \qquad \text{by } (*), \qquad \text{as } n \to \infty$$

$$\widehat{\boldsymbol{\theta}}_n \equiv \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \widehat{Q}_n(\boldsymbol{\theta})$$
$$\widehat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n} g(\mathbf{X}_j, \boldsymbol{\theta})$$

Want to estimate $\boldsymbol{\theta}_0$.
Assume sufficient conditions for consistency:

- $\Theta \subset \mathbb{R}^m$ is closed and bounded
- $g(\mathbf{x},\boldsymbol{\theta})$ is Borel-measurable on $\mathbb{R}^k \times \Theta$ and $g(\mathbf{x},\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ $\forall \mathbf{x} \in \mathbb{R}^k$
- $E\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|g(\mathbf{X}_j, \boldsymbol{\theta})\right|\right] < \infty$
- $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are iid random k-vectors
- For $\overline{Q}(\boldsymbol{\theta}) \equiv E[g(\mathbf{X}_1, \boldsymbol{\theta})]$ suppose that $\boldsymbol{\theta}_0 \equiv \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \overline{Q}(\boldsymbol{\theta})$ is unique.

Add the following conditions that together with the conditions suffice for asymptotic Normality

- Θ is convex
- $\boldsymbol{\theta}_0$ is an interior point of Θ
- $\forall \mathbf{x} \in \mathbb{R}^k$, $g(\mathbf{x},\boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta} \in \Theta$
- $E\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{\partial^2 g(\mathbf{X}_1,\boldsymbol{\theta})}{\partial\theta_{i_1}\partial\theta_{i_2}}\right|\right] < \infty$ for any components (not necessarily distinct) θ_{i_1}, θ_{i_2} of $\boldsymbol{\theta}$
- $\mathbf{A}_0 \equiv E\left[\frac{\partial^2 g(\mathbf{X}_1,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0^T}\right]$ is nonsingular
- $\mathbf{B}_0 \equiv E\left[\frac{\partial g(\mathbf{X}_1,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0^T}\frac{\partial g(\mathbf{X}_1,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_0}\right]$ is finite

Recall: $\widehat{\boldsymbol{\theta}}_n$ is <u>asymptotically Normal</u> iff there is an increasing sequence of positive numbers $\{a_n\}$ and a positive semidefinite matrix $\boldsymbol{\Sigma}$ s.t.

$$a_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$$

(this implies consistency of $\widehat{\boldsymbol{\theta}}_n$, since)

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = O_p(a_n^{-1}) = o_p(1)$$

Here we consider the case of extremum estimators that are <u>root-n-consistent</u> and asymptotically Normal, i.e.

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$$

Theorem The eleven conditions given earlier imply

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_m\left(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\right)$$

Proof Consider only the case m=1. Recall that the first five conditions imply the consistency $\hat{\theta}_n \xrightarrow{p} \theta_0$. Since $\theta_0 \in \text{int}\Theta$, then $P\left[\hat{\theta}_n \in \text{int}\Theta\right] \to 1$. Therefore $P\left[\hat{Q}_n'\left(\hat{\theta}_n\right) = 0\right] \to 1$ (where we have applied differentiability of $\hat{Q}_n(\theta)$). Apply the two-times continuous differentiability of $\hat{Q}_n(\boldsymbol{\theta})$ and the mean-value theorem to deduce that $\exists\hat{\lambda}_n \in (0,1)$ s.t.

$$\sqrt{n}\hat{Q}_n'\left(\hat{\theta}_n\right) = \sqrt{n}\hat{Q}_n'(\theta_0) + \hat{Q}_n''\left(\theta_0 + \hat{\lambda}_n\left(\hat{\theta}_n - \theta_0\right)\right)\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$$

Since Θ is convex, $P\left[\theta_0 + \hat{\lambda}_n\left(\hat{\theta}_n - \theta_0\right) \in \Theta\right] = 1$ and since $\hat{\theta}_n \xrightarrow{p} \theta_0$, we have $\theta_0 + \hat{\lambda}_n\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{p} \theta_0$. By the UWLLN and the twice continuous differentiability of $\overline{Q}(\theta)$ (dominate convergence→check, second-order partial derivatives?) we have

$$\sup_{\theta\in\Theta}\left|\hat{Q}_n''(\theta) - \overline{Q}''(\theta)\right| \xrightarrow{p} 0$$

Therefore $\hat{Q}_n''\left(\theta_0 + \hat{\lambda}_n\left(\hat{\theta}_n - \theta_0\right)\right) \xrightarrow{p} \overline{Q}''(\theta_0) < 0$ since θ_0 is the unique maximizer of $\overline{Q}(\theta)$. Therefore

$$\left(\hat{Q}_n''\left(\theta_0 + \hat{\lambda}_n\left(\hat{\theta}_n - \theta_0\right)\right)\right)^{-1} \xrightarrow{p} \left(\overline{Q}''(\theta_0)\right)^{-1} = A_0^{-1}$$

Go back to the mean-value expansion, we have

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) &= -\left(\hat{Q}_n''\left(\theta_0 + \hat{\lambda}_n\left(\hat{\theta}_n - \theta_0\right)\right)\right)^{-1}\sqrt{n}\hat{Q}_n'(\theta_0) + \left(\hat{Q}_n''\left(\theta_0 + \hat{\lambda}_n\left(\hat{\theta}_n - \theta_0\right)\right)\right)^{-1}\sqrt{n}\hat{Q}_n'\left(\hat{\theta}_n\right)\\ &= -\left(\hat{Q}_n''\left(\theta_0 + \hat{\lambda}_n\left(\hat{\theta}_n - \theta_0\right)\right)\right)^{-1}\sqrt{n}\hat{Q}_n'(\theta_0) + o_p(1)\end{aligned}$$

since $P\left[\hat{Q}_n'\left(\hat{\theta}_n\right) = 0\right] \to 1$. We have $\theta_0 \in \text{int}\Theta$, so

$$0 = \overline{Q}'(\theta_0) = E\left[\frac{dg(X_1, \theta_0)}{d\theta_0}\right]$$

In addition, we have

$$B_0 = \text{Var}\left[\frac{dg(X_1, \theta_0)}{d\theta_0}\right] \in (0, \infty), \qquad \text{by assumption}$$

Therefore

$$\sqrt{n}\hat{Q}_n'(\theta_0) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{dg\left(X_j, \theta_0\right)}{d\theta_0} \xrightarrow{d} N(0, B_0)$$

By Slutsky's theorem, $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, A_0^{-1}B_0A_0^{-1})$

In general, if an extremum estimator $\widehat{\boldsymbol{\theta}}_n$ satisfies

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_m\left(0, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\right)$$

Need to estimate the asymptotic variance

$$\mathbf{V}\left[\widehat{\boldsymbol{\theta}}_n\right] = \frac{1}{n}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$$

Let

$$\widehat{\mathbf{A}}_n \equiv \frac{1}{n}\sum_{j=1}^{n} \left.\frac{\partial^2 g\left(\mathbf{X}_j, \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$$

$$\widehat{\mathbf{B}}_n \equiv \frac{1}{n}\sum_{j=1}^{n} \left.\frac{\partial g\left(\mathbf{X}_j, \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \left.\frac{\partial g\left(\mathbf{X}_j, \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n}$$

Under the eleven conditions sufficient for the result

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_m\left(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\right)$$

We have $\widehat{\mathbf{A}}_n \xrightarrow{p} \mathbf{A}_0$

Impose the additional condition

$$E\left[\sup_{\theta\in\Theta}\left\|\frac{\partial g(\mathbf{X}_1, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\|^2\right] < \infty$$

Then if the eleven conditions sufficient for asymptotic Normality hold, $\widehat{\mathbf{B}}_n \xrightarrow{p} \mathbf{B}_0$ (exercise: check the applicability of the appropriate UWLLN). Therefore, under the twelve previous conditions

$$\widehat{\mathbf{V}}_n\left[\widehat{\boldsymbol{\theta}}_n\right] \equiv \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n\widehat{\mathbf{A}}_n^{-1} \text{ satisfies } \left\|\widehat{\mathbf{V}}_n\left[\widehat{\boldsymbol{\theta}}_n\right] - \mathbf{V}\left[\widehat{\boldsymbol{\theta}}_n\right]\right\| \xrightarrow{p} 0$$

So: If the twelve conditions hold, then it is straightforward to test hypothesis of the form

$$H_0: \mathbf{R}\boldsymbol{\theta}_0 = \mathbf{q}, \qquad \text{vs } H_1: \mathbf{R}\boldsymbol{\theta}_0 \neq \mathbf{q}$$

where $\mathbf{R}$ is an (r×m)-matrix of rank r, $\mathbf{q} \in \mathbb{R}^r$

To see this, under H_0,

$$\sqrt{n}\left(\mathbf{R}\widehat{\boldsymbol{\theta}}_n - \mathbf{q}\right) \xrightarrow{d} N_r\left(\mathbf{0}, \mathbf{R}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\mathbf{R}^\top\right)$$

(also if $\mathbf{B}_0$ is nonsingular, then $\mathbf{R}\mathbf{A}_0{}^{-1}\mathbf{B}_0\mathbf{A}_0{}^{-1}\mathbf{R}^\top$ is invertible, otherwise can consider a generalized inverse).

By Slutsky's theorem:

$$W_n = n\left(\mathbf{R}\widehat{\boldsymbol{\theta}}_n - \mathbf{q}\right)^\top\left(\mathbf{R}\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n\widehat{\mathbf{A}}_n^{-1}\mathbf{R}^\top\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\theta}}_n - \mathbf{q}\right) \xrightarrow{d} \chi_r^2, \qquad \text{under } H_0$$

This allows one to formulate a test of (approximate) level α when n is large.

We have $P[W_n > B_\alpha] \to \alpha$ under H_0 as $n\to\infty$, where B_α=(1−α)-quantile of a χ_r^2.

So: If rejection of H_0: $\mathbf{R\theta}_0=\mathbf{q}$ occurs iff $W_n>B_\alpha$ then the probability of type-I error is approximately α when n is large.

To see that this "Wald-type" test has power, suppose H_0 is false, then under our twelve conditions

$$\frac{W_n}{n} \xrightarrow{p} (\mathbf{R\theta}_0 - \mathbf{q})^\top (\mathbf{R}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\mathbf{R}^\top)^{-1}(\mathbf{R\theta}_0 - \mathbf{q}) > 0$$

e.g. Poisson MLE: Suppose we observe $(\mathbf{X}_1^\top, Y_1)^\top, \ldots, (\mathbf{X}_n^\top, Y_n)^\top$ iid. Assume $E[Y_i|\mathbf{X}_i] = e^{\mathbf{X}_i^\top \boldsymbol{\beta}_0}$ w.p. 1. Let

$$\widehat{Q}_n(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^{n}\left(-e^{\mathbf{X}_i^\top \boldsymbol{\beta}} + Y_i \mathbf{X}_i^\top \boldsymbol{\beta} - \log Y_i!\right), \qquad \text{i.e. the Poisson likelihood}$$

Let

$$\overline{Q}(\boldsymbol{\beta}) \equiv E\left[\widehat{Q}_n(\boldsymbol{\beta})\right]$$

By Kolmogorov's LLN

$$\overline{Q}(\boldsymbol{\beta}) = \operatorname{plim} \widehat{Q}_n(\boldsymbol{\beta}) = -E\left[e^{\mathbf{X}_1^\top \boldsymbol{\beta}}\right] + E[Y_1 \mathbf{X}_1^\top \boldsymbol{\beta}] - E[\log Y_1!]$$

provided that each of these expectations exists. Since $E[Y_i|\mathbf{X}_i] = e^{\mathbf{X}_i^\top \boldsymbol{\beta}_0}$ w.p. 1, so

$$\overline{Q}(\boldsymbol{\beta}) = -E\left[e^{\mathbf{X}_1^\top \boldsymbol{\beta}}\right] + E\left[e^{\mathbf{X}_1^\top \boldsymbol{\beta}_0} \mathbf{X}_1^\top \boldsymbol{\beta}\right] - E[\log Y_1!]$$

Therefore, if dominated convergence is applicable (say $\boldsymbol{\beta} \in B$, where B is compact and $E[\|\mathbf{X}_1\|^2] < \infty$) then

$$\frac{\partial}{\partial \boldsymbol{\beta}^\top}\overline{Q}(\boldsymbol{\beta}) = -E\left[e^{\mathbf{X}_1^\top \boldsymbol{\beta}} \mathbf{X}_1\right] + E\left[e^{\mathbf{X}_1^\top \boldsymbol{\beta}_0} \mathbf{X}_1\right], \qquad \text{i.e.} \left.\frac{\partial}{\partial \boldsymbol{\beta}^\top}\overline{Q}(\boldsymbol{\beta})\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{0}$$

In addition,

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top}\overline{Q}(\boldsymbol{\beta}) = -E\left[e^{\mathbf{X}_1^\top \boldsymbol{\beta}} \mathbf{X}_1 \mathbf{X}_1^\top\right]$$

which is negative definite if there is no perfect multicollinearity.

Therefore $\overline{Q}(\boldsymbol{\beta})$ is globally maximized at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

Consistency of the Poisson MLE follows from what we've discussed.

What about asymptotic Normality? Take a Taylor expansion to first-order of the FOCs to get

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top \boldsymbol{\beta}}\right)\mathbf{X}_i\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_n} = \mathbf{0}$$

to get

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) = -\left(-\frac{1}{n}\sum_{i=1}^{n} e^{\mathbf{X}_i^\top \overline{\boldsymbol{\beta}}_n}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top \boldsymbol{\beta}_0}\right)\mathbf{X}_i$$

where $\overline{\boldsymbol{\beta}}_n$ is between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. We have $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$, so $\overline{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$, so if $E[\|\mathbf{X}_1\|^4] < \infty$ (why?) then

$$-\frac{1}{n}\sum_{i=1}^{n} e^{\mathbf{X}_i^\top \overline{\boldsymbol{\beta}}_n}\mathbf{X}_i\mathbf{X}_i^\top \xrightarrow{p} -E\left[e^{\mathbf{X}_1^\top \boldsymbol{\beta}_0}\mathbf{X}_1\mathbf{X}_1^\top\right] \equiv \mathbf{A}_0, \qquad \cdots (*)$$

Next, suppose $\mathbf{X}_i$ is a scalar, i.e. $\mathbf{X}_i = X_i$. Then

$$E\left[\left(Y_1 - e^{X_1\beta_0}\right)X_1\right] = 0$$

and

$$\begin{aligned}\mathrm{Var}\left[\left(Y_1 - e^{X_1\beta_0}\right)X_1\right] &= E\left[\mathrm{Var}\left[\left(Y_1 - e^{X_1\beta_0}\right)X_1\middle|X_1\right]\right] + \mathrm{Var}\left[E\left[\left(Y_1 - e^{X_1\beta_0}\right)X_1\middle|X_1\right]\right] \\ &= E[\mathrm{Var}[Y_1|X_1]X_1^2] < \infty\end{aligned}$$

if $E[X_1^2]<\infty$ and $\mathrm{Var}[Y_1|X_1]<\infty$. By Lindberg–Lévy CLT,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i - e^{X_i\beta_0}\right)X_i \xrightarrow{d} N(0, E[\mathrm{Var}[Y_1|X_1]X_1^2]), \qquad \cdots (**)$$

Can extend (*) to the case where $\mathbf{X}_1$ is vector-valued by Cramér–Wold device.
In particular,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top \boldsymbol{\beta}_0}\right)\mathbf{X}_1 \xrightarrow{d} N_d(\mathbf{0}, E[\mathrm{Var}[Y_1|\mathbf{X}_1]\mathbf{X}_1\mathbf{X}_1^\top])$$

Let

$$\mathbf{B}_0 \equiv E[\mathrm{Var}[Y_1|\mathbf{X}_1]\mathbf{X}_1\mathbf{X}_1^\top]$$

Combining results, we get

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \xrightarrow{d} N_d(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1})$$

by Slutsky's theorem.

Poisson MLE

Observe $[\mathbf{X}_i^\top \; Y_i]$ (i=1,…,n) iid

$$\begin{aligned}\widehat{Q}_n(\boldsymbol{\beta}) &\equiv \frac{1}{n}\sum_{i=1}^{n}\left(-e^{\mathbf{X}_i^\top\boldsymbol{\beta}} + Y_i\mathbf{X}_i^\top\boldsymbol{\beta} - \log Y_i!\right)\\ \bar{Q}(\boldsymbol{\beta}) &\equiv \operatorname{plim}\widehat{Q}_n(\boldsymbol{\beta})\\ &= E\left[\widehat{Q}_n(\boldsymbol{\beta})\right]\\ &= -E\left[e^{\mathbf{X}_i^\top\boldsymbol{\beta}}\right] + E\left[Y_i\mathbf{X}_i^\top\boldsymbol{\beta}\right] - E[\log Y_i!]\end{aligned}$$

Assume $E[Y_1|\mathbf{X}_1] = e^{\mathbf{X}_1^\top\boldsymbol{\beta}_0}$ w.p. 1. We showed that

$$\frac{\partial}{\partial\boldsymbol{\beta}^\top}\bar{Q}(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{0}$$

(assuming that $E[Y_1|\mathbf{X}_1] = e^{\mathbf{X}_1^\top\boldsymbol{\beta}_0}$ and dominated convergence applies)

Also

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top}\bar{Q}(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = -E\left[e^{\mathbf{X}_1^\top\boldsymbol{\beta}_0}\mathbf{X}_1\mathbf{X}_1^\top\right]$$

i.e. negative definite if no perfect multicollinearity.

Therefore under appropriate conditions $\bar{Q}(\boldsymbol{\beta})$ is globally maximized at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ and consistency of the Poisson MLE follows.

What about asymptotic Normality?

Take a Taylor expansion to first order of the left hand side of the system

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top\boldsymbol{\beta}}\right)\mathbf{X}_i\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_n} = \mathbf{0}$$

to get

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) = -\left(-\frac{1}{n}\sum_{i=1}^{n}e^{\mathbf{X}_i^\top\bar{\boldsymbol{\beta}}_n}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top\boldsymbol{\beta}_0}\right)\mathbf{X}_i$$

where $\bar{\boldsymbol{\beta}}_n$ is between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$.

Since $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$, $\bar{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$ and therefore if e.g. $E[\|\mathbf{X}_1\|^4] < \infty$ we get

$$-\frac{1}{n}\sum_{i=1}^{n}e^{\mathbf{X}_i^\top\bar{\boldsymbol{\beta}}_n}\mathbf{X}_i\mathbf{X}_i^\top \xrightarrow{p} -E\left[e^{\mathbf{X}_1^\top\boldsymbol{\beta}_0}\mathbf{X}_1\mathbf{X}_1^\top\right] \equiv \mathbf{A}_0, \qquad \begin{pmatrix}\text{exercise:}\\ \text{supply details}\end{pmatrix}$$

Next suppose $\mathbf{X}_i$ is scalar, then

$$\mathrm{E}\left[\left(\mathrm{Y}_1 - \mathrm{e}^{\mathrm{X}_1\beta_0}\right)\mathrm{X}_1\right] = 0$$

and

$$\mathrm{Var}\left[\left(\mathrm{Y}_1 - \mathrm{e}^{\mathrm{X}_1\beta_0}\right)\mathrm{X}_1\right] = \mathrm{E}[\mathrm{Var}[\mathrm{Y}_1|\mathrm{X}_1]\mathrm{X}_1^2], \qquad \begin{pmatrix}\text{exercise:}\\ \text{supply details}\end{pmatrix}$$

Assume $\mathrm{P}[\mathrm{Var}[\mathrm{Y}_1|\mathrm{X}_1] < \infty] = 1$, $\mathrm{E}[\mathrm{X}_1^2] < \infty$. Then by the Lindberg–Lévy CLT

$$\frac{1}{\sqrt{\mathrm{n}}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{Y}_\mathrm{i} - \mathrm{e}^{\mathrm{X}_\mathrm{i}\beta_0}\right)\mathrm{X}_\mathrm{i} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathrm{E}[\mathrm{Var}[\mathrm{Y}_1|\mathrm{X}_1]\mathrm{X}_1^2])$$

The case with $\mathbf{X}_i$ not necessarily scalar follows the Cramér–Wold device (exercise: review).

Let $\mathbf{B}_0 \equiv \mathrm{E}[\mathrm{Var}[\mathrm{Y}_1|\mathbf{X}_1]\mathbf{X}_1\mathbf{X}_1^\top]$

$$\sqrt{\mathrm{n}}\left(\widehat{\boldsymbol{\beta}}_\mathrm{n} - \boldsymbol{\beta}_0\right) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}), \qquad \cdots (*)$$

Important: Convergence (*) holds event if $\mathrm{Y}_i|\mathbf{X}_i$ is not Poisson($\mathrm{e}^{\mathbf{X}_i^\top\boldsymbol{\beta}_0}$) all that is required for consistency that $\mathrm{E}[\mathrm{Y}_\mathrm{i}|\mathbf{X}_\mathrm{i}] = \mathrm{e}^{\mathbf{X}_\mathrm{i}^\top\boldsymbol{\beta}_0}$ (along with other conditions like $\mathrm{E}[\|\mathbf{X}_1\|^4] < \infty$). Asymptotic Normality also requires that $\mathrm{Var}[\mathrm{Y}_\mathrm{i}|\mathbf{X}_\mathrm{i}] < \infty$ w.p. 1. On the other hand, if $\mathrm{Var}[\mathrm{Y}_\mathrm{i}|\mathbf{X}_\mathrm{i}] = \mathrm{e}^{\mathbf{X}_\mathrm{i}^\top\boldsymbol{\beta}_0}$ (for example, if $\mathrm{Y}_i|\mathbf{X}_i \sim$ Poisson($\mathrm{e}^{\mathbf{X}_\mathrm{i}^\top\boldsymbol{\beta}_0}$) then $\mathbf{A}_0 = -\mathbf{B}_0$ and the convergence in (*) simplifies to)

$$\sqrt{\mathrm{n}}\left(\widehat{\boldsymbol{\beta}}_\mathrm{n} - \boldsymbol{\beta}_0\right) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, -\mathbf{A}_0^{-1})$$

In general, the Poisson MLE $\widehat{\boldsymbol{\beta}}_\mathrm{n}$ is really a quasi-MLE, i.e. don't need to assume the conditional distribution of $\mathrm{Y}_i|\mathbf{X}_i$ in order to estimate parameter of $\mathrm{E}[\mathrm{Y}_i|\mathbf{X}_i] = \mathrm{e}^{\mathbf{X}_\mathrm{i}^\top\boldsymbol{\beta}_0}$.

But: Assuming correctness of the nonlinear regression $\mathrm{E}[\mathrm{Y}_i|\mathrm{X}_i] = \mathrm{e}^{\mathbf{X}_\mathrm{i}^\top\boldsymbol{\beta}_0}$ is the asymptotic variance in (*) the best possible?

(In other words, is it possible for another M-estimator of $\mathbf{B}_0$ to be $\sqrt{\mathrm{n}}$-consistent and asymptotically Normal, but generate smaller standard errors?)

For example, nonlinear least squares, i.e.

$$\widehat{\boldsymbol{\beta}}_{\mathrm{NLLS}} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}}\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{Y}_\mathrm{i} - \mathrm{e}^{\mathbf{X}_\mathrm{i}^\top\boldsymbol{\beta}}\right)^2$$

The answer to this question involves the theory of efficient GMM.

On the other hand, suppose $\mathrm{Y}_i|\mathbf{X}_i \sim$ Poisson($\mathrm{e}^{\mathbf{X}_\mathrm{i}^\top\boldsymbol{\beta}_0}$) holds for each i.

In this case, the MLE $\widehat{\boldsymbol{\beta}}_\mathrm{n}$ is asymptotically efficient its asymptotic variance $-\frac{1}{\mathrm{n}}\mathbf{A}_0^{-1}$ dominates the asymptotic variance of any other "regular" estimator of $\mathbf{B}_0$.

(we call $\widetilde{\boldsymbol{\beta}}_\mathrm{n}$ is regular estimator of $\boldsymbol{\beta}_0$ if it is $\sqrt{\mathrm{n}}$-consistent and asymptotically Normal with $\sqrt{\mathrm{n}}\left(\widetilde{\boldsymbol{\beta}}_\mathrm{n} - \boldsymbol{\beta}_0\right) \xrightarrow{\mathrm{d}} \mathrm{N}_\mathrm{d}\left(\mathbf{0}, \mathbf{V}(\boldsymbol{\beta})\right)$ for all $\boldsymbol{\beta}$ in an open neighborhood of $\boldsymbol{\beta}_0$, i.e. $\sqrt{\mathrm{n}}$-consistency and asymptotic Normality of regular estimators are maintained for all $\boldsymbol{\beta}$ near $\boldsymbol{\beta}_0$)

Theory of Maximum Likelihood

e.g. Let $(S,\mathcal{A},P)$ be a probability space

Let $Z_1,\ldots,Z_n$ be random variables defined on $(S,\mathcal{A},P)$

Where $Z_1,\ldots,Z_n$ are iid Uniform $(0,\theta_0)$ for some $\theta_0>0$

The common density of the Z_is

$$f(z,\theta_0) \equiv \frac{1}{\theta_0} 1\{0 \le z \le \theta_0\}$$

So the likelihood function is

$$\hat{L}_n(\theta) \equiv \theta^{-1} \prod_{j=1}^{n} 1\{0 \le z_j \le \theta\}, \qquad \left(= \prod_{j=1}^{n} f(z_j,\theta) \right)$$

e.g. Let $\mathbf{Z}_j \equiv \begin{bmatrix} \mathbf{X}_j^\top & Y_j \end{bmatrix}^\top$ (j=1,…,n) be independent with

$$Y_j = \alpha_0 + \mathbf{X}_j^\top \boldsymbol{\beta}_0 + u_j, \qquad \text{where } u_j | \mathbf{X}_j \sim N(0,\sigma_0^2)$$

(i.e. Normal linear model)

The conditional density of Y_j given $\mathbf{X}_j$ is

$$f(y|\mathbf{X}_j;\boldsymbol{\theta}_0) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y-\alpha_0-\mathbf{X}_j^\top\boldsymbol{\beta}_0)^2}{2\sigma_0^2}}$$

Assume that the $\mathbf{X}_j$s are iid with absolutely continuous distribution having density $g(\cdot)$. Then the likelihood function

$$\begin{aligned} \hat{L}_n(\boldsymbol{\theta}) &= \prod_{j=1}^{n} f(Y_j|\mathbf{X}_j;\boldsymbol{\theta}) g(\mathbf{X}_j) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{j=1}^{n}(Y_j-\alpha-\mathbf{X}_j^\top\boldsymbol{\beta})^2}{2\sigma^2}} \prod_{j=1}^{n} g(\mathbf{X}_j) \end{aligned}$$

where $\boldsymbol{\theta} = [\alpha \quad \boldsymbol{\beta}^\top \quad \sigma^2]^\top$

Assuming that $g(\cdot)$ does not depend on $\boldsymbol{\theta}_0$ can work with the conditional likelihood function

$$\hat{L}_n^c(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(Y_j|\mathbf{X}_j;\boldsymbol{\theta})$$

In general, writing down a likelihood function is nontrivial.

- The data may not be iid.
- The joint distribution of the data may not be absolutely continuous or discrete.

Better to define a likelihood function implicitly.

Definition 1 (likelihood function) A sequence $\{\hat{L}_n(\boldsymbol{\theta})\text{: n=1,2,…}\}$ of a nonnegative random functions in $\theta\in\Theta$ is a sequence of likelihood functions iff

L.1. $\exists$ an increasing sequence $\{\mathcal{A}_n\text{: n} = 1,2,\dots\}$ of σ-algebras such that $\forall\ \theta\in\Theta$ and n≥1, $\hat{L}_n(\boldsymbol{\theta})$ is measurable $\mathcal{A}_n$

L.2. $\exists\ \boldsymbol{\theta}_0\in\Theta$ s.t. $\forall\ \theta\in\Theta$

$$\mathrm{P}\left[\mathrm{E}\left[\frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_n(\boldsymbol{\theta})}\middle|\mathcal{A}_0\right]\leq 1\right]=1$$

and $\forall$ n≥2

$$\mathrm{P}\left[\mathrm{E}\left[\frac{\frac{\hat{L}_n(\boldsymbol{\theta})}{\hat{L}_{n-1}(\boldsymbol{\theta})}}{\frac{\hat{L}_n(\boldsymbol{\theta}_0)}{\hat{L}_{n-1}(\boldsymbol{\theta}_0)}}\middle|\mathcal{A}_{n-1}\right]\leq 1\right]=1$$

L.3. $\forall\ \boldsymbol{\theta}_1\neq\boldsymbol{\theta}_2$,

$$\mathrm{P}\left[\hat{L}_1(\boldsymbol{\theta}_1)=\hat{L}_1(\boldsymbol{\theta}_2)\middle|\mathcal{A}_0\right]<1$$

and n≥2

$$\mathrm{P}\left[\frac{\hat{L}_n(\boldsymbol{\theta}_1)}{\hat{L}_{n-1}(\boldsymbol{\theta}_1)}=\frac{\hat{L}_n(\boldsymbol{\theta}_2)}{\hat{L}_{n-1}(\boldsymbol{\theta}_2)}\middle|\mathcal{A}_{n-1}\right]<1$$

(L3 rules out functions $\hat{L}_n(\boldsymbol{\theta})$ that are w.p. 1 on Θ)
L2, L3 guarantee that the point $\boldsymbol{\theta}_0\in\Theta$ (defined in L2) is unique.

Theorem 1 $\forall\ \boldsymbol{\theta}\in\Theta\backslash\{\boldsymbol{\theta}_0\}$ and $\forall$ n≥1,

$$\mathrm{E}\left[\log\left(\frac{\hat{L}_n(\boldsymbol{\theta})}{\hat{L}_n(\boldsymbol{\theta}_0)}\right)\right]<0$$

Proof Let n=1, we have

$$\log\left(\frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}\right)<\frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}-1,\qquad \text{by the property of } \log(\cdot)$$
$$\text{if } \frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}\neq 1$$

Let

$$Y(\boldsymbol{\theta}) \equiv \frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)} - \log\left(\frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}\right) - 1, \qquad X(\boldsymbol{\theta}) \equiv \frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}$$

Then

$$Y(\boldsymbol{\theta}) \geq 0 \text{ with } Y(\boldsymbol{\theta}) > 0 \text{ iff } X(\boldsymbol{\theta}) \neq 1$$

Suppose

$$P[E[Y(\boldsymbol{\theta})|\mathcal{A}_0] = 0] = 1$$

Then

$$P[Y(\theta) = 0|\mathcal{A}_0] = 1 \text{ a.s.}$$

because

$$Y(\boldsymbol{\theta}) \geq 0, \qquad > 0 \text{ in this case,} \qquad P[X(\boldsymbol{\theta}) = 1|\mathcal{A}_0] = 1 \text{ a.s.}$$

By L3 it cannot be true that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ so

$$P\left[E\left[\log\left(\frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}\right)\middle|\mathcal{A}_0\right] < 0\right] = 1 \text{ iff } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

So if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then

$$E\left[\log\left(\frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}\right)\right] < 0$$

A similar argument shows that if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then

$$E\left[\log\left(\frac{\hat{L}_n(\boldsymbol{\theta})}{\hat{L}_{n-1}(\boldsymbol{\theta})}\right) - \log\left(\frac{\hat{L}_n(\boldsymbol{\theta}_0)}{\hat{L}_{n-1}(\boldsymbol{\theta}_0)}\right)\right] < 0$$

Definition (likelihood function) A sequence $\{\hat{L}_n(\boldsymbol{\theta}): n=1,2,\ldots\}$ of non-negative random functions on a parameter space Θ is a sequence of likelihood functions iff

L1 $\exists$ increasing sequence $\{\mathcal{A}_n: n=0,1,2,\ldots\}$ of σ-algebras such that $\forall\ \boldsymbol{\theta}\in\Theta$ and $\forall\ n\geq 1$ $\hat{L}_n(\boldsymbol{\theta})$ is measurable $\mathcal{A}$.

L2 $\boldsymbol{\theta}_0\in\Theta$ such that

$$P\left[E\left[\frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}\middle|\mathcal{A}_0\right]\leq 1\right]=1$$

and $\forall\ n\geq 2$

$$P\left[E\left[\frac{\frac{\hat{L}_n(\boldsymbol{\theta})}{\hat{L}_{n-1}(\boldsymbol{\theta})}}{\frac{\hat{L}_n(\boldsymbol{\theta}_0)}{\hat{L}_{n-1}(\boldsymbol{\theta}_0)}}\middle|\mathcal{A}_{n-1}\right]\leq 1\right]=1$$

L3 $\forall\ \boldsymbol{\theta}_1\neq\boldsymbol{\theta}_2$, $P[\hat{L}_1(\boldsymbol{\theta}_1)=\hat{L}_1(\boldsymbol{\theta}_2)|\mathcal{A}_0]<1$, and $\forall\ n\geq 2$

$$P\left[\frac{\hat{L}_n(\boldsymbol{\theta}_1)}{\hat{L}_{n-1}(\boldsymbol{\theta}_1)}=\frac{\hat{L}_n(\boldsymbol{\theta}_2)}{\hat{L}_{n-1}(\boldsymbol{\theta}_2)}\middle|\mathcal{A}_{n-1}\right]<1$$

(L3 rules out functions $\hat{L}_n(\boldsymbol{\theta})$ that are almost surely constant in $\boldsymbol{\theta}$ on Θ)
L3 implies that the point $\boldsymbol{\theta}_0$ defined in L2 is unique.

Theorem 1 $\forall\ \boldsymbol{\theta}\in\Theta\backslash\{\boldsymbol{\theta}_0\}$ and $\forall\ n\geq 1$

$$E\left[\log\left(\frac{\hat{L}_n(\boldsymbol{\theta})}{\hat{L}_n(\boldsymbol{\theta}_0)}\right)\right]<0$$

Also, if the data are iid from an absolutely continuous distribution with density $f(\cdot;\boldsymbol{\theta})$ the assumption that

$$\operatorname{supp} f(\cdot;\boldsymbol{\theta})\equiv\left\{\mathbf{z}\in\mathbb{R}^k: f(\mathbf{z};\boldsymbol{\theta})>0\right\}$$

is the same for all $\boldsymbol{\theta}\in\Theta$ implies that

$$E\left[\frac{\hat{L}_n(\boldsymbol{\theta})}{\hat{L}_n(\boldsymbol{\theta}_0)}\right]=1$$

$\forall\ \theta\in\Theta$ and all $n\geq 1$; this implication also goes in the other direction.

Definition 2 (invariant support) A sequence $\{\hat{L}_n(\boldsymbol{\theta}); n=1,2,\ldots\}$ of likelihood functions has invariant support iff $\forall\ \boldsymbol{\theta}\in\Theta$

$$P\left[E\left[\frac{\hat{L}_1(\boldsymbol{\theta})}{\hat{L}_1(\boldsymbol{\theta}_0)}\middle|\mathcal{A}_0\right]=1\right]=1$$

and $\forall\ n\geq 2$

$$P\left[E\left[\frac{\frac{\hat{L}_n(\boldsymbol{\theta})}{\hat{L}_{n-1}(\boldsymbol{\theta})}}{\frac{\hat{L}_n(\boldsymbol{\theta}_0)}{\hat{L}_{n-1}(\boldsymbol{\theta}_0)}}\middle|\mathcal{A}_{n-1}\right]=1\right]=1$$

Note: Most likelihood functions in Econometrics have invariant support.

In general, L1–L3 allow one to treat MLE as special case of extremum estimation with

$$\boldsymbol{\theta}_0=\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\,E\left[\log\hat{L}_n(\boldsymbol{\theta})\right]$$

(recall: need L2, L3 for this)

This may require sometimes one to choose Θ sufficiently small.

e.g. Suppose $Z_j=\cos(X_j+\theta_0)$ where $X_1, X_2, \ldots$ are iid with an absolutely continuous distribution with density f.

Then the density of Zj has the form $f(z;\theta_0)=f(z;\theta_0+2\pi s)$ for any $s\in\mathbb{Z}\equiv$set of integers.

In this case, need Θ to be small, say $\Theta=[0,2\pi]$ in order to make θ_0 unique.

In addition: MLE cannot be an extremum estimator if $E[\log\hat{L}_n(\boldsymbol{\theta})]$ does not satisfy first- and second-order conditions for the existence of a global maximum at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ when $\boldsymbol{\theta}_0$ is in the interior of Θ.

However, suppose $\boldsymbol{\theta}_0$ is on the boundary of Θ.

e.g. $Z_1,\ldots,Z_n$ are iid Uniform $(0,\theta_0)$ for some $\theta_0>0$. Then we know

$$\hat{L}_n(\theta)=\frac{1}{\theta^n}\prod_{j=1}^{n}1\{Z_j\in[0,\theta]\}$$

We have

$$E\left[\log\hat{L}_n(\theta)\right]=\begin{cases}-\infty, & \theta<\theta_0\\ -n\log\theta, & \theta\geq\theta_0\end{cases}$$

The left derivative of $E[\log\hat{L}_n(\theta)]$ at $\theta=\theta_0$ is

$$\lim_{\delta\downarrow 0}\frac{1}{\delta}\left(E\left[\log\hat{L}_n(\theta_0)\right]-E\left[\log\hat{L}_n(\theta_0-\delta)\right]\right)=\infty$$

The right derivative of $E[\log\hat{L}_n(\theta)]$ at $\theta=\theta_0$ is

$$\lim_{\delta\downarrow 0}\frac{1}{\delta}\left(E\left[\log\hat{L}_n(\theta_0+\delta)\right]-E\left[\log\hat{L}_n(\theta_0)\right]\right)=-\frac{n}{\theta_0}$$

So $E[\log\hat{L}_n(\theta)]$ is not differentiable at θ_0.

Therefore the MLE $\hat{\theta}_n = Z_{(n)}$ of θ_0 does not behave as a standard extremum estimator; cannot derive its asymptotic properties from first and second order conditions.

Henceforth: Only consider MLEs for problems where θ_0 is defined in first- and second-order conditions.

Assumption 1 Θ is convex; $\boldsymbol{\theta}_0 \in \text{interior}\Theta$. Also $\hat{L}_n(\boldsymbol{\theta})$ is (w.p. 1) twice continuously differentiable in an open neighborhood Θ_0 of $\boldsymbol{\theta}_0$.

If $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ then for each $i_1, i_2 \in \{1,\ldots,d\}$ assume

$$E\left[\sup_{\theta\in\Theta_0}\left|\frac{\partial^2 \hat{L}_n(\boldsymbol{\theta})}{\partial\theta_{i_1}\partial\theta_{i_2}}\right|\right] < \infty, \qquad \cdots (22)$$

and

$$E\left[\sup_{\theta\in\Theta_0}\left|\frac{\partial^2 \log \hat{L}_n(\boldsymbol{\theta})}{\partial\theta_{i_1}\partial\theta_{i_2}}\right|\right] < \infty, \qquad \cdots (23)$$

Theorem 2 (score; information identities)

$$E\left[\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = \mathbf{0}, \qquad \cdots (\text{score identity})$$

$$E\left[\frac{\partial^2 \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = -\text{Var}\left[\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right], \qquad \cdots (\text{information identity})$$

(Note: Without the score identity, the MLE is inconsistent. If the score identity holds, but the information identity fails, the MLE is consistent but asymptotically inefficient)

Proof Case d=1 only. Also assume that $\mathbf{Z} \equiv [\mathbf{z}_1 \quad \cdots \quad \mathbf{z}_n]^\top$ is a random sample from an absolutely continuous distribution with density $f(\mathbf{z};\theta_0)$. Then

$$E\left[\frac{1}{n}\log \hat{L}_n(\theta)\right] = \frac{1}{n}\sum_{i=1}^{n} E[\log f(\mathbf{z}_1;\theta)] = \int (\log f(\mathbf{z};\theta)) f(\mathbf{z};\theta_0) d\mathbf{z}$$

By Taylor's theorem, $\forall\ \theta\in\Theta_0$ and $\forall\ \delta\neq 0$ with $\theta+\delta\in\Theta_0$, $\exists\ \lambda(\mathbf{z},\delta)\in[0,1]$ such that

$$\log f(\mathbf{z};\theta+\delta) = \log f(\mathbf{z};\theta) + \delta\frac{d}{d\theta}\log f(\mathbf{z};\theta) + \frac{\delta^2}{2}\frac{d^2}{d\tilde{\theta}^2}\log f(\mathbf{z};\tilde{\theta})\bigg|_{\tilde{\theta}=\theta+\lambda(z;\delta)\delta}$$

Since Θ is convex $\theta+\lambda(\mathbf{z},\delta)\delta\in\Theta$. By (23) and the dominated convergence theorem

$$\frac{d}{d\theta}\int (\log f(\mathbf{z};\theta)) f(\mathbf{z};\theta_0) d\mathbf{z} = \int \frac{d\log f(\mathbf{z};\theta)}{d\theta} f(\mathbf{z};\theta_0) d\mathbf{z}, \qquad \cdots (26)$$

By condition (22) and the dominated convergence theorem, we have

$$\int \frac{df(\mathbf{z};\theta)}{d\theta} d\mathbf{z} = \frac{d}{d\theta}\int f(\mathbf{z};\theta) d\mathbf{z} = \frac{d}{d\theta} 1 = 0, \qquad \cdots (27)$$

Also,

$$\int \frac{d \log f(\mathbf{z};\theta)}{d\theta}\bigg|_{\theta=\theta_0} f(\mathbf{z};\theta_0) d\mathbf{z} = \int \frac{1}{f(\mathbf{z};\theta)} \frac{df(\mathbf{z};\theta)}{d\theta}\bigg|_{\theta=\theta_0} f(\mathbf{z};\theta_0) d\mathbf{z}$$
$$= \frac{d}{d\theta}\int f(\mathbf{z};\theta_0) d\mathbf{z}, \qquad \cdots (26)$$

Combine (24)–(28) to get

$$E\left[\frac{d}{d\theta}\frac{1}{n}\log \hat{L}_n(\theta)\bigg|_{\theta=\theta_0}\right] = E\left[\frac{d \log f(\mathbf{z}_1;\theta)}{d\theta}\bigg|_{\theta=\theta_0}\right] = \int \left(\frac{d}{d\theta}\log f(\mathbf{z};\theta)\bigg|_{\theta=\theta_0}\right) f(\mathbf{z};\theta_0) d\mathbf{z} = 0$$

i.e. the score identity.

Next: Consider the information identity. By (22), (23) and the dominated convergence theorem we have

$$\frac{d^2}{d\theta^2}\int (\log f(\mathbf{z};\theta)) f(\mathbf{z};\theta_0) d\mathbf{z} = \int \left(\frac{d^2}{d\theta^2}\log f(\mathbf{z};\theta)\right) f(\mathbf{z};\theta_0) d\mathbf{z}, \qquad \cdots (29)$$

and

$$\int \frac{d^2 f(\mathbf{z};\theta)}{d\theta^2} d\mathbf{z} = \frac{d^2}{d\theta^2}\int f(\mathbf{z};\theta_0) d\mathbf{z} = 0, \qquad \cdots (30)$$

(29), (30) imply the information identity because

$$\int \left(\frac{d^2}{d\theta^2}\log f(\mathbf{z};\theta)\bigg|_{\theta=\theta_0}\right) f(\mathbf{z};\theta_0) d\mathbf{z} = \int \left(\frac{\frac{d^2}{d\theta^2} f(\mathbf{z};\theta)}{f(\mathbf{z};\theta)}\bigg|_{\theta=\theta_0}\right) f(\mathbf{z};\theta_0) d\mathbf{z}$$
$$- \int \left(\frac{\frac{d}{d\theta} f(\mathbf{z};\theta)}{f(\mathbf{z};\theta)}\right)^2 \bigg|_{\theta=\theta_0} f(\mathbf{z};\theta_0) d\mathbf{z}$$
$$= \int \frac{d^2}{d\theta^2} f(\mathbf{z};\theta)\bigg|_{\theta=\theta_0} d\mathbf{z} - \int \left(\frac{d}{d\theta}\log f(\mathbf{z};\theta)\right)^2 \bigg|_{\theta=\theta_0} f(\mathbf{z};\theta_0) d\mathbf{z}$$
$$= -\int \left(\frac{d}{d\theta}\log f(\mathbf{z};\theta)\bigg|_{\theta=\theta_0}\right)^2 f(\mathbf{z};\theta_0) d\mathbf{z}$$
$$= -\mathrm{Var}\left[\frac{d}{d\theta}\log f(\mathbf{z}_1;\theta)\bigg|_{\theta=\theta_0}\right] \text{ by the score identity}$$

i.e. the information identity.

In general, the (d×d) matrix

$$\mathbf{H} \equiv \mathrm{Var}\left[\frac{\partial}{\partial \boldsymbol{\theta}^{\top}} \log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

is the Fischer information (matrix) for $\boldsymbol{\theta}_0$.
(Recall: If θ is scalar and the observations are iid, then H^{-1} is the Cramér–Rao lower bound for the variance of an unbiased estimator of θ_0)

Consistency and Asymptotic Normality of MLEs

- Consider MLE as a special case of extremum estimation
- Need to assume that

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\Theta}{\mathrm{argmax}}\, \mathrm{E}\left[\log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta})\right]$$

can be characterized in terms of first- and second-order conditions.

Also need the following.

Assumption 2 As n→∞

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{\mathrm{n}}\log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta}) - \mathrm{E}\left[\log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta})\right]\right| \xrightarrow{\mathrm{p}} 0$$

and

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{\mathrm{n}}\mathrm{E}\left[\log\left(\frac{\hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta})}{\hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta}_0)}\right)\right] - \ell(\boldsymbol{\theta};\boldsymbol{\theta}_0)\right| \to 0$$

for some continuous function $\ell(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}$ s.t. ∀ δ>0

$$\sup_{\boldsymbol{\theta}\in\Theta,\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\geq\delta} \ell(\boldsymbol{\theta};\boldsymbol{\theta}_0) < 0, \qquad \cdots(*)$$

Theorem 3 (consistency) Under assumptions 1, 2

$$\widehat{\boldsymbol{\theta}}_{\mathrm{n}} \xrightarrow{\mathrm{p}} \boldsymbol{\theta}$$

Note: Assumption 2 can only be verified in the context of specific examples.

- Can consider conditions that suffice for a ULLN: e.g. Θ compact, $\log\hat{\mathrm{L}}\mathrm{n}(\theta)$ to be continuous and uniformly bounded on Θ w.p. 1
- Condition (*) in Assumption 2 will hold if Θ is compact and $\ell(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ is continuous on Θ and $\boldsymbol{\theta}_0$ is unique ($\boldsymbol{\theta}_0$ is unique by Theorem 1; Theorem 1 holds by L2, L3)

Recall certain assumptions that allow one to treat MLE as "just another" M-estimator

Assumption 1

- The parameter space Θ is convex
- $\boldsymbol{\theta}_0 \in \text{int}\Theta$
- The likelihood function $\hat{L}_n(\boldsymbol{\theta})$ is (w.p. 1) twice continuously differentiable in an open neighborhood Θ_0 of $\boldsymbol{\theta}_0$
- If $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, then for each pair $i_1, i_2 \in \{1,\ldots,d\}$

$$E\left[\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\frac{\partial^2 \hat{L}_n(\boldsymbol{\theta})}{\partial\theta_{i_1}\partial\theta_{i_2}}\right|\right] < \infty, \qquad \text{and } E\left[\sup_{\boldsymbol{\theta}\in\Theta_0}\left|\frac{\partial^2 \ln \hat{L}_n(\boldsymbol{\theta})}{\partial\theta_{i_1}\partial\theta_{i_2}}\right|\right] < \infty$$

(Recall that Assumption 1 implies the score and information identities: score identity→ consistency, information identity → asymptotic efficiency)

Assumption 2 As $n\to\infty$

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\log \hat{L}_n(\boldsymbol{\theta}) - E\left[\frac{1}{n}\log \hat{L}_n(\boldsymbol{\theta})\right]\right| \xrightarrow{p} 0, \qquad \text{and } \sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}E\left[\log\left(\frac{\hat{L}_n(\boldsymbol{\theta})}{\hat{L}_n(\boldsymbol{\theta}_0)}\right)\right] - \ell(\boldsymbol{\theta};\boldsymbol{\theta}_0)\right| \to 0$$

for some continuous function $\ell(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}$ s.t. for any $\delta>0$,

$$\sup_{\boldsymbol{\theta}\in\Theta, \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\geq\delta} \ell(\boldsymbol{\theta};\boldsymbol{\theta}_0) < 0, \qquad \cdots(*)$$

Theorem 3 (consistency of the MLE) Assumption 2 implies

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$$

Note

- Assumption 2 can replace Assumption 1 if we only care about consistency of the MLE
- Can verify Assumption 2 for specific cases by checking that the two conditions hold
  - The ULLN holds if Θ is compact, $\log\hat{L}_n(\boldsymbol{\theta})$ is continuous
  - (*) holds if Θ is compact, $\ell(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ is continuous on Θ and $\boldsymbol{\theta}_0$ is unique

Assumption 3 As $n\to\infty$, for each pair $i_1, i_2 \in \{1,\ldots,d\}$

$$\underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{\partial^2 \log \hat{L}_n(\boldsymbol{\theta})}{\partial\theta_{i_1}\partial\theta_{i_2}} - E\left[\frac{\partial^2 \frac{1}{n}\log \hat{L}_n(\boldsymbol{\theta})}{\partial\theta_{i_1}\partial\theta_{i_2}}\right]\right| \xrightarrow{p} 0}_{\cdots(32)}, \qquad \text{and } \underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\left|E\left[\frac{\partial^2 \frac{1}{n}\log \hat{L}_n(\boldsymbol{\theta})}{\partial\theta_{i_1}\partial\theta_{i_2}}\right] - h_{i_1,i_2}(\theta)\right| \to 0}_{\cdots(33)}$$

for some function $h_{i_1,i_2}(\boldsymbol{\theta})$ that is continuous at $\boldsymbol{\theta}_0$.

In addition, the (d×d)-matrix $\bar{\mathbf{H}}$ with

$$\underbrace{\bar{\mathbf{H}}_0}_{(d\times d)} = \left[h_{i_1,i_2}(\theta_0)\right] \text{ is nonsingular,} \qquad \text{and } \frac{\partial \frac{1}{\sqrt{n}} \log \hat{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0^\top} \xrightarrow{d} N_d(\mathbf{0}, \bar{\mathbf{H}}_0)$$

Note: If $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ (possibly implied by Assumption 2) then Assumption 1, 3 imply that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_d(\mathbf{0}, \bar{\mathbf{H}}_0^{-1})$$

Note: $\bar{\mathbf{H}}_0 = \lim_{n\to\infty} \frac{1}{n}\mathbf{H}_n$ where $\mathbf{H}_n$ is the Fisher information

$$\mathbf{H}_n = \mathrm{Var}\left[\left.\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

Note: Verifying (32) implies verification of a ULLN
Note: (33) allows for non-iid data. In the case of iid data we simply have

$$h_{i_1,i_2} = -\frac{1}{n} E\left[\frac{\partial^2 \log \hat{L}_n(\boldsymbol{\theta})}{\partial \theta_{i_1} \partial \theta_{i_2}}\right]$$

Theorem 4 (asymptotic Normality of the MLE) Suppose $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ then Assumption 1, 3 imply $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_d(\mathbf{0}, \bar{\mathbf{H}}_0^{-1})$

Proof By the mean-value theorem for each $i\in\{1,\ldots,d\}$ $\exists\, \hat{\lambda}_i \in (0,1)$ such that

$$\left.\frac{1}{\sqrt{n}}\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = \left.\frac{1}{\sqrt{n}}\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \left.\frac{1}{n}\frac{\partial^2 \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0+\hat{\lambda}_i(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})} \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$$

By FOC for log-likelihood maximization (this requires $\boldsymbol{\theta}_0 \in \mathrm{int}\Theta$ and other conditions of Assumption 1), we have

$$\left.\frac{1}{\sqrt{n}}\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = 0, \qquad \cdots(36)$$

Since Θ is convex (Assumption 1) the mean value satisfies $\boldsymbol{\theta}_0 + \hat{\lambda}_i\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \in \Theta$, since $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, the mean value is also consistent for $\boldsymbol{\theta}_0$ and

$$\tilde{\mathbf{H}}_n \equiv \frac{1}{n}\begin{bmatrix} \left.\frac{\partial^2 \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \theta_1}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0+\hat{\lambda}_1(\boldsymbol{\theta}_n-\boldsymbol{\theta}_0)} \\ \vdots \\ \left.\frac{\partial^2 \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \theta_d}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0+\hat{\lambda}_d(\boldsymbol{\theta}_n-\boldsymbol{\theta}_0)} \end{bmatrix} \xrightarrow{p} \mathbf{H}_0$$

by the continuous mapping theorem (CMT hereafter) and information identity (implied by Assumption 1).

Since $\bar{\mathbf{H}}_0$ is nonsingular (Assumption 3) then by the CMT

$$\tilde{\mathbf{H}}_n^{-1} \xrightarrow{p} \bar{\mathbf{H}}_0^{-1}, \qquad \cdots (38)$$

By the original mean value expansion and (36), we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) = \tilde{\mathbf{H}}_n^{-1} \frac{1}{\sqrt{n}} \frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

So by (38) and Slutsky's theorem

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_d(\mathbf{0}, \bar{\mathbf{H}}_0^{-1})$$

(obviously used the assumption that)

$$\frac{1}{\sqrt{n}} \frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{H}}_0)$$

Note: Suppose iid data, i.e. suppose $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ are iid k variate with common density $f(\mathbf{z};\boldsymbol{\theta}_0)$.

In this case, Assumption 1 implies

$$\frac{1}{n} E\left[\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = E\left[\frac{\partial \log f(\mathbf{Z}_1;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = \mathbf{0}$$

and

$$\frac{1}{n} \mathrm{Var}\left[\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = \mathrm{Var}\left[\frac{\partial \log f(\mathbf{Z}_1;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = \bar{\mathbf{H}}_0$$

In this case, it is clear that

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}^\top} \log \hat{L}_n(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}^\top} \log f(\mathbf{Z}_i;\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{d} N_d(\mathbf{0}, \bar{\mathbf{H}}_0)$$

Stylized fact: Subject to regularity conditions (e.g. consistency, Assumption 1 & 3) the MLE is asymptotically efficient over the class of consistent and asymptotically Normal extremum estimators of $\boldsymbol{\theta}_0$.

Consider

$$\tilde{\boldsymbol{\theta}}_n \equiv \underset{\boldsymbol{\theta}\in\Theta}{\mathrm{argmax}} \frac{1}{n} \sum_{j=1}^{n} g\left(\mathbf{Z}_j;\boldsymbol{\theta}\right), \qquad \text{i. e. an extremum estimator of}$$

$$\boldsymbol{\theta}_0 \equiv \underset{\boldsymbol{\theta}\in\Theta}{\mathrm{argmax}}\, E[g(\mathbf{Z}_1,\boldsymbol{\theta})], \qquad \text{where } \mathbf{Z}_1,\ldots,\mathbf{Z}_n \text{ are iid k vectors}$$

from an absolutely continuous distribution with density $f(\mathbf{z};\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0 \in \mathrm{int}\Theta$, where $\Theta \subset \mathbb{R}^d$ and where Θ is convex.

Then $\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_d(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1})$ (subject to the right conditions on g(.,.))

$$\text{where } \mathbf{A}_0 = E\left[\left.\frac{\partial^2 g(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

$$\mathbf{B}_0 = E\left[\left.\frac{\partial g(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\frac{\partial g(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

Then $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1} - \bar{\mathbf{H}}_0^{-1}$ is positive semidefinite, i.e. the MLE standard errors are no larger than the $\widetilde{\boldsymbol{\theta}}_n$ standard errors.

To see this: By Assumption 1 and the FOC for

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, E[g(\mathbf{Z}_1, \boldsymbol{\theta})]$$

We have

$$\int \left.\frac{\partial g(\mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f(\mathbf{z}; \boldsymbol{\theta}_0) d\mathbf{z} = \mathbf{0}, \qquad \cdots (49)$$

(obviously) (49) holds regardless of the actual value of $\boldsymbol{\theta}_0$, so

$$\int \frac{\partial g(\mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} f(\mathbf{z}; \boldsymbol{\theta}) d\mathbf{z} = \mathbf{0}, \qquad \forall \boldsymbol{\theta} \in \Theta, \qquad \cdots (50)$$

Under the conditions of Assumption 1, we have

$$\begin{aligned}\mathbf{0} &= \int \frac{\partial^2 g(\mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} f(\mathbf{z}; \boldsymbol{\theta}) d\mathbf{z} + \int \frac{\partial g(\mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\frac{\partial f(\mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} d\mathbf{z} \\ &= \int \frac{\partial^2 g(\mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} f(\mathbf{z}; \boldsymbol{\theta}) d\mathbf{z} + \int \frac{\partial g(\mathbf{z}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\frac{\partial \log f(\mathbf{z}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\mathbf{z}; \boldsymbol{\theta}) d\mathbf{z}, \qquad \cdots (51)\end{aligned}$$

Let $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, then from (51) we have

$$E\left[\left.\frac{\partial g(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\frac{\partial \log f(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = -\mathbf{A}_0, \qquad \cdots (52)$$

But $\left.\frac{\partial g(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ has mean zero by (49). While $\left.\frac{\partial \log f(\mathbf{Z}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ also has mean zero by the score identity. Therefore, (52) is equivalent to

$$\operatorname{Cov}\left[\left.\frac{\partial g(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \left.\frac{\partial \log f(\mathbf{Z}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = -\mathbf{A}_0$$

So by Assumption 3 we have

$$\mathrm{Var}\left[\underbrace{\begin{bmatrix} \left.\frac{\partial g(\mathbf{Z}_1,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ \left.\frac{\partial \log f(\mathbf{Z}_1;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \end{bmatrix}}_{2d\times 1}\right] = \underbrace{\begin{bmatrix} \mathbf{B}_0 & -\mathbf{A}_0 \\ -\mathbf{A}_0 & \bar{\mathbf{H}}_0 \end{bmatrix}}_{2d\times 2d}$$

which (by virtue of being a covariance matrix) is positive semidefinite. Therefore

$$\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1} - \bar{\mathbf{H}}_0^{-1} = [\mathbf{A}_0^{-1} \quad \bar{\mathbf{H}}_0^{-1}] \begin{bmatrix} \mathbf{B}_0 & -\mathbf{A}_0^{-1} \\ -\mathbf{A}_0^{-1} & \bar{\mathbf{H}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{A}_0^{-1} \\ \bar{\mathbf{H}}_0^{-1} \end{bmatrix}$$

is also positive semidefinite.
(This argument also requires one to check that)

$$\frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{j=1}^{1} \left.\frac{\partial g(\mathbf{Z}_j,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ \left.\frac{\partial \log \hat{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \end{bmatrix} \xrightarrow{d} N_{2d}\left(\mathbf{0}, \begin{bmatrix} \mathbf{B}_0 & -\mathbf{A}_0 \\ -\mathbf{A}_0 & \bar{\mathbf{H}}_0 \end{bmatrix}\right)$$

Estimators based on estimating Equations (i.e. Z-estimators)

- An extremum estimator maximizes a criterion function $\hat{Q}_n(\boldsymbol{\theta})$ globally over a compact set Θ
- This may be impractical
- It may be different to maximize $\hat{Q}_n(\boldsymbol{\theta})$ globally, if indeed such a global maximum exists (exception: $\hat{Q}_n(\boldsymbol{\theta})$ known to be globally concave)
- It will be difficult to prove that $\bar{Q}(\boldsymbol{\theta})\equiv E[\hat{Q}_n(\boldsymbol{\theta})]$ has a unique global maximum on Θ at $\boldsymbol{\theta}_0$
- Specifying Θ as compact may be implausible

e.g. Suppose that

$$E[Y_i|\mathbf{X}_i] = e^{\mathbf{X}_i^\top \boldsymbol{\theta}_0}$$

Why should $\boldsymbol{\theta}_0$ be restricted to lie in a compact subset of $\mathbb{R}^d$?

- Proving asymptotic Normality is easiest if $\hat{\boldsymbol{\theta}}_n$ is simply a local maximizer of $\hat{Q}_n(\boldsymbol{\theta})$

For all these reasons, it may be better to consider estimators defined as the solutions to FOCs for maximization of $\hat{Q}_n(\boldsymbol{\theta})$.

$$\text{Let } \hat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{Z}_i, \boldsymbol{\theta})$$
$$\text{where } \mathbf{Z}_i = [\mathbf{X}_i^\top \quad Y_i]^\top$$
$$\text{and } \boldsymbol{\theta} \in \mathbb{R}^d$$
$$\text{Assume } \{\mathbf{Z}_i : i = 1, \dots, n\} \text{ is iid}$$
$$\text{Let } \bar{Q}(\boldsymbol{\theta}) \equiv E\left[\hat{Q}_n(\boldsymbol{\theta})\right] = E[g(\mathbf{Z}_1, \boldsymbol{\theta})]$$

Theorem 1 Assume

(1) $\Theta \subset \mathbb{R}^d$ is an open set so $\boldsymbol{\theta}_0 \in \text{int}\Theta$

(2) $g(\mathbf{Z}_1,\boldsymbol{\theta})$ is Borel measurable $\forall \boldsymbol{\theta}\in\Theta$, $\frac{\partial g}{\partial \boldsymbol{\theta}^\top}$ exists and is continuous $\forall \boldsymbol{\theta}\in N_1(\boldsymbol{\theta}_0)$, where $N_1(\boldsymbol{\theta})$ is an open neighborhood of $\boldsymbol{\theta}_0$ (so $\hat{Q}_n(\boldsymbol{\theta})$ is continuous $\forall \boldsymbol{\theta}\in N_1(\boldsymbol{\theta}_0)$)

(3) There is an open neighborhood $N_2(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ s.t.

$$\sup_{\boldsymbol{\theta}\in N_2(\boldsymbol{\theta}_0)} \left|\hat{Q}_n(\boldsymbol{\theta}) - \bar{Q}(\boldsymbol{\theta})\right| \xrightarrow{p} 0$$

and where $\boldsymbol{\theta}_0$ is the unique (local) maximizer of $\bar{Q}(\boldsymbol{\theta})$ on $N_2(\boldsymbol{\theta}_0)$

Let Θ_n be the set of solutions to $\frac{\partial \hat{Q}_n}{\partial \boldsymbol{\theta}^\top} = \mathbf{0}$

(i.e. Θ_n be the set of local extrema of $\hat{Q}_n(\boldsymbol{\theta})$)

(if $\Theta_n=\emptyset$, then define $\Theta_n\equiv\{0\}$)

Then

$$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} P\left[\inf_{\boldsymbol{\theta}\in\Theta_n} (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0) > \varepsilon\right] = \lim_{n\to\infty} P\left[\inf_{\boldsymbol{\theta}\in\Theta_n} \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 > \varepsilon\right] = 0$$

(Therefore there is a consistent root for the system $\frac{\partial \hat{Q}_n}{\partial \boldsymbol{\theta}^\top} = \mathbf{0}$)

Proof Let $S \subset N_1(\boldsymbol{\theta}_0) \cap N_2(\boldsymbol{\theta}_0)$ be compact. Let $\tilde{\boldsymbol{\theta}}_n$ be the maximizer of $\hat{Q}_n(\boldsymbol{\theta})$ on S. Then $\tilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ (see Theorem 2 in the handout on extremum estimation). But since $\hat{Q}_n(\boldsymbol{\theta})$ attains a (local) maximum at $\hat{\boldsymbol{\theta}}_n$ with probability approaching 1 as $n \to \infty$, we have

$$\lim_{n\to\infty} P\left[\tilde{\boldsymbol{\theta}}_n \in \Theta_n\right] = 1$$

A consistent root of $\frac{\partial \hat{Q}_n}{\partial \boldsymbol{\theta}^\top} = \mathbf{0}$ is asymptotically Normal under additional conditions.

Theorem 2 Assume the conditions of Theorem 1 and the following additional conditions

(4) $\frac{\partial^2 \hat{Q}_n}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}$ exists and is continuous on an open convex neighborhood of $\boldsymbol{\theta}_0$

(5)

$$\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 g(\mathbf{Z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_n} \xrightarrow{p} \mathbf{A}_0 \equiv E\left[\left.\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} g(\mathbf{Z}_1, \boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

Where $\mathbf{A}_0$ is finite and nonsingular for any $\tilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$

(6)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left.\frac{\partial g(\mathbf{Z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{d} N(\mathbf{0}, \mathbf{B}_0)$$

$$\text{where } \mathbf{B}_0 = E\left[\left.\frac{\partial g(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left.\frac{\partial g(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

Let $\{\hat{\boldsymbol{\theta}}_n : n = 1,2,\ldots\}$ be a sequence formed by taking one element from Θ_n for each n such that $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$. Then $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N_d(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1})$

Proof Take a Taylor expansion to get

$$(0 =) \left.\frac{\partial \hat{Q}_n}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} = \left.\frac{\partial \hat{Q}_n}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \left.\frac{\partial^2 \hat{Q}_n}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

Where $\boldsymbol{\theta}_n^*$ is a convex combination of $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. But $\hat{\boldsymbol{\theta}}_n \in \Theta_n$, so

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\left(\left.\frac{\partial^2 \hat{Q}_n}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*}\right)^{-1} \sqrt{n} \left.\frac{\partial \hat{Q}_n}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

But $\boldsymbol{\theta}_n^* \xrightarrow{p} \boldsymbol{\theta}_0$, so by Assumptions (4), (5)

$$\left.\frac{\partial^2 \widehat{Q}_n}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*} \xrightarrow{p} \mathbf{A}_0$$

The conclusion follows from Assumption (6) and Slutsky's theorem.

Note: Assumption (5) in Theorem 2 is usually difficult to verify.

Consider two results using different assumptions that imply Assumption (5).

Let $\mathbf{Z} \equiv [Z_1 \ \ldots \ Z_n]^\top$. Define $g_n(\boldsymbol{\theta}) \equiv g(\mathbf{Z}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$, suppose $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, where Θ open.

Assume that $g(\mathbf{Z}, \boldsymbol{\theta})$ is Borel measurable $\forall \boldsymbol{\theta} \in \Theta$.

Theorem 3 Suppose Θ is convex and that $\frac{\partial g_n}{\partial \boldsymbol{\theta}}$ exists $\forall \boldsymbol{\theta} \in \Theta$. Suppose that $\forall \varepsilon > 0 \ \exists M_2$ s.t. for each θ_j (where θ_j is the j-th element of $\boldsymbol{\theta}$)

$$P\left[\left|\frac{\partial g_n}{\partial \theta_j}\right| < M_2\right] \geq 1 - \varepsilon$$

$\forall n \in \{1,2,\ldots\}$ and $\forall \boldsymbol{\theta} \in \Theta$. Then

$$\left|g_n(\widehat{\boldsymbol{\theta}}_n) - g_n(\boldsymbol{\theta}_0)\right| \xrightarrow{p} 0, \qquad \text{if } \widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0 \in \Theta$$

Proof Take a Taylor expansion

$$g_n(\widehat{\boldsymbol{\theta}}_n) = g_n(\boldsymbol{\theta}_0) + \left.\frac{\partial g_n}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

Where $\boldsymbol{\theta}_n^*$ is between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. By convexity, $\boldsymbol{\theta}_n^* \in \Theta$.

Theorem 4 Suppose $\exists$an open neighborhood $N(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ such that

$$\sup_{\boldsymbol{\theta} \in N(\boldsymbol{\theta}_0)} \left|g_n(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}_0}[g_n(\boldsymbol{\theta})]\right| \xrightarrow{p} 0, \qquad (*)$$

Then, $\left|g_n(\widehat{\boldsymbol{\theta}}_n) - E_{\theta_0}[g_n(\boldsymbol{\theta}_0)]\right| \xrightarrow{p} 0$ if $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ and $E_{\boldsymbol{\theta}_0}[g_n(\boldsymbol{\theta})]$ is continuous at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$.

Proof Because of the ULLN in (*) and also because $\widehat{\boldsymbol{\theta}}_n \in N(\boldsymbol{\theta}_0)$ with probability approaching 1 as $n \to \infty$, then $\forall \varepsilon > 0$, $\forall \delta > 0$, $\exists M_1$ s.t. $\forall n \geq M_1$,

$$P\left[\left|g_n(\widehat{\boldsymbol{\theta}}_n) - E_{\boldsymbol{\theta}_0}[g_n(\widehat{\boldsymbol{\theta}}_n)]\right| \geq \frac{\varepsilon}{2}\right] \leq \frac{\delta}{2}, \qquad \cdots (\dagger)$$

(exercise: supply details)

Because $E_{\boldsymbol{\theta}_0}[g_n(\boldsymbol{\theta})]$ is continuous at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, so

$$\left|E_{\boldsymbol{\theta}_0}[g_n(\widehat{\boldsymbol{\theta}}_n)] - E_{\boldsymbol{\theta}_0}[g_n(\boldsymbol{\theta}_0)]\right| \xrightarrow{p} 0, \qquad \text{by the continuous mapping theorem}$$

So $\forall\varepsilon>0\ \forall\delta>0$, $\exists M_2$ s.t. $\forall n\geq M_2$,

$$P\left[\left|E_{\boldsymbol{\theta}_0}\left[g_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right]-E_{\boldsymbol{\theta}_0}[g_n(\boldsymbol{\theta}_0)]\right|\geq\frac{\varepsilon}{2}\right]\leq\frac{\delta}{2},\qquad\cdots(\dagger\dagger)$$

From (†), (††), we have $\forall n\geq\max\{M_1,M_2\}$

$$P\left[\left|g_n\left(\widehat{\boldsymbol{\theta}}_n\right)-g_n(\boldsymbol{\theta}_0)\right|\leq\varepsilon\right]\geq 1-\delta$$

Note: $E_{\boldsymbol{\theta}_0}\left[g_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right]\approx\frac{1}{n}\sum_{i=1}^{n}E_{\boldsymbol{\theta}_0}\left[g_n\left(\widehat{\boldsymbol{\theta}}_n\right)\middle|\widehat{\boldsymbol{\theta}}_n\right]$

Exercise: Use Theorem 1, 2 in this lecture to rederive the consistency and asymptotic Normality of estimators defined as solutions of

$$\frac{\partial}{\partial\boldsymbol{\theta}^\top}\log\hat{L}_n(\boldsymbol{\theta})=\mathbf{0}$$

Where $\hat{L}_n(\boldsymbol{\theta})$ satisfies L1–L3.

Quasi-Maximum Likelihood (also known as Pseudo-Maximum Likelihood)

- In practice, one can never be sure that an MLE actually maximizes a random function satisfying L1–L3
- This can be serious because, misspecification of $\hat{L}_n(\boldsymbol{\theta})$ may cause the score identity to fail, leading the MLE to be inconsistent
- On the other hand, a quasi-MLE, defined as the maximizer of a possibly misspecified log-likelihood will generally be consistent for the pseudo-true value, i.e. for

$$\overline{\boldsymbol{\theta}}_0\equiv\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\,E_{\boldsymbol{\theta}_0}\left[\frac{1}{n}\log\hat{L}_n(\boldsymbol{\theta})\right],\qquad\cdots(*)$$

where the expectation is taken with respect to the time data density indexed by $\boldsymbol{\theta}_0$. From (*), we see that if $\hat{L}_n(\boldsymbol{\theta})$ is correctly satisfied, then $\overline{\boldsymbol{\theta}}_0=\boldsymbol{\theta}_0$; otherwise the quasi-MLE is inconsistent.

Huber (1962), White (1982) showed that a quasi-MLE given as

$$\widehat{\boldsymbol{\theta}}_{\mathrm{QML}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \frac{1}{\mathrm{n}} \log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta})$$

is consistent for the pseudo-true value, i.e. for

$$\overline{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \mathrm{E}_{\boldsymbol{\theta}_0}\left[\frac{1}{\mathrm{n}} \log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta})\right]$$

(where the expectation is taken w.r.t. the true joint data density indexed by $\boldsymbol{\theta}_0$)

In fact, we have

$$\sqrt{\mathrm{n}}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{QML}} - \overline{\boldsymbol{\theta}}_0\right) \xrightarrow{\mathrm{d}} \mathrm{N}_{\mathrm{d}}(\mathbf{0}, \overline{\mathbf{A}}_0^{-1}\overline{\mathbf{B}}_0\overline{\mathbf{A}}_0^{-1})$$

where

$$\overline{\mathbf{A}}_0 = \mathrm{E}_{\boldsymbol{\theta}_0}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\top}} \log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right], \qquad \overline{\mathbf{B}}_0 = \mathrm{E}_{\boldsymbol{\theta}_0}\left[\frac{\partial}{\partial\boldsymbol{\theta}^{\top}} \log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta}) \frac{\partial}{\partial\boldsymbol{\theta}} \log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

If $\overline{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$, i.e. the likelihood is correctly specified, then by the information identity

$$\sqrt{\mathrm{n}}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{QML}} - \boldsymbol{\theta}_0\right) \xrightarrow{\mathrm{d}} \mathrm{N}_{\mathrm{d}}(\mathbf{0}, \overline{\mathbf{A}}_0^{-1})$$

How to interpret $\overline{\boldsymbol{\theta}}_0$?

White (1982) proposed the interpretation.

- Let $\mathrm{f}(\mathbf{z};\boldsymbol{\theta})$ denote the assumed joint density of $\mathbf{Z}_1,\ldots,\mathbf{Z}_{\mathrm{n}}$ (where $\boldsymbol{\theta}\in\Theta$)
- Let $\mathrm{g}(\mathbf{z})$ denote the true but unknown density of $\mathbf{Z}_1,\ldots,\mathbf{Z}_{\mathrm{n}}$

Then the Kullback–Leibler information criterion (KLIC) as

$$\mathrm{KL}(\mathrm{g};\mathrm{f}) \equiv \mathrm{E}_{\mathrm{g}}\left[\log\left(\frac{\mathrm{g}(\mathbf{z})}{\mathrm{f}(\mathbf{z};\boldsymbol{\theta})}\right)\right] = \mathrm{E}_{\mathrm{g}}[\log \mathrm{g}(\mathbf{z})] - \mathrm{E}_{\mathrm{g}}[\log \mathrm{f}(\mathbf{z};\boldsymbol{\theta})]$$

(the expectation w.r.t. $\mathrm{g}(\mathbf{z})$)

The KLIC is minimized at 0 if $\exists\boldsymbol{\theta}_0$ s.t.

$$\mathrm{g}(\mathbf{z}) = \mathrm{f}(\mathbf{z};\boldsymbol{\theta}_0), \qquad \forall\mathbf{z}$$

i.e. the model $\{\mathrm{f}(\mathbf{z};\boldsymbol{\theta}):\boldsymbol{\theta}\in\Theta\}$ is correctly specified White (1982) showed that the QMLE $\widehat{\boldsymbol{\theta}}_{\mathrm{QML}}$ minimizes

$$\mathrm{KL}\big(\mathrm{g};\mathrm{f}(\cdot;\boldsymbol{\theta})\big), \qquad \text{over } \boldsymbol{\theta}\in\Theta$$

because

$$\frac{1}{\mathrm{n}} \log \hat{\mathrm{L}}_{\mathrm{n}}(\boldsymbol{\theta}) \xrightarrow{\mathrm{p}} \mathrm{E}_{\mathrm{g}}[\log \mathrm{f}(\mathbf{z};\boldsymbol{\theta})], \qquad \text{under suitable conditions}$$

and $\mathrm{E}_{\mathrm{g}}[\log\mathrm{f}(\mathbf{z};\boldsymbol{\theta})]$ is maximized over Θ by $\overline{\boldsymbol{\theta}}_0$, i.e. $\overline{\boldsymbol{\theta}}_0$ minimizes $\mathrm{KL}\big(\mathrm{g};\mathrm{f}(\cdot;\boldsymbol{\theta})\big)$ over $\boldsymbol{\theta}\in\Theta$.

Therefore, $\widehat{\boldsymbol{\theta}}_{\text{QML}}$ is consistent for the minimizer of $\text{KL}\big(\text{g};\text{f}(\cdot;\boldsymbol{\theta})\big)$, which "may be useful" if $\text{KL}\big(\text{g};\text{f}(\cdot;\boldsymbol{\theta})\big)$ is small.

Next: Nonlinear least squares

Suppose we observe $[\mathbf{X}_i^\top \, Y_i]^\top$ (i=1,…,n) iid, where $\mathbf{X}_i$ is k-variate.

Assume $\exists\boldsymbol{\beta}\in\mathbb{R}^d$ and a function $g:\mathbb{R}^k\times\mathbb{R}^d\to\mathbb{R}$ s.t. $E[Y_i|\mathbf{X}_i]=g(\mathbf{X}_i;\boldsymbol{\beta}_0)$ w.p. 1, i.e. a nonlinear regression model.

In this case, a natural estimator of $\boldsymbol{\beta}_0$ is the nonlinear least squares (NLLS) estimator.

$$\widetilde{\boldsymbol{\beta}}_{\text{NLLS}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{argmax}}\, \widehat{\text{Q}}_n(\boldsymbol{\beta})$$

where

$$\widehat{\text{Q}}_n(\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^{n}\big(Y_i - g(\mathbf{X}_i,\boldsymbol{\beta})\big)^2$$

Somewhat confusingly, "nonlinear least squares" usually refers to the following Z-estimator of $\boldsymbol{\beta}_0$:

$$\left.\frac{\partial\widehat{\text{Q}}_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{NLLS}}} = \mathbf{0}$$

Under the conditions for Z-estimation $\widehat{\boldsymbol{\beta}}_{\text{NLLS}}$ is $\sqrt{n}$-consistent with

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_{\text{NLLS}} - \boldsymbol{\beta}_0\big) \xrightarrow{d} \text{N}_d(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1})$$

where

$$\mathbf{A}_0 = -\text{E}\left[\frac{\partial g(\mathbf{X}_1,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\frac{\partial g(\mathbf{X}_1,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right], \qquad \mathbf{B}_0 = \text{E}\left[u_1^2\frac{\partial g(\mathbf{X}_1,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\frac{\partial g(\mathbf{X}_1,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]$$

where $u_1=Y_1-g(\mathbf{X}_1,\boldsymbol{\beta}_0)$

(Exercise: Fill in the details and give specific assumptions)

e.g. Suppose linear regression, i.e. k=d,

$$g(\mathbf{x},\boldsymbol{\beta}_0) = \mathbf{x}^\top\boldsymbol{\beta}_0, \qquad \forall\mathbf{x}\in\mathbb{R}^k$$

Then $\widehat{\boldsymbol{\beta}}_{\text{NLLS}} = \widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\mathbf{A}_0=-\text{E}[\mathbf{X}_1\mathbf{X}_1^\top]$, $\mathbf{B}_0=\text{E}[u_1^2\mathbf{X}_1\mathbf{X}_1^\top]$.

e.g. Suppose $\text{E}[Y_i|\mathbf{X}_i] = e^{\mathbf{X}_i^\top\boldsymbol{\beta}_0}$ (k=d). Then,

$$\left.\frac{\partial\widehat{\text{Q}}_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{NLLS}}} = \mathbf{0} \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}\right)e^{\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}\mathbf{X}_i = \mathbf{0}$$

i.e. no closed-form expression for $\widehat{\boldsymbol{\beta}}_{\text{NLLS}}$

Also,

$$\mathbf{A}_0 = -\text{E}\left[e^{2\mathbf{X}_1^\top\boldsymbol{\beta}_0}\mathbf{X}_1\mathbf{X}_1^\top\right], \qquad \mathbf{B}_0 = \text{E}\left[u_1^2e^{2\mathbf{X}_1^\top\boldsymbol{\beta}_0}\mathbf{X}_1\mathbf{X}_1^\top\right]$$

In this case, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{NLLS}}$ is consistently estimated as

$$\widehat{\mathbf{V}}_{\text{n}}\left[\widehat{\boldsymbol{\beta}}_{\text{NLLS}}\right] = \frac{1}{\text{n}}\widehat{\mathbf{A}}_{\text{n}}^{-1}\widehat{\mathbf{B}}_{\text{n}}\widehat{\mathbf{A}}_{\text{n}}^{-1}$$

where

$$\widehat{\mathbf{A}}_{\text{n}} = -\frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}} \text{e}^{2\mathbf{X}_{\text{i}}^{\top}\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}\mathbf{X}_{\text{i}}\mathbf{X}_{\text{i}}^{\top}, \qquad \widehat{\mathbf{B}}_{\text{n}} = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}} \hat{\text{u}}_{\text{i}}^{2}\text{e}^{2\mathbf{X}_{\text{i}}^{\top}\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}\mathbf{X}_{\text{i}}\mathbf{X}_{\text{i}}^{\top}, \qquad \hat{\text{u}}_{\text{i}} = \text{Y}_{\text{i}} - \text{e}^{\mathbf{X}_{\text{i}}^{\top}\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}$$

Note: $\widehat{\mathbf{V}}_{\text{n}}\left[\widehat{\boldsymbol{\beta}}_{\text{NLLS}}\right]$ is heteroskedasticity-consistent, i.e. $\text{E}[\text{u}_{\text{i}}^{2}|\mathbf{X}_{\text{i}}]$ can vary with i.
(On the other hand, if $\text{E}[\text{u}_{\text{i}}^{2}|\mathbf{X}_{\text{i}}] \equiv \sigma_{\text{u}}^{2}$ w.p. 1 (i.e. homoskedastic), then can estimate the NLLS asymptotic variance via $\text{s}_{\text{n}}^{2}\widehat{\mathbf{A}}_{\text{n}}^{-1}$ where)

$$\text{s}_{\text{n}}^{2} = \frac{1}{\text{n}-\text{k}}\sum_{\text{i}=1}^{\text{n}} \hat{\text{u}}_{\text{i}}^{2}$$

In general, let

$$\widehat{\mathbf{A}}_{\text{n}} = -\frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}} \frac{\partial \text{g}(\mathbf{X}_{\text{i}},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\top}}\frac{\partial \text{g}(\mathbf{X}_{\text{i}},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}$$

$$\widehat{\mathbf{B}}_{\text{n}} = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}} \hat{\text{u}}_{\text{i}}^{2}\frac{\partial \text{g}(\mathbf{X}_{\text{i}},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\top}}\frac{\partial \text{g}(\mathbf{X}_{\text{i}},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}$$

where $\hat{\text{u}}_{\text{i}} = \text{Y}_{\text{i}} - \text{g}\left(\mathbf{X}_{\text{i}},\widehat{\boldsymbol{\beta}}_{\text{NLLS}}\right)$
Convenient to have an expression for the heteroskedasticity-consistent asymptotic variance estimator $\widehat{\mathbf{V}}_{\text{n}}\left[\widehat{\boldsymbol{\beta}}_{\text{NLLS}}\right]$ in terms of matrices

$$\text{Let } \widehat{\boldsymbol{\Omega}}_{\text{n}} \equiv \text{diag}[\hat{\text{u}}_{1}^{2},\ldots,\hat{\text{u}}_{\text{n}}^{2}]$$

$$\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}} = \left[\frac{\partial \text{g}(\mathbf{X}_{\text{i}},\boldsymbol{\beta})}{\partial \beta_{\text{j}}}\right]_{\text{i}=1,\text{j}=1}^{\text{n,d}}, \qquad \widehat{\mathbf{G}}_{\text{n}} \equiv \frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}$$

Then,

$$\widehat{\mathbf{A}}_{\text{n}} = -\frac{1}{\text{n}}\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\mathbf{G}}_{\text{n}}, \qquad \widehat{\mathbf{B}}_{\text{n}} = \frac{1}{\text{n}}\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\boldsymbol{\Omega}}_{\text{n}}\widehat{\mathbf{G}}_{\text{n}}$$

So

$$\widehat{\mathbf{V}}_{\text{n}}\left[\widehat{\boldsymbol{\beta}}_{\text{NLLS}}\right] = \frac{1}{\text{n}}\widehat{\mathbf{A}}_{\text{n}}^{-1}\widehat{\mathbf{B}}_{\text{n}}\widehat{\mathbf{A}}_{\text{n}}^{-1} = \left(\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\mathbf{G}}_{\text{n}}\right)^{-1}\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\boldsymbol{\Omega}}_{\text{n}}\widehat{\mathbf{G}}_{\text{n}}\left(\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\mathbf{G}}_{\text{n}}\right)^{-1}$$

Note: One often uses $\widehat{\mathbf{B}}_{\text{n}} = \frac{1}{\text{n}-\text{d}}\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\boldsymbol{\Omega}}_{\text{n}}\widehat{\mathbf{G}}_{\text{n}}$, which makes

$$\widehat{\mathbf{V}}_{\text{n}}\left[\widehat{\boldsymbol{\beta}}_{\text{NLLS}}\right] = \frac{\text{n}}{\text{n}-\text{d}}\left(\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\mathbf{G}}_{\text{n}}\right)^{-1}\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\boldsymbol{\Omega}}_{\text{n}}\widehat{\mathbf{G}}_{\text{n}}\left(\widehat{\mathbf{G}}_{\text{n}}^{\top}\widehat{\mathbf{G}}_{\text{n}}\right)^{-1}$$

Consider feasible generalized nonlinear least squares (FGNLLS)

Suppose we model heteroskedasticity as

$$\mathrm{E}[\mathbf{u}\mathbf{u}^{\top}|\mathbf{X}] = \boldsymbol{\Omega}(\boldsymbol{\gamma}_0), \qquad \mathbf{u} = [\mathrm{u}_1 \quad \cdots \quad \mathrm{u}_\mathrm{n}]^{\top}, \qquad \mathbf{X} = [\mathbf{X}_1 \quad \cdots \quad \mathbf{X}_\mathrm{n}]^{\top}$$

Suppose $\exists\hat{\boldsymbol{\gamma}}_\mathrm{n} \xrightarrow{\mathrm{p}} \boldsymbol{\gamma}_0$. The FGNLLS estimator of $\boldsymbol{\beta}$ maximizes the weighted sum of squares

$$\widehat{\mathrm{Q}}_\mathrm{n}(\boldsymbol{\beta}) = -\frac{1}{2\mathrm{n}}\big(\mathbf{y} - \mathbf{g}(\boldsymbol{\beta})\big)^{\top}[\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}}_\mathrm{n})]^{-1}\big(\mathbf{y} - \mathbf{g}(\boldsymbol{\beta})\big)$$

where $\mathbf{y} = [\mathrm{y}_1 \quad \cdots \quad \mathrm{y}_\mathrm{n}]^{\top}$, $\mathbf{g}(\boldsymbol{\beta}) = [\mathrm{g}(\mathbf{X}_1, \boldsymbol{\beta}) \quad \cdots \quad \mathrm{g}(\mathbf{X}_\mathrm{n}, \boldsymbol{\beta})]^{\top}$

(Exercise: Specify appropriate assumptions such that)

$$\sqrt{\mathrm{n}}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{FGNLLS}} - \boldsymbol{\beta}_0\big) \xrightarrow{\mathrm{d}} \mathrm{N}_\mathrm{d}\left(\mathbf{0}, \left(\operatorname{plim}\frac{1}{\mathrm{n}}\widehat{\mathbf{G}}_\mathrm{n}^{\top}\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})\widehat{\mathbf{G}}_\mathrm{n}\right)^{-1}\right)$$

where

$$\widehat{\mathbf{G}}_\mathrm{n} \equiv \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{g}(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\mathrm{FGNLLS}}}$$

Note: FGNLLS delivers smaller standard errors then available using the heteroskedasticity-consistent NLLS asymptotic variance estimator, but at the price of requiring a model for $\mathrm{E}[\mathbf{u}\mathbf{u}^{\top}|\mathbf{X}]$.

FGNLLS asymptotic variance estimate

$$\left(\widehat{\mathbf{G}}_\mathrm{n}^{\top}\big(\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}}_\mathrm{n})\big)^{-1}\widehat{\mathbf{G}}_\mathrm{n}\right)^{-1}$$

is inconsistent.

Weighted Nonlinear Least Squares (WNLLS) can avoid having to specify a model for $\mathrm{E}[\mathbf{u}\mathbf{u}^{\top}|\mathbf{X}]$ by adopting a working error variance matrix. In particular, let $\boldsymbol{\Sigma}_0 \equiv \boldsymbol{\Sigma}(\boldsymbol{\gamma}_0)$ be a possible model for $\mathrm{E}[\mathbf{u}\mathbf{u}^{\top}|\mathbf{X}]$ (i.e. Σ is a working error variance matrix). Assume

$$\hat{\boldsymbol{\gamma}}_\mathrm{n} \xrightarrow{\mathrm{p}} \boldsymbol{\gamma}_0, \qquad \text{Let } \widehat{\boldsymbol{\Sigma}}_\mathrm{n} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}_\mathrm{n})$$

The weighted nonlinear least squares (WNLLS) estimator of $\boldsymbol{\beta}_0$ is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}} \equiv \underset{\boldsymbol{\beta}\in\mathbb{R}^\mathrm{d}}{\operatorname{argmax}}\,\widehat{\mathrm{Q}}_\mathrm{n}(\boldsymbol{\beta}), \qquad \widehat{\mathrm{Q}}_\mathrm{n}(\boldsymbol{\beta}) = -\frac{1}{2\mathrm{n}}\big(\mathbf{y} - \mathbf{g}(\boldsymbol{\beta})\big)^{\top}\widehat{\boldsymbol{\Sigma}}_\mathrm{n}^{-1}\big(\mathbf{y} - \mathbf{g}(\boldsymbol{\beta})\big)$$

Exercise: Specify assumptions such that $\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}}$ is $\sqrt{\mathrm{n}}$-consistent with

$$\sqrt{\mathrm{n}}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}} - \boldsymbol{\beta}_0\big) \xrightarrow{\mathrm{d}} \mathrm{N}_\mathrm{d}\left(0, \operatorname{plim}\mathrm{n}\big(\widehat{\mathbf{G}}_\mathrm{n}^{\top}\widehat{\boldsymbol{\Sigma}}_\mathrm{n}^{-1}\widehat{\mathbf{G}}_\mathrm{n}\big)^{-1}\widehat{\mathbf{G}}_\mathrm{n}^{\top}\widehat{\boldsymbol{\Sigma}}_\mathrm{n}^{-1}\widehat{\boldsymbol{\Omega}}_\mathrm{n}\widehat{\boldsymbol{\Sigma}}_\mathrm{n}^{-1}\widehat{\mathbf{G}}_\mathrm{n}\big(\widehat{\mathbf{G}}_\mathrm{n}^{\top}\widehat{\boldsymbol{\Sigma}}_\mathrm{n}^{-1}\widehat{\mathbf{G}}_\mathrm{n}\big)^{-1}\right)$$

Where $\widehat{\boldsymbol{\Omega}}_\mathrm{n} = \mathrm{diag}[\hat{\mathrm{u}}_1^2 \quad \cdots \quad \hat{\mathrm{u}}_\mathrm{n}^2]$

Note: If $\boldsymbol{\Sigma}(\boldsymbol{\gamma}_0) = \boldsymbol{\Omega} \equiv \mathrm{E}[\mathbf{u}\mathbf{u}^{\top}|\mathbf{X}]$, then the WNLLS asymptotic variance will be close to the FGNLLS asymptotic variance when n is large.

Alarm: Midterm November 3, 2016

NLLS with a working error variance matrix

Let $\boldsymbol{\Sigma}_0 \equiv \boldsymbol{\Sigma}(\boldsymbol{\gamma}_0)$ be a possible model for $\mathrm{E}[\mathbf{u}\mathbf{u}^\top|\mathbf{X}]$. Assume $\hat{\boldsymbol{\gamma}}_n \xrightarrow{p} \boldsymbol{\gamma}_0$. Let $\widehat{\boldsymbol{\Sigma}}_n \equiv \boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}_n)$.
Weighted NLLS is given by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}} \equiv \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmax}}\, \widehat{\mathrm{Q}}_n(\boldsymbol{\beta})$$

where

$$\widehat{\mathrm{Q}}_n(\boldsymbol{\beta}) = -\frac{1}{2n}\big(\mathbf{y}-\mathbf{g}(\boldsymbol{\beta})\big)^\top \widehat{\boldsymbol{\Sigma}}_n^{-1}\big(\mathbf{y}-\mathbf{g}(\boldsymbol{\beta})\big)$$

Then under appropriate assumptions,

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}}-\boldsymbol{\beta}_0\big) \xrightarrow{d} \mathrm{N}_d\left(0, \operatorname{plim} n\big(\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\big)^{-1}\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\Omega}}_n\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\big(\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\big)^{-1}\right)$$
$$\text{where } \widehat{\boldsymbol{\Omega}}_n = \operatorname{diag}[\hat{u}_1^2,\dots,\hat{u}_n^2]$$

e.g. Consider WNLLS with a working model of multiplicative heteroskedasticity, i.e.

$$\sigma_i^2 \equiv \mathrm{E}\big[u_i^2\big|\mathbf{x}_i\big] = e^{\mathbf{z}_i^\top\boldsymbol{\gamma}_0}, \qquad \text{for each } i\in\{1,\dots,n\}$$
$$\text{where } \mathbf{z}_i = \text{a subvector of } \mathbf{x}_i$$

Then

$$\boldsymbol{\Sigma}_0 = \operatorname{diag}\left[e^{\mathbf{z}_i^\top\boldsymbol{\gamma}_0}: i=1,\dots,n\right]$$
$$\text{and } \widehat{\boldsymbol{\Sigma}}_n = \operatorname{diag}\left[e^{\mathbf{z}_i^\top\hat{\boldsymbol{\gamma}}_n}: i=1,\dots,n\right]$$

where $\hat{\boldsymbol{\gamma}}_n$ can be obtained by NLLS of the squared NLLS residuals $(y_i-g(\mathbf{x}_i,\widehat{\boldsymbol{\beta}}_{\mathrm{NLLS}}))^2$ on $e^{\mathbf{z}_i^\top\boldsymbol{\gamma}}$.
We have

$$\boldsymbol{\Sigma}_0^{-1} = \operatorname{diag}\left[e^{-\mathbf{z}_i^\top\boldsymbol{\gamma}_0}: i=1,\dots,n\right]$$

And the WNLLS estimator maximizes

$$\widehat{\mathrm{Q}}_n(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}\frac{\big(y_i-g(\mathbf{x}_i,\boldsymbol{\beta})\big)^2}{e^{\mathbf{z}_i^\top\hat{\boldsymbol{\gamma}}_n}}$$

The asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}}$ is consistently estimated by

$$\left(\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\right)^{-1}\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\Omega}}_n\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\left(\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\right)^{-1} = \left(\sum_{i=1}^n \frac{1}{\widehat{\sigma}_i^2}\frac{\partial}{\partial\boldsymbol{\beta}^\top}g(\mathbf{x}_i,\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}}}\right)^{-1}$$
$$\times\left(\sum_{i=1}^n \frac{\hat{u}_i^2}{\hat{\sigma}_i^4}\frac{\partial}{\partial\boldsymbol{\beta}^\top}g(\mathbf{x}_i,\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}}}\right)$$
$$\times\left(\sum_{i=1}^n \frac{1}{\widehat{\sigma}_i^2}\frac{\partial}{\partial\boldsymbol{\beta}^\top}g(\mathbf{x}_i,\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}}}\right)^{-1}$$

where

$$\widehat{\sigma}_i^2 = e^{\mathbf{z}_i^\top\hat{\boldsymbol{\gamma}}_n}, \qquad \hat{u}_i = y_i - g\left(\mathbf{x}_i,\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}}\right)$$

Note: If one assumes that

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Omega}_0 \equiv E[\mathbf{u}\mathbf{u}^\top|\mathbf{X}]$$

then WNLLS conincides with FGNLLS and the asymptotic variance matrix is consistently estimated by

$$\left(\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\right)^{-1} = \left(\sum_{i=1}^n \frac{1}{\widehat{\sigma}_i^2}\frac{\partial}{\partial\boldsymbol{\beta}^\top}g(\mathbf{x}_i,\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}}g(\mathbf{x}_i,\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\mathrm{WNLLS}}}\right)^{-1}$$

e.g. Suppose $y|\mathbf{x} \sim \text{exponential}\left(e^{\mathbf{x}^\top\boldsymbol{\beta}_0}\right)$, i.e. y given $\mathbf{x}$ has a continuous distribution with density

$$f_{Y|\mathbf{x}}(y;\boldsymbol{\beta}_0) = \begin{cases} e^{\mathbf{x}^\top\boldsymbol{\beta}_0}e^{-e^{\mathbf{x}^\top\boldsymbol{\beta}_0}y}, & y > 0 \\ 0, & y \le 0 \end{cases}$$

Then $E[Y|\mathbf{x}] = e^{-\mathbf{x}^\top\boldsymbol{\beta}_0}$. Suppose we observe $[\mathbf{x}_i^\top\ y_i]^\top$ (i=1,…,n) iid. In this case, the OLS estimator of $\boldsymbol{\beta}_0$ is inconsistent (exercise: verify this).
The MLE is obtained as follows.

$$\log f_{Y|\mathbf{x}}(y;\boldsymbol{\beta}_0) = \mathbf{x}^\top\boldsymbol{\beta}_0 - e^{\mathbf{x}^\top\boldsymbol{\beta}_0}y, \qquad \text{if } y > 0$$

So

$$\hat{L}_n(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^n \log f_{Y|\mathbf{x}}(y_i;\boldsymbol{\beta})$$
$$= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^\top\boldsymbol{\beta} - \frac{1}{n}\sum_{i=1}^n e^{\mathbf{x}_i^\top\boldsymbol{\beta}}y_i$$
$$\Rightarrow \frac{\partial\hat{L}_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i - \frac{1}{n}\sum_{i=1}^n e^{\mathbf{x}_i^\top\boldsymbol{\beta}}y_i\mathbf{x}_i$$

The FOCs for likelihood maximization are

$$\frac{\partial \hat{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = \mathbf{0} \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\left(1 - e^{\mathbf{x}_i^\top \boldsymbol{\beta}} y_i\right)\mathbf{x}_i = \mathbf{0}$$

which has no closed-form solution (exercise: check the second-order condition, give a sufficient condition for $\frac{\partial^2 \hat{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top}$ to be negative definite).

Since $E[Y|\mathbf{x}] = e^{-\mathbf{x}^\top \boldsymbol{\beta}_0}$, we can write $Y = e^{-\mathbf{x}^\top \boldsymbol{\beta}_0} + u$, where u is a random error with $E[u|\mathbf{x}]=0$. Note that

$$E[u^2|\mathbf{x}] = Var[u|\mathbf{x}] = Var[y|\mathbf{x}] = e^{-2\mathbf{x}^\top \boldsymbol{\beta}_0}$$

So (conditionally) heteroskedastic

$\boldsymbol{\beta}_0$ can be consistently estimated by NLLS by maximizing

$$\hat{Q}_{NLLS}(\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}\right)^2$$

So

$$\frac{\partial \hat{Q}_{NLLS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}\right)e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}\mathbf{x}_i \neq \frac{\partial \hat{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

Therefore $\hat{\boldsymbol{\beta}}_{NLLS} \neq \hat{\boldsymbol{\beta}}_{MLE}$.

Now consider WNLLS. Choose an error variance proportional to the mean. In particular, set

$$\boldsymbol{\Sigma}_0 = \text{diag}\left[e^{-\mathbf{x}_i^\top \boldsymbol{\beta}_0} : i = 1, \dots, n\right]$$

and $\hat{\boldsymbol{\beta}}_{WNLLS}$ to maximize

$$\hat{Q}_{WNLLS}(\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^{n}\frac{\left(y_i - e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}\right)^2}{e^{-\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{NLLS}}}$$

$$\Rightarrow \frac{\partial \hat{Q}_{WNLLS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}\right)e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}e^{\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{NLLS}}\mathbf{x}_i$$

$$\Rightarrow \left.\frac{\partial \hat{Q}_{WNLLS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \approx -\frac{1}{n}\sum_{i=1}^{n}\left(y_i e^{\mathbf{x}_i^\top \boldsymbol{\beta}_0} - 1\right)\mathbf{x}_i = -\left.\frac{\partial \hat{L}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$$

when n is large (given the consistency of $\hat{\boldsymbol{\beta}}_{NLLS}$).

Therefore, $\hat{\boldsymbol{\beta}}_{FGNLLS} \approx \hat{\boldsymbol{\beta}}_{MLE}$ when n is large, i.e. $\hat{\boldsymbol{\beta}}_{FGNLLS}$ is asymptotically efficient assuming a correctly specified likelihood.

Consider some simulations.

Suppose $Y|x \sim \text{exponential}\left(e^{\beta_1+\beta_2 x}\right)$, where $X \sim N(1,1^2)$ and $\beta_1=2$, $\beta_2=-1$. Draw $[x_i\ y_i]^\top$ $(i=1,\ldots,10{,}000)$ from this model. Find $\bar{y}=0.62$ with standard deviation $s_y=1.29$.

1. OLS

$$\hat{\beta}_{1,\text{OLS}} = -0.0093, \qquad \hat{\beta}_{2,\text{OLS}} = 0.6198, \qquad \text{inconsistent}$$

2. MLE

$$\hat{\beta}_{1,\text{MLE}} = \underset{(0.0144)}{1.9829}, \qquad \hat{\beta}_{2,\text{MLE}} = \underset{(0.0099)}{-0.9896}$$

The MLE standard errors are derived from

$$\frac{1}{n}\widehat{\mathbf{H}}_n^{-1}\left(\widehat{\boldsymbol{\beta}}_{\text{MLE}}\right), \qquad \text{where } n = 10{,}000 \text{ and}$$

$$\widehat{\mathbf{H}}_n\left(\widehat{\boldsymbol{\beta}}_{\text{MLE}}\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top}\log f_{Y|x}(y_i;\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{MLE}}}$$

3. NLLS

$$\hat{\beta}_{1,\text{NLLS}} = \underset{(0.1421)}{1.8876}, \qquad \hat{\beta}_{2,\text{NLLS}} = \underset{(0.0612)}{0.9575}$$

The NLLS standard errors come from the heteroskedasticity-consistent estimate

$$\begin{aligned}\widehat{\mathbf{V}}\left[\widehat{\boldsymbol{\beta}}_{\text{NLLS}}\right] &= \frac{1}{n}\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n\widehat{\mathbf{A}}_n^{-1} \\ &= \left(\widehat{\mathbf{G}}_n^\top\widehat{\mathbf{G}}_n\right)^{-1}\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Omega}}_n\widehat{\mathbf{G}}_n\left(\widehat{\mathbf{G}}_n^\top\widehat{\mathbf{G}}_n\right)^{-1}\end{aligned}$$

where

$$\widehat{\mathbf{A}}_n = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(x_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\frac{\partial g(x_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}, \qquad \widehat{\mathbf{B}}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\frac{\partial g(x_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\frac{\partial g(x_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{NLLS}}}$$

$$\hat{u}_i = y_i - g\left(x_i,\widehat{\boldsymbol{\beta}}_{\text{NLLS}}\right), \qquad g(x_i,\boldsymbol{\beta}) = e^{-\beta_1-\beta_2 x_i}$$

4. WNLLS with the working error variance

$$\hat{\sigma}_i^2 = e^{-\hat{\beta}_{1,\text{NLLS}}-\hat{\beta}_{2,\text{NLLS}}x_i}$$

we get

$$\hat{\beta}_{1,\text{WNLLS}} = \underset{(0.0359)}{1.9906}, \qquad \hat{\beta}_{2,\text{WNLLS}} = \underset{(0.0224)}{-0.9961}$$

The WNLLS standard errors come from the matrix

$$\left(\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\right)^{-1}\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\Omega}}_n\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\left(\widehat{\mathbf{G}}_n^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\mathbf{G}}_n\right)^{-1}$$

where

$$\widehat{\boldsymbol{\Sigma}}_n = \text{diag}\left[\hat{\sigma}_i^2 : i = 1,\ldots,n\right], \qquad \widehat{\boldsymbol{\Omega}}_n = \text{diag}\left[\hat{u}_i^2 : i = 1,\ldots,n\right]$$

5. FGNLLS

$$\hat{\beta}_{1,\text{FGNLLS}} = \underset{(0.0146)}{1.9840}, \qquad \hat{\beta}_{2,\text{FGNLLS}} = \underset{(0.0101)}{-0.9907}$$

FGNLLS in this case is basically equal to MLE in terms of performance.
The FGNLLS standard errors come from

$$\left(\hat{\mathbf{G}}_n^{\top}\hat{\boldsymbol{\Sigma}}_n^{-1}\hat{\mathbf{G}}_n\right)^{-1}$$

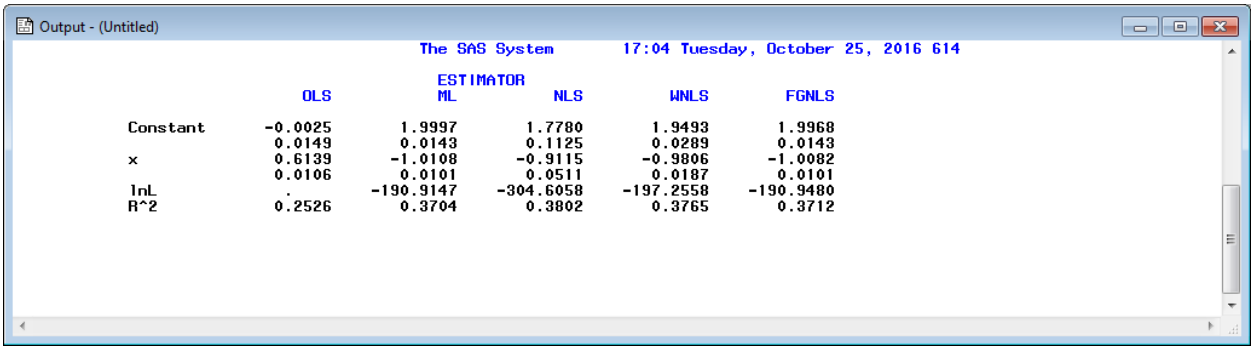
SAS Program for the Simulation

```
/************************************************
Replication of Table 5.7 from Cameron and Trivedi "Microeconometrics" (p. 161)
Written by Junyong Kim, University of Wisconsin-Milwaukee on October 25, 2016
************************************************/
resetline;
ods html close;
ods graphics off;
ods listing;

/************************************************
1. Generate 10,000 observations from X~Normal(1,1), Y|x~Exponential(exp(2-1*x))
************************************************/
data _01;
   do I=1 to 10000;
      X=1+rannor(1);
      Y=-log(1-ranuni(2))/exp(2-X);
      output;
   end;
run;

/************************************************
2. Compute sample mean and standard deviation of yi
************************************************/
proc means mean std maxdec=2;
   var Y;
run;

/************************************************
3. Compute OLS, ML, NLS, WNLS, FGNLS estimates from (yi,xi)
************************************************/
proc iml;
   use _01;
   read all var{Y} into Y;
   read all var{X} into X;
   X2=j(nrow(X),1,1)||X;
   /*********************************************
   3.1. Compute OLS estimates (from the closed-form solution)
   *********************************************/
   BETA_OLS=inv(X2`*X2)*X2`*Y;
   U_OLS=Y-X2*BETA_OLS;
   S2_OLS=U_OLS`*U_OLS/(nrow(X2)-ncol(X2));
   VAR_OLS=S2_OLS*inv(X2`*X2);
   STE_OLS=sqrt(vecdiag(VAR_OLS));
   RSQ_OLS=1-U_OLS`*U_OLS/((Y-mean(Y))`*(Y-mean(Y)));
   /*********************************************
   3.2. Compute ML estimates (by maximizing the log-likelihood with Quasi-Newton method)
   *********************************************/
   start LNHAT(BETA) global(X,Y);
      LNHAT=0;
      do I=1 to nrow(Y);
         LNHAT=LNHAT+BETA[1]+BETA[2]*X[I]-exp(BETA[1]+BETA[2]*X[I])*Y[I];
      end;
      return(LNHAT);
   finish LNHAT;
   call nlpqn(RC1,BETA_MLE,"LNHAT",BETA_OLS,{1,0});
```

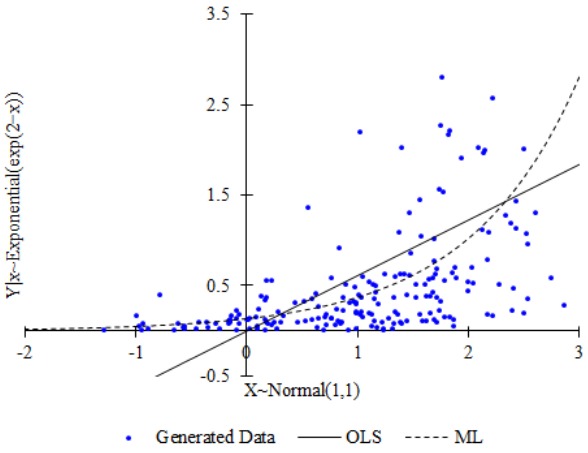
```
BETA_MLE=BETA_MLE`;
H_MLE={0 0,0 0};
do I=1 to nrow(Y);
   H_MLE=H_MLE+exp(BETA_MLE[1]+BETA_MLE[2]*X[I])*Y[I]*(1//X[I])*(1||X[I])/nrow(Y);
end;
VAR_MLE=inv(H_MLE)/nrow(Y);
STE_MLE=sqrt(vecdiag(VAR_MLE));
/************************************************
3.3. Compute Nonlinear Least Squares estimates (by maximizing Q with QN method)
************************************************/
start QNLLS(BETA) global(X,Y);
   QNLLS=0;
   do I=1 to nrow(Y);
      QNLLS=QNLLS-(Y[I]-exp(-BETA[1]-BETA[2]*X[I]))**2/(2*nrow(Y));
   end;
   return(QNLLS);
finish QNLLS;
call nlpqn(RC2,BETA_NLLS,"QNLLS",BETA_OLS,{1,0});
BETA_NLLS=BETA_NLLS`;
A_NLLS=0;
B_NLLS=0;
do I=1 to nrow(Y);
   A_NLLS=A_NLLS-exp(-2*BETA_NLLS[1]-2*BETA_NLLS[2]*X[I])*(1//X[I])*(1||X[I])/nrow(Y);
   B_NLLS=B_NLLS+(Y[I]-exp(-BETA_NLLS[1]-BETA_NLLS[2]*X[I]))**2
      *exp(-2*BETA_NLLS[1]-2*BETA_NLLS[2]*X[I])*(1//X[I])*(1||X[I])/nrow(Y);
end;
VAR_NLLS=inv(A_NLLS)*B_NLLS*inv(A_NLLS)/nrow(Y);
STE_NLLS=sqrt(vecdiag(VAR_NLLS));
/************************************************
3.4. Compute Weighted NLS estimates (by maximizing Q with QN method)
************************************************/
start QWNLLS(BETA) global(X,Y,BETA_NLLS);
   QWNLLS=0;
   do I=1 to nrow(Y);
      QWNLLS=QWNLLS-(Y[I]-exp(-BETA[1]-BETA[2]*X[I]))**2
         /(2*nrow(Y)*exp(-BETA_NLLS[1]-BETA_NLLS[2]*X[I]));
   end;
   return(QWNLLS);
finish QWNLLS;
call nlpqn(RC3,BETA_WNLLS,"QWNLLS",BETA_NLLS,{1,0});
BETA_WNLLS=BETA_WNLLS`;
U_WNLLS=j(nrow(Y),1,.);
do I=1 to nrow(Y);
   U_WNLLS[I]=Y[I]-exp(-BETA_WNLLS[1]-BETA_WNLLS[2]*X[I]);
end;
U2_WNLLS=U_WNLLS##2;
A_WNLLS=0;
B_WNLLS=0;
do I=1 to nrow(Y);
   A_WNLLS=A_WNLLS+exp(-2*BETA_WNLLS[1]-2*BETA_WNLLS[2]*X[I])*(1//X[I])*(1||X[I])
      /exp(-BETA_WNLLS[1]-BETA_WNLLS[2]*X[I]);
   B_WNLLS=B_WNLLS+exp(-2*BETA_WNLLS[1]-2*BETA_WNLLS[2]*X[I])*(1//X[I])*(1||X[I])
      *U2_WNLLS[I]/exp(2*(-BETA_WNLLS[1]-BETA_WNLLS[2]*X[I]));
end;
VAR_WNLLS=inv(A_WNLLS)*B_WNLLS*inv(A_WNLLS);
STE_WNLLS=sqrt(vecdiag(VAR_WNLLS));
/************************************************
3.5. Compute Feasible Generalized NLS estimates (by maximizing Q with QN method)
************************************************/
start QFGNLLS(BETA) global(X,Y,BETA_NLLS);
   QFGNLLS=0;
   do I=1 to nrow(Y);
      QFGNLLS=QFGNLLS-(Y[I]-exp(-BETA[1]-BETA[2]*X[I]))**2
         /(2*nrow(Y)*exp(-2*BETA_NLLS[1]-2*BETA_NLLS[2]*X[I]));
   end;
   return(QFGNLLS);
finish QFGNLLS;
call nlpqn(RC3,BETA_FGNLLS,"QFGNLLS",BETA_NLLS,{1,0});
BETA_FGNLLS=BETA_FGNLLS`;
A_FGNLLS=0;
do I=1 to nrow(Y);
```

```
      A_FGNLLS=A_FGNLLS+exp(-2*BETA_FGNLLS[1]-2*BETA_FGNLLS[2]*X[I])*(1//X[I])*(1||X[I])
         /exp(-2*BETA_FGNLLS[1]-2*BETA_FGNLLS[2]*X[I]);
   end;
   VAR_FGNLLS=inv(A_FGNLLS);
   STE_FGNLLS=sqrt(vecdiag(VAR_FGNLLS));
   start RSQ_NL(BETA) global(X,Y);
      U=j(nrow(Y),1,.);
      do I=1 to nrow(Y);
         U[I]=Y[I]-exp(-BETA[1]-BETA[2]*X[I]);
      end;
      RSQ_NL=1-U`*U/((Y-mean(Y))`*(Y-mean(Y)));
      return(RSQ_NL);
   finish RSQ_NL;
   *print BETA_OLS[format=12.4] STE_OLS[format=12.4],
      BETA_MLE[format=12.4] STE_MLE[format=12.4],
      BETA_NLLS[format=12.4] STE_NLLS[format=12.4],
      BETA_WNLLS[format=12.4] STE_WNLLS[format=12.4],
      BETA_FGNLLS[format=12.4] STE_FGNLLS[format=12.4];
   ESTIMATOR=(BETA_OLS[1]||BETA_MLE[1]||BETA_NLLS[1]||BETA_WNLLS[1]||BETA_FGNLLS[1])//
      (STE_OLS[1]||STE_MLE[1]||STE_NLLS[1]||STE_WNLLS[1]||STE_FGNLLS[1])//
      (BETA_OLS[2]||BETA_MLE[2]||BETA_NLLS[2]||BETA_WNLLS[2]||BETA_FGNLLS[2])//
      (STE_OLS[2]||STE_MLE[2]||STE_NLLS[2]||STE_WNLLS[2]||STE_FGNLLS[2])//
      (.||LNHAT(BETA_MLE)||LNHAT(BETA_NLLS)||LNHAT(BETA_WNLLS)||LNHAT(BETA_FGNLLS))//
      (RSQ_OLS||RSQ_NL(BETA_MLE)||RSQ_NL(BETA_NLLS)||RSQ_NL(BETA_WNLLS)||RSQ_NL(BETA_FGNLLS));
   print ESTIMATOR[colname={"OLS" "ML" "NLS" "WNLS" "FGNLS"}
      rowname={"Constant" "" "x" "" "lnL" "R^2"} format=12.4];
quit;
```

SAS Result for the Simulation



Scatter Plot



Alarm: Midterm November 3, 2016

- Cover many things from nonlinear estimation
- Also cover sketches of proofs of consistency, Normality…

Numerical Methods of Optimization

Nonlinear estimators are usually only implicitly defined.
e.g.

$$\widehat{\boldsymbol{\theta}}_n \equiv \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \widehat{Q}_n(\boldsymbol{\theta}), \qquad \text{(M Estimation)}$$

$$\left.\frac{\partial \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} = \mathbf{0}, \qquad \text{(Z Estimation)}$$

Important to know how computers compute $\widehat{\boldsymbol{\theta}}_n$: essentially two possibilities
1. Grid Search

- Pick many possible values of $\boldsymbol{\theta}$ in some set
- Compute $\hat{Q}_n(\boldsymbol{\theta})$ for each values
- Pick $\widehat{\boldsymbol{\theta}}_n$ to be the value of $\boldsymbol{\theta}$ that maximizes $\hat{Q}_n(\boldsymbol{\theta})$/makes $\frac{\partial \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}$ close to $\mathbf{0}$ on the grid
- This method works if the grid is adequate
- This method is very inefficient if $\boldsymbol{\theta}$ is event of moderately high dimensional

e.g. Suppose $\boldsymbol{\theta}\in\mathbb{R}^{10}$. Pick a grid of just 10 points (this grid is likely to be too sparse). Then $\hat{Q}_n(\boldsymbol{\theta})$ or $\frac{\partial \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}$ will need to be evaluated 10^{10}=10 billion times

2. Iterative Methods

- Much more popular than grid search
- Suppose we want to estimate

$$\boldsymbol{\theta}_0 \equiv \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \mathrm{E}\left[\widehat{Q}_n(\boldsymbol{\theta})\right], \qquad \text{via } \widehat{\boldsymbol{\theta}}_n \equiv \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \widehat{Q}_n(\boldsymbol{\theta})$$

- The idea is to update estimates of $\boldsymbol{\theta}_0$ via an algorithm
- Suppose $\widehat{\boldsymbol{\theta}}_s$ estimates $\boldsymbol{\theta}_0$ at iteration s
- The idea is to use $\widehat{\boldsymbol{\theta}}_s$ to generate another estimate $\widehat{\boldsymbol{\theta}}_{s+1}$ such that $\widehat{Q}_n(\widehat{\boldsymbol{\theta}}_{s+1}) > \widehat{Q}_n(\widehat{\boldsymbol{\theta}}_s)$
- Most iterative methods are gradient methods, i.e. we have for each iteration s∈{1,…,S}

$$\widehat{\boldsymbol{\theta}}_{s+1} = \widehat{\boldsymbol{\theta}}_s + \mathbf{A}_s\mathbf{g}_s$$

where $\mathbf{A}_s$ is a matrix that depends on $\widehat{\boldsymbol{\theta}}_s$ and

$$\mathbf{g}_s = \left.\frac{\partial \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_s}, \qquad \text{i.e. the gradient vector at } \widehat{\boldsymbol{\theta}}_s$$

- Different gradient methods are distinguished by different choice of weighting matrix $\mathbf{A}_s$

e.g. the Newton–Raphson method uses

$$\mathbf{A}_s = -\mathbf{H}_s^{-1}, \qquad \text{where } \mathbf{H}_s = \left.\frac{\partial^2 \hat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_s}$$

- In general, want $\mathbf{A}_s$ to be positive definite with high probability. This makes $\hat{Q}_n(\hat{\boldsymbol{\theta}}_{s+1}) > \hat{Q}_n(\hat{\boldsymbol{\theta}}_s)$ with high probability
- To see this, $\hat{Q}_n(\hat{\boldsymbol{\theta}}_{s+1}) = \hat{Q}_n(\hat{\boldsymbol{\theta}}_s) + \mathbf{g}_s^\top \mathbf{A}_s \mathbf{g}_s + R_n$, where $R_n = o_p(\mathbf{g}_s^\top \mathbf{A}_s \mathbf{g}_s)$

Note: If $\|\mathbf{A}_s\|$ is too small, then the algorithm might a slow rate of convergence. If $\|\mathbf{A}_s\|$ is too large, then the algorithm might cause $\{\hat{\boldsymbol{\theta}}_s\}$ to overshoot the actual maximizer of $\hat{Q}_n(\boldsymbol{\theta})$

- Gradient methods often use <u>step-size adjustments</u> to reduce the risk of overshooting/undershooting

Take

$$\hat{\boldsymbol{\theta}}_{s+1} = \hat{\boldsymbol{\theta}}_s + \hat{\lambda}_s \mathbf{A}_s \mathbf{g}_s$$

where $\hat{\lambda}_s$ is a scalar step size chosen to maximize $\hat{Q}_n(\hat{\boldsymbol{\theta}}_{s+1})$

- This works by calculating $\mathbf{A}_s\mathbf{g}_s$, followed by calculating $\hat{Q}_n(\hat{\boldsymbol{\theta}}_s + \lambda \mathbf{A}_s \mathbf{g}_s)$ $\forall$ values of λ in some chosen grid Λ. The final choice of $\hat{\lambda}_s$ is the true value $\lambda \in \Lambda$ that maximizes $\hat{Q}_n(\hat{\boldsymbol{\theta}}_s + \lambda \mathbf{A}_s \mathbf{g}_s)$
- This procedure minimizes the total number of iterations at the cost of doing grid searches between iterations
- Gradient methods are tend to converge to local maxima close to the chosen starting value $\hat{\boldsymbol{\theta}}_1$. If $\hat{Q}_n(\boldsymbol{\theta})$ is suspected to have local maxima, then one should try different starting values spaced far apart
- Iterations continued until (in the ideal case) converge, which is defined to occur when

(a) $\left|\hat{Q}_n(\hat{\boldsymbol{\theta}}_{s+1}) - \hat{Q}_n(\hat{\boldsymbol{\theta}}_s)\right|$ is small

(b) $\|\mathbf{A}_{s+1}\mathbf{g}_{t+1} - \mathbf{A}_s\mathbf{g}_s\|$ is small

(c) $\left\|\hat{\boldsymbol{\theta}}_{s+1} - \hat{\boldsymbol{\theta}}_s\right\|$ is small

- Ideally want all three of these convergence criteria to be very small, say $<10^{-6}$
- Gradient methods also typically involve a preset maximum number of iterations: if the number of iterations exceeds this maximum without converging, then the algorithm stops and is said not to have converged
- If the algorithm takes S iterations to converge, then $\hat{\boldsymbol{\theta}}_s$ is a local maximum of $\hat{Q}_n(\boldsymbol{\theta})$: no guarantee that $\hat{\boldsymbol{\theta}}_s$ is a global maximum unless $\hat{Q}_n(\boldsymbol{\theta})$ is globally concave
- Can reduce the total number of iterations by picking a starting value $\hat{\boldsymbol{\theta}}_1$, that is a consistent estimate of $\boldsymbol{\theta}_0$

e.g. Newton–Raphson (NR)
This is a gradient method appropriate for cases where $\hat{Q}_n(\boldsymbol{\theta})$ is globally concave

$$\mathbf{A}_s = -\mathbf{H}_s^{-1}, \qquad \text{where } \mathbf{H}_s = \left.\frac{\partial^2 \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_s}$$

(Note: H_s^{-1} is guaranteed to exist for all possible values of $\widehat{\boldsymbol{\theta}}_s$ if $\hat{Q}_n(\boldsymbol{\theta})$ is globally concave)
The NR method is motivated via a Taylor expansion

$$\widehat{Q}_n(\boldsymbol{\theta}) = \widehat{Q}_n(\widehat{\boldsymbol{\theta}}_s) + \mathbf{g}_s^\top(\theta - \hat{\theta}_s) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s)^\top \mathbf{H}_s(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) + \hat{R}_n$$

which if $\hat{R}_n$ is small, yields the approximation

$$\widehat{Q}_n(\boldsymbol{\theta}) \approx \widehat{Q}_n^*(\boldsymbol{\theta}) \equiv \widehat{Q}_n(\widehat{\boldsymbol{\theta}}_s) + \mathbf{g}_s^\top(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s)^\top \mathbf{A}_s(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s)$$

We see that $\hat{Q}_n^*(\boldsymbol{\theta})$ is globally maximized by solving

$$\frac{\partial \widehat{Q}_n^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} = \mathbf{0} \Leftrightarrow \mathbf{g}_s + \mathbf{H}_s(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s) = \mathbf{0}$$
$$\Leftrightarrow \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_s - \mathbf{H}_s^{-1}\mathbf{g}_s$$

which implies that $\widehat{\boldsymbol{\theta}}_{s+1} = \widehat{\boldsymbol{\theta}}_s - \mathbf{H}_s^{-1}\mathbf{g}_s$ is a global maximum (since Hs is negative definite) of $\hat{Q}_n^*(\boldsymbol{\theta})$.
Note:

$$\widehat{Q}_n(\widehat{\boldsymbol{\theta}}_{s+1}) - \widehat{\mathbf{Q}}_n(\widehat{\boldsymbol{\theta}}_s) = -\mathbf{g}_s^\top \mathbf{H}_s^{-1}\mathbf{g}_s + \frac{1}{2}(-\mathbf{H}_s^\top \mathbf{g}_s)^\top \mathbf{H}_s(-\mathbf{H}_s^\top \mathbf{g}_s) + \hat{R}_n$$
$$= -\frac{1}{2}\mathbf{g}_s^\top \mathbf{H}_s^{-1}\mathbf{g}_s + \hat{R}_n$$

So if $\hat{R}_n$ is small, then $\widehat{Q}_n(\widehat{\boldsymbol{\theta}}_{s+1}) - \widehat{Q}_n(\widehat{\boldsymbol{\theta}}_s) \approx -\frac{1}{2}\mathbf{g}_s^\top \mathbf{H}_s^{-1}\mathbf{g}_s$, which is positive if $\mathbf{H}_s$ is negative definite (exercise: prove this).
Note: If $\widehat{\boldsymbol{\theta}}_s$ is a local maximum, then $\mathbf{H}_s$ is negative semidefinite by definition.

- If $\widehat{\boldsymbol{\theta}}_s$ is far from a local maximum, then it is unclear if $\mathbf{H}_s$ is even a definite matrix without assumptions on the shape of $\hat{Q}_n(\boldsymbol{\theta})$: in such cases, NR may fail
- If $\mathbf{H}_s$ is singular, then $\widehat{\boldsymbol{\theta}}_{s+1}$ is not computable and algorithm fails
- For these reasons, NR works best when $\hat{Q}_n(\boldsymbol{\theta})$ is globally concave

e.g. Method of Scoring (MS)
This is a variant of NR, where instead of $\mathbf{A}_s = -\mathbf{H}_s^{-1}$, we use $\mathbf{A}_s = -\bar{\mathbf{H}}_s^{-1}$, where

$$\bar{\mathbf{H}}_s = \mathrm{E}\left[\left.\frac{\partial^2 \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right]\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_s}, \qquad \text{or a consistent estimate of } \mathrm{E}\left[\left.\frac{\partial^2 \widehat{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

MS is often applied to compute MLEs, i.e. cases where

$$\widehat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\log \hat{L}_n(\boldsymbol{\theta})$$

Recall by the information identity,

$$-\mathrm{E}\left[\left.\frac{\partial^2 \log \hat{\mathrm{L}}_\mathrm{n}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = \mathrm{E}\left[\left.\frac{\partial \log \hat{\mathrm{L}}_\mathrm{n}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\frac{\partial \log \hat{\mathrm{L}}_\mathrm{n}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$$

which is positive definite in view of the score identity.

Therefore, $\bar{\mathbf{H}}_\mathrm{s}$ is likely to be negative definite (for MLE computation) when $\hat{\boldsymbol{\theta}}_\mathrm{s}$ is close to $\boldsymbol{\theta}_0$ (e.g. $\hat{\boldsymbol{\theta}}_\mathrm{s}$ is consistent).

e.g. BHHH

Another variant of NR, due to Berndt, Hall, Hall and Hausman (1974): Suppose

$$\hat{\mathrm{Q}}_\mathrm{n}(\boldsymbol{\theta}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{q}(\mathbf{Z}_\mathrm{i}, \boldsymbol{\theta})$$

where q(z,$\boldsymbol{\theta}$) to be continuously differentiable in $\boldsymbol{\theta}$ for all $\mathbf{z}$. The BHHH algorithm uses $\mathbf{A}_\mathrm{s}=-\hat{\mathbf{H}}_\mathrm{s}^{-1}$, where

$$\hat{\mathbf{H}}_\mathrm{s} = -\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \left.\frac{\partial \mathrm{q}(\mathbf{z}_\mathrm{i}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\frac{\partial \mathrm{q}(\mathbf{z}_\mathrm{i}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\mathrm{s}}$$

(i.e. no second-order derivative)

BHHH, like MS, is usually applied to compute MLEs: Suppose

$$\hat{\mathrm{Q}}_\mathrm{n}(\boldsymbol{\theta}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \log \mathrm{f}(\mathbf{Z}_\mathrm{i}, \boldsymbol{\theta}), \qquad \text{where } \mathbf{Z}_1, \ldots, \mathbf{Z}_\mathrm{n} \text{ are iid with density } \mathrm{f}(\cdot; \boldsymbol{\theta})$$

By the information identity, we have

$$\begin{aligned}\frac{1}{\mathrm{n}}\mathrm{E}\left[\left.\frac{\partial^2 \log \hat{\mathrm{L}}_\mathrm{n}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] &= -\frac{1}{\mathrm{n}}\mathrm{E}\left[\left.\frac{\partial \log \hat{\mathrm{L}}_\mathrm{n}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\frac{\partial \log \hat{\mathrm{L}}_\mathrm{n}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] \\ &= -\mathrm{E}\left[\left.\frac{\partial \log \mathrm{f}(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\frac{\partial \log \mathrm{f}(\mathbf{Z}_1, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]\end{aligned}$$

which is negative definite by score identity.

Alarm: Midterm November 3, 2016

- Cover many things from nonlinear estimation
  - Especially, one should know how to derive the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ or the sandwich-style variance $A_0^{-1}B_0A_0^{-1}$, etc.
- Also cover sketches of proofs of consistency, Normality…

Gradient methods: Goal is to compute

$$\widehat{\boldsymbol{\theta}}_n \equiv \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \widehat{Q}_n(\boldsymbol{\theta})$$

by means of an algorithm that generates a sequence

$$\widehat{\boldsymbol{\theta}}_s \to \widehat{\boldsymbol{\theta}}_n$$

where for each s∈{1,…,S}

$$\widehat{\boldsymbol{\theta}}_{s+1} = \widehat{\boldsymbol{\theta}}_s + \mathbf{A}_s\mathbf{g}_s$$

e.g. c) Steepest ascent: The simplest of all gradient methods
Set $\mathbf{A}_s=\mathbf{I}_d$ and $\widehat{\boldsymbol{\theta}}_{s+1} = \widehat{\boldsymbol{\theta}}_s + \hat{\lambda}_s\mathbf{g}_s$ where the step size $\hat{\lambda}_s$ maximizes $\widehat{Q}_n(\widehat{\boldsymbol{\theta}}_s + \lambda\mathbf{g}_s)$ over λ in some grid Λ. Usually,

$$\hat{\lambda}_s = \frac{\mathbf{g}_s^\top\mathbf{g}_s}{\mathbf{g}_s^\top\mathbf{H}_s\mathbf{g}_s}, \qquad \text{where } \mathbf{H}_s = \left.\frac{\partial^2\widehat{Q}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_s}$$

(for this setting of $\hat{\lambda}_s$ to work, need $\hat{\lambda}_s\mathbf{I}_d$ to be negative definite, i.e. $\hat{\lambda}_s < 0$, i.e. $\mathbf{H}_s$ to be negative definite)

Review: Why does $\mathbf{A}_s$ need to be positive definite?

$$\begin{aligned}\widehat{Q}_n(\widehat{\boldsymbol{\theta}}_{s+1}) - \widehat{Q}_n(\widehat{\boldsymbol{\theta}}_s) &= \mathbf{g}_s^\top\mathbf{A}_s\mathbf{g}_s + \widehat{R}_n \\ &\approx \mathbf{g}_s^\top\mathbf{A}_s\mathbf{g}_s, \qquad \text{if } \widehat{R}_n \text{ is small}\end{aligned}$$

Need $\mathbf{g}_s^\top\mathbf{A}_s\mathbf{g}_s > 0$, i.e. $\mathbf{A}_s$ should be positive definite.

e.g. d) Gauss–Newton (GN): Gradient method applicable to NLLS

Suppose $E[Y|\mathbf{X}]=g(\mathbf{X},\boldsymbol{\beta})$. Observe $[\mathbf{X}_i^\top\ Y_i]^\top$ (i=1,…,n) iid. The GN method sets $\widehat{\boldsymbol{\beta}}_{s+1}-\widehat{\boldsymbol{\beta}}_s$ equal to the OLS estimates in a regression given by

$$Y_i-g\left(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_s\right)=\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\boldsymbol{\beta}+v_i$$

i.e. $\widehat{\boldsymbol{\beta}}_{s+1}$ is the OLS estimate for a regression given by

$$Y_i-g\left(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_s\right)-\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\widehat{\boldsymbol{\beta}}_s=\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\boldsymbol{\beta}+v_i$$

GN is motivated by the following Taylor expansion

$$g(\mathbf{X}_i,\boldsymbol{\beta})=g\left(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_s\right)+\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_s\right)+R_n$$

Assume that R_n is small, substitute the approximation

$$g^*(\mathbf{X}_i,\boldsymbol{\beta})\equiv g\left(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_s\right)+\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_s\right)$$

into the NLLS criterion function to get

$$\widehat{Q}_n(\boldsymbol{\beta})\approx-\frac{1}{2n}\sum_{i=1}^{n}\left[Y_i-g\left(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_s\right)-\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_s\right)\right]^2$$

which is equivalent to the OLS criterion function for a regression of $Y_i-g\left(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_s\right)$ on $\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}$ with parameter vector $\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_s$. This implies that

$$\widehat{\boldsymbol{\beta}}_{s+1}=\widehat{\boldsymbol{\beta}}_s+\left(\sum_{i=1}^{n}\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\right)^{-1}\sum_{i=1}^{n}\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\left(Y_i-g\left(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_s\right)\right)$$

This shows that GN is a special case of MS applied to NLLS.

$$\text{For } \widehat{Q}_n(\boldsymbol{\beta})=-\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i-g(\mathbf{X}_i,\boldsymbol{\beta})\right)^2,\qquad \text{we have}$$

$$\frac{\partial\widehat{Q}_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-g(\mathbf{X}_i,\boldsymbol{\beta})\right)\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}$$

$$\frac{\partial^2\widehat{Q}_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top}=-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}+\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-g(\mathbf{X}_i,\boldsymbol{\beta})\right)\frac{\partial^2 g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top}$$

$$\Rightarrow E\left[\frac{\partial^2\widehat{Q}_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top}\right]=-E\left[\frac{\partial g(\mathbf{X}_1,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^\top}\frac{\partial g(\mathbf{X}_1,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right]$$

which is approximable as

$$\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}\left.\frac{\partial g(\mathbf{X}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_s}$$

(compare with the expression for $\widehat{\boldsymbol{\beta}}_{s+1}$)

Example: Suppose we observe $[\mathbf{X}_i^\top\ Y_i]^\top$ (i=1,…,n) iid from a model where

$$E[Y_i|\mathbf{X}_i] = e^{\mathbf{X}_i^\top\boldsymbol{\beta}_0}, \qquad \text{where } \boldsymbol{\beta}_0 \in \mathbb{R}^d$$

The NLLS estimator of $\boldsymbol{\beta}_0$ is

$$\widehat{\boldsymbol{\beta}}_{NLLS} \equiv \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmax}}\,\widehat{Q}_n(\boldsymbol{\beta}), \qquad \text{where } \widehat{Q}_n(\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top\boldsymbol{\beta}}\right)^2$$

We have

$$\mathbf{g}(\boldsymbol{\beta}) \equiv \frac{\partial \widehat{Q}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top\boldsymbol{\beta}}\right)e^{\mathbf{X}_i^\top\boldsymbol{\beta}}\mathbf{X}_i$$

$$\mathbf{H}(\boldsymbol{\beta}) \equiv \frac{\partial^2 \widehat{Q}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^\top} = -\frac{1}{n}\sum_{i=1}^{n}e^{2\mathbf{X}_i^\top\boldsymbol{\beta}}\mathbf{X}_i\mathbf{X}_i^\top + \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top\boldsymbol{\beta}}\right)e^{\mathbf{X}_i^\top\boldsymbol{\beta}}\mathbf{X}_i\mathbf{X}_i^\top$$

NR sets

$$\widehat{\boldsymbol{\beta}}_{s+1} = \widehat{\boldsymbol{\beta}}_s - \mathbf{H}^{-1}(\widehat{\boldsymbol{\beta}}_s)\mathbf{g}(\widehat{\boldsymbol{\beta}}_s)$$

MS exploits the fact that

$$E_{\boldsymbol{\beta}_0}[\mathbf{H}(\boldsymbol{\beta}_0)] = -E_{\boldsymbol{\beta}_0}\left[e^{2\mathbf{X}_i^\top\boldsymbol{\beta}_0}\mathbf{X}_1\mathbf{X}_1^\top\right]$$

which can be approximated by

$$\bar{\mathbf{H}}(\widehat{\boldsymbol{\beta}}_s) \equiv -\frac{1}{n}\sum_{i=1}^{n}e^{2\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}_s}\mathbf{X}_i\mathbf{X}_i^\top$$

Therefore MS sets

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{s+1} &= \widehat{\boldsymbol{\beta}}_s - \bar{\mathbf{H}}^{-1}(\widehat{\boldsymbol{\beta}}_s)\mathbf{g}(\widehat{\boldsymbol{\beta}}_s)\\ &= \widehat{\boldsymbol{\beta}}_s + \left(\frac{1}{n}\sum_{i=1}^{n}e^{2\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}_s}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - e^{\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}_s}\right)e^{\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}_s}\mathbf{X}_i\end{aligned}$$

i.e. $\widehat{\boldsymbol{\beta}}_{s+1} - \widehat{\boldsymbol{\beta}}_s$ is the OLS estimate for a regression of $Y_i - e^{\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}}_s}$ on $e^{\mathbf{X}_i^\top\boldsymbol{\beta}}\mathbf{X}_i$. This shows that in this case, MS is GN.

3. Non-Gradient Methods

a) EM algorithm: The expectation maximization (EM) algorithm is an iterative method for computing MLEs (Hartley, 1958, Biometrics).

In Econometrics, the EM algorithm is usually associated with MLEs of latent-variable models, i.e. a model where a matrix $\mathbf{Z}$ of observable variables is related to a vector $\mathbf{y}^*$ of unobservable variables in such a way that $\mathbf{y}^*$ of unobservable variables in such a way that $\mathbf{y}^*$ uniquely determines $\mathbf{Z}$, but $\mathbf{Z}$ does not uniquely define y^*.

e.g. Suppose we observe $[D_i\ \mathbf{X}_i^\top\ Y_i]^\top$ (i=1,…,n) where for each i

$$Y_i^* = \mathbf{X}_i^\top \boldsymbol{\beta}_0 + u_i, \qquad D_i = 1\{Y_i^* > 0\}, \qquad Y_i = D_i Y_i^*$$

(and where $u_1,\ldots,u_n$ are iid $N(0,\sigma_0^2)$ and unobservable)

This is the standard Tobit model. In this case,

$$\mathbf{Y}^* = \begin{bmatrix} Y_1^* \\ \vdots \\ Y_n^* \end{bmatrix}, \qquad \mathbf{Z} = \begin{bmatrix} D_1 & \mathbf{X}_1^\top & Y_1 \\ \vdots & \vdots & \vdots \\ D_n & \mathbf{X}_n^\top & Y_n \end{bmatrix}$$

Note that observations for which $D_i=0$ do not allow one to infer anything about $\mathbf{Y}_i^*$ apart from the fact that $Y_i^*<0$. In general, suppose $\mathbf{Y}^*$ (i.e. the latent variable) has joint density $f(\mathbf{y}^*)$. Let $\mathbf{Z}$ have density $g(\mathbf{z})$.

$$\text{Let } k(\mathbf{y}^*|\mathbf{z}) \equiv \frac{f(\mathbf{y}^*)}{g(\mathbf{z})}$$

The joint density of $\mathbf{Y}^*$ and $\mathbf{Z}$ is

$$f(\mathbf{y}^*, \mathbf{z}) = f(\mathbf{z}|\mathbf{y}^*)f(\mathbf{y}^*) = f(\mathbf{y}^*)$$

because $f(\mathbf{z}|\mathbf{y}^*) \equiv 1$ in view of the fact that $\mathbf{y}^*$ uniquely determines $\mathbf{Z}$.

Assume that f(.), g(.) and k(.|.) all depend on a parameter $\boldsymbol{\theta}\in\mathbb{R}^d$, i.e.

$$f(\mathbf{y}^*) = f(\mathbf{y}^*; \boldsymbol{\theta}), \qquad g(\mathbf{z}) = g(\mathbf{z}; \boldsymbol{\theta}), \qquad k(\mathbf{y}^*|\mathbf{z}) = k(\mathbf{y}^*|\mathbf{z}; \boldsymbol{\theta})$$

Want to compute the MLE, i.e.

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} \equiv \underset{\boldsymbol{\theta}}{\text{argmax}} \frac{1}{n} \log \hat{L}_n(\boldsymbol{\theta})$$

where $\hat{L}_n(\boldsymbol{\theta}) = g(\mathbf{z}; \boldsymbol{\theta}) = \frac{f(\mathbf{y}^*;\boldsymbol{\theta})}{k(\mathbf{y}^*|\mathbf{z};\boldsymbol{\theta})}$. So

$$\frac{1}{n} \log \hat{L}_n(\boldsymbol{\theta}) = \frac{1}{n} \log f(\mathbf{y}^*; \boldsymbol{\theta}) - \frac{1}{n} \log k(\mathbf{y}^*|z; \boldsymbol{\theta})$$

$$\text{Let } \bar{R}(\boldsymbol{\theta}; \boldsymbol{\theta}_s) \equiv E_{\boldsymbol{\theta}_s}\left[\frac{1}{n} \log f(\mathbf{y}^*; \boldsymbol{\theta}) \,\middle|\, \mathbf{Z}\right]$$

The EM algorithm maximizes $\frac{1}{n} \log \hat{L}_n(\boldsymbol{\theta})$ by maximizing $\bar{R}(\boldsymbol{\theta};\boldsymbol{\theta}_s)$ with respect to $\boldsymbol{\theta}$.

Let

$$\bar{S}(\boldsymbol{\theta};\boldsymbol{\theta}_s) \equiv E_{\boldsymbol{\theta}_s}\left[\frac{1}{n}\log k(\mathbf{y}^*|\mathbf{z};\boldsymbol{\theta})\,\middle|\,\mathbf{Z}\right]$$

Note:

$$\bar{Q}(\boldsymbol{\theta};\boldsymbol{\theta}_s) \equiv E_{\boldsymbol{\theta}_s}\left[\frac{1}{n}\log g(\mathbf{z};\boldsymbol{\theta})\,\middle|\,\mathbf{Z}\right] = \frac{1}{n}\log \hat{L}_n(\boldsymbol{\theta}) = \hat{Q}_n(\boldsymbol{\theta})$$

So

$$\hat{Q}_n(\boldsymbol{\theta}) = \bar{R}(\boldsymbol{\theta};\boldsymbol{\theta}_s) - \bar{S}(\boldsymbol{\theta};\boldsymbol{\theta}_s)$$

By Jensen's inequality

$$\bar{S}(\boldsymbol{\theta};\boldsymbol{\theta}_s) < \bar{S}(\boldsymbol{\theta}_s;\boldsymbol{\theta}_s), \qquad \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}_s, \qquad \cdots(*)$$

$$\left(\text{notice: } E_{\boldsymbol{\theta}_0}\left[\log\frac{f(\mathbf{z};\boldsymbol{\theta})}{f(\mathbf{z};\boldsymbol{\theta}_0)}\right] < \log E_{\boldsymbol{\theta}_0}\left[\frac{f(\mathbf{z};\boldsymbol{\theta})}{f(\mathbf{z};\boldsymbol{\theta}_0)}\right] = 0, \qquad \forall\boldsymbol{\theta} \neq \boldsymbol{\theta}_0\right)$$

For a given $\boldsymbol{\theta}_s$, let $\boldsymbol{\theta}_{s+1}^*$ maximizes $\bar{R}(\boldsymbol{\theta};\boldsymbol{\theta}_s)$
Then

$$\hat{Q}_n(\boldsymbol{\theta}_{s+1}^*) = \bar{R}(\boldsymbol{\theta}_{s+1}^*;\boldsymbol{\theta}_s) - \bar{S}(\boldsymbol{\theta}_{s+1}^*;\boldsymbol{\theta}_s)$$
$$\text{but } \bar{R}(\boldsymbol{\theta}_{s+1}^*;\boldsymbol{\theta}_s) \geq \bar{R}(\boldsymbol{\theta};\boldsymbol{\theta}_s), \qquad \forall\boldsymbol{\theta}$$
$$\text{while } \bar{S}(\boldsymbol{\theta}_{s+1}^*;\boldsymbol{\theta}_s) \leq \bar{S}(\boldsymbol{\theta}_s;\boldsymbol{\theta}_s), \qquad \text{by } (*)$$
$$\text{hence } \hat{Q}_n(\boldsymbol{\theta}_{s+1}^*) \geq \hat{Q}_n(\boldsymbol{\theta}_s)$$

This shows that the log-likelihood never decreases under the EM algorithm. In addition, if $\hat{L}_n(\boldsymbol{\theta})$ is bounded, then $\lim_{s\to\infty}\frac{1}{n}\log\hat{L}_n(\hat{\boldsymbol{\theta}}_s)$ exists.
To see this, suppose $\boldsymbol{\theta}^*$ is a stationary point of the EM algorithm. Then $\boldsymbol{\theta}^*$ is a stationary point of $\hat{Q}_n(\boldsymbol{\theta})$. Suppose that $\hat{Q}_n(\boldsymbol{\theta})$ is differentiable. Then

$$\left.\frac{\partial\hat{Q}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_s} = \left.\frac{\partial\bar{R}(\boldsymbol{\theta};\boldsymbol{\theta}_s)}{\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_s} - \left.\frac{\partial\bar{S}(\boldsymbol{\theta};\boldsymbol{\theta}_s)}{\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_s}$$

But

$$\left.\frac{\partial\bar{S}(\boldsymbol{\theta};\boldsymbol{\theta}_s)}{\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_s} = \mathbf{0}, \qquad (\text{because } \bar{S}(\boldsymbol{\theta};\boldsymbol{\theta}_s) < \bar{S}(\boldsymbol{\theta}_s;\boldsymbol{\theta}_s)\ \forall\boldsymbol{\theta} \in \boldsymbol{\theta}_s)$$

So if $\boldsymbol{\theta}_s$ is a stationary point of $\bar{R}(\boldsymbol{\theta};\boldsymbol{\theta}_s)$ (i.e. a stationary point of the EM algorithm, it is also a stationary point of $\hat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\log\hat{L}_n(\boldsymbol{\theta})$)

Latent Variable Models: Are characterized by a random matrix $\underline{\underline{Z}}$ of observables uniquely determined by another random matrix $\underline{\underline{Z}}^*$ of un-observables and such that $\underline{\underline{Z}}$ does not uniquely determine $\underline{\underline{Z}}^*$

Notation: c, $\underline{c}$, $\underline{\underline{c}}$ for scalar, column vector, matrix; Y, $\underline{Y}$, $\underline{\underline{Y}}$ for random variable, column vector, matrix, respectively. Until now I have used expressions such as c, **c**, **C** and $\tilde{y}$, $\tilde{\mathbf{y}}$, $\tilde{\mathbf{Y}}$, respectively.

e.g. Standard Tobit model

Observe $[D_i \ \underline{X}_i^\top \ Y_i]^\top$ (i=1,…,n) where for each i

$$Y_i^* = \underline{X}_i^\top \underline{\beta}_0 + U_i, \qquad Y_i = D_i Y_i^*, \qquad D_i = 1\{Y_i^* > 0\}$$

and where $U_1,\ldots,U_n$ iid $N(0,\sigma_0^2)$. In this case

$$\underline{\underline{Z}}^* = \begin{bmatrix} \underline{X}_1^\top & Y_1^* \\ \vdots & \vdots \\ \underline{X}_n^\top & Y_n^* \end{bmatrix}, \qquad \underline{\underline{Z}} = \begin{bmatrix} D_1 & \underline{X}_1^\top & Y_1 \\ \vdots & \vdots & \vdots \\ D_n & \underline{X}_n^\top & Y_n \end{bmatrix}$$

Note: Observations for which $D_i=0$ do not allow one to infer anything about the corresponding Y_i^* apart from the fact that $Y_i^* \leq 0$.

EM algorithm for computing the MLE of latent variable models

Let $\underline{Y}^*$ (the latent variable) have joint density $f(\underline{y}^*)$. Let $\underline{\underline{Z}}$ (observed) have density $g(\underline{\underline{z}})$.

Let $k(\underline{y}^* | \underline{\underline{z}}) \equiv \frac{f(\underline{y}^*)}{g(\underline{\underline{z}})}$. Note that the joint density of $\underline{Y}^*$ and $\underline{\underline{Z}}$ is $f(\underline{y}^*,\underline{\underline{z}})=f(\underline{\underline{z}}|\underline{y}^*)f(\underline{y}^*)=f(\underline{y}^*)$ since $f(\underline{\underline{z}}|\underline{y}^*)\equiv 1$.

Assume that $f(\underline{y}^*)$, $g(\underline{\underline{z}})$, $k(\underline{y}^*|\underline{\underline{z}})$ all depend on a parameter $\underline{\theta}_0 \in \mathbb{R}^d$. Let $\underline{\theta} \in \mathbb{R}^d$ and write

$$f(\underline{y}^*) = f(\underline{y}^*; \underline{\theta}), \qquad g(\underline{\underline{z}}) = g(\underline{\underline{z}}; \underline{\theta}), \qquad k(\underline{y}^* | \underline{\underline{z}}) = k(\underline{y}^* | \underline{\underline{z}}; \underline{\theta})$$

Let the MLE be

$$\hat{\underline{\theta}}_{ML} \equiv \underset{\underline{\theta} \in \mathbb{R}^d}{\operatorname{argmax}} \frac{1}{n} \log \hat{L}_n(\underline{\theta})$$

where

$$\hat{L}_n(\underline{\theta}) = g(\underline{\underline{Z}}; \underline{\theta}) = \frac{f(\underline{Y}^*; \underline{\theta})}{k(\underline{Y}^* | \underline{\underline{Z}}; \underline{\theta})}, \qquad \Rightarrow \frac{1}{n} \log \hat{L}_n(\underline{\theta}) = \frac{1}{n} \log f(\underline{Y}^*; \underline{\theta}) - \frac{1}{n} \log k(\underline{Y}^* | \underline{\underline{Z}}; \underline{\theta})$$

Let

$$\overline{R}(\underline{\theta}; \underline{\theta}_s) \equiv E_{\underline{\theta}_s}\left[\frac{1}{n} \log f(\underline{Y}^*; \underline{\theta}) \,\middle|\, \underline{\underline{Z}} \right]$$

i.e. a conditional mean given $\underline{\underline{z}}$ where $\underline{\theta}_s$ is treated as the true value of $\underline{\theta}$.

The EM algorithm maximizes $\frac{1}{n}\log \hat{L}_n(\underline{\theta})$ by maximizing $\bar{R}(\underline{\theta};\underline{\theta}_s)$ with respect to $\underline{\theta}$. To see why this this works, let

$$\bar{S}(\underline{\theta};\underline{\theta}_s) \equiv E_{\underline{\theta}_s}\left[\frac{1}{n}\log k(\underline{Y}^*|\underline{\underline{Z}};\underline{\theta})\,\middle|\,\underline{\underline{Z}}\right]$$

and observe that

$$\bar{Q}(\underline{\theta};\underline{\theta}_s) \equiv E_{\underline{\theta}_s}\left[\frac{1}{n}\log g(\underline{\underline{Z}};\underline{\theta})\,\middle|\,\underline{\underline{Z}}\right] = \frac{1}{n}\log \hat{L}_n(\underline{\theta})$$

So

$$\hat{Q}_n(\underline{\theta}) = \frac{1}{n}\log \hat{L}_n(\underline{\theta}) = \bar{R}(\underline{\theta};\underline{\theta}_s) - \bar{S}(\underline{\theta};\underline{\theta}_s)$$

By Jensen's inequality

$$\bar{S}(\underline{\theta};\underline{\theta}_s) < \bar{S}(\underline{\theta}_s;\underline{\theta}_s), \qquad \forall \underline{\theta} \neq \underline{\theta}_s, \qquad \cdots(*)$$

This is because

$$E_{\underline{\theta}_0}\left[\log\left(\frac{f(\underline{\underline{Z}};\underline{\theta})}{f(\underline{\underline{Z}};\underline{\theta}_0)}\right)\right] < \log E_{\underline{\theta}_0}\left[\frac{f(\underline{\underline{Z}};\underline{\theta})}{f(\underline{\underline{Z}};\underline{\theta}_0)}\right] = 0, \qquad \text{when } \underline{\theta} = \underline{\theta}_0$$

since $\log E_{\underline{\theta}_0}[1] = \log 1 = 0$.
For a given $\underline{\theta}_s$, let $\underline{\theta}^*_{s+1}$ maximizes $\bar{R}(\underline{\theta},\underline{\theta}_s)$. Then

$$\hat{Q}_n(\underline{\theta}^*_{s+1}) = \bar{R}(\underline{\theta}^*_{s+1};\underline{\theta}_s) - \bar{S}(\underline{\theta}^*_{s+1};\underline{\theta}_s)$$

But

$$\bar{R}(\underline{\theta}^*_{s+1};\underline{\theta}_s) \geq \bar{R}(\underline{\theta};\underline{\theta}_s), \qquad \forall \underline{\theta} \in \mathbb{R}^d$$

while

$$\bar{S}(\underline{\theta}^*_{s+1};\underline{\theta}_s) \leq \bar{S}(\underline{\theta}_s;\underline{\theta}_s), \qquad \text{by } (*)$$

which implies that $\hat{Q}_n(\underline{\theta}^*_{s+1}) \geq \hat{Q}_n(\underline{\theta}_s)$
In addition, if $\hat{L}_n(\underline{\theta})$ is bounded, then $\lim_{s\to\infty}\frac{1}{n}\log \hat{L}_n(\underline{\theta}_s)$ exists. To see this, suppose $\underline{\theta}^*$ is a stationary point of the EM algorithm. Then $\underline{\theta}^*$ is a stationary point of $\hat{Q}_n(\underline{\theta})$.
If $\hat{Q}_n(\underline{\theta})$ is twice differentiable, then

$$\left.\frac{\partial \hat{Q}_n(\underline{\theta})}{\partial \underline{\theta}^\top}\right|_{\underline{\theta}=\underline{\theta}_s} = \left.\frac{\partial \bar{R}(\underline{\theta};\underline{\theta}_s)}{\partial \underline{\theta}^\top}\right|_{\underline{\theta}=\underline{\theta}_s} - \left.\frac{\partial \bar{S}(\underline{\theta};\underline{\theta}_s)}{\partial \underline{\theta}^\top}\right|_{\underline{\theta}=\underline{\theta}_s}$$

But $\underline{\theta}_s$ is a global maximum of $\bar{S}(\underline{\theta},\underline{\theta}_s)$, so

$$\left.\frac{\partial \bar{S}(\underline{\theta};\underline{\theta}_s)}{\partial \underline{\theta}^\top}\right|_{\underline{\theta}=\underline{\theta}_s} = \underline{0}$$

So if $\underline{\theta}_s$ is a stationary point of $\bar{R}(\underline{\theta};\underline{\theta}_s)$, it is also a stationary point of $\hat{Q}_n(\underline{\theta})$.

e.g. Consider the standard Tobit model

Suppose we observe $[D_i \ \underline{X}_i^\top \ Y_i]^\top$ (i=1,…,n) where for each i

$$\begin{aligned} Y_i^* &= \underline{X}_i^\top \underline{\beta}_0 + U_i \\ D_i &= 1\{Y_i^* > 0\} \\ Y_i &= Y_i^* D_i \end{aligned}$$

where $U_1,\ldots,U_n$ are iid $N(0,\sigma_0^2)$ and unobserved. Let

$$\begin{aligned} \underline{\theta} &\equiv \left[\underline{\beta}^\top \quad \sigma^2\right]^\top \\ \underline{\theta}_0 &\equiv \left[\underline{\beta}_0^\top \quad \sigma^2\right]^\top \\ \underline{Y}^* &\equiv [Y_1^* \quad \cdots \quad Y_n^*]^\top \\ \underline{\underline{Z}} &= \begin{bmatrix} D_1 & \underline{X}_1^\top & Y_1 \\ \vdots & \vdots & \vdots \\ D_n & \underline{X}_n^\top & Y_n \end{bmatrix} \end{aligned}$$

We have

$$\begin{aligned} \log f(\underline{Y}^*;\underline{\theta}) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i^* - \underline{X}_i^\top\underline{\beta}\right)^2 \\ E_{\underline{\theta}_s}\left[\frac{1}{n}\log f(\underline{Y}^*;\underline{\theta})\middle|\underline{\underline{Z}}\right] &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2 n}\sum_{\{i:D_i=1\}}\left(Y_i - \underline{X}_i^\top\underline{\beta}\right)^2 \\ &\quad - \frac{1}{2\sigma^2 n}\sum_{\{i;D_i=0\}} E_{\underline{\theta}_s}\left[\left(Y_i^* - \underline{X}_i^\top\underline{\beta}\right)^2\middle| D_i = 0, \underline{X}_i\right] \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 \\ &\quad - \frac{1}{2n\sigma^2}\sum_{\{i;D_i=1\}}\left(Y_i - \underline{X}_i^\top\underline{\beta}\right)^2 - \frac{1}{2\sigma^2 n}\sum_{\{i;D_i=0\}} \mathrm{Var}_{\underline{\theta}_s}[Y_i^*|D_i=0,\underline{X}_i] \\ &\quad - \frac{1}{2\sigma^2 n}\sum_{\{i;D_i=0\}}\left(E_{\underline{\theta}_s}[Y_i^*|D_i=0,\underline{X}_i] - \underline{X}_i^\top\underline{\beta}\right)^2, \qquad \cdots (**) \end{aligned}$$

where

$$\begin{aligned} E_{\underline{\theta}_s}[Y_i^*|D_i=0,\underline{X}_i] &= \underline{X}_i^\top\underline{\beta}_s - \frac{\sigma_s\phi_{is}}{1-\Phi_{is}} \equiv \overline{Y}_{i0} \\ \mathrm{Var}_{\underline{\theta}_s}[Y_i^*|D_i=0,\underline{X}_i] &= \sigma_s^2 + \underline{X}_i^\top\underline{\beta}_s\frac{\sigma_s\phi_{is}}{1-\Phi_{is}} - \left(\frac{\sigma_s\phi_{is}}{1-\Phi_{is}}\right)^2 \end{aligned}$$

where

$$\phi_{is} = \phi\left(\frac{\underline{X}_i^\top\underline{\beta}_s}{\sigma_s}\right), \qquad \Phi_{is} = \Phi\left(\frac{\underline{X}_i^\top\underline{\beta}_s}{\sigma_s}\right)$$

(e.g. Hayashi (2000), §8.2)

Since $Y_i^*|\underline{X}_i \sim N\left(\underline{X}_i^\top \underline{\beta}_s, \sigma_s^2\right)$ when computing $E_{\underline{\theta}_s}[Y_i^*|D_i = 0, \underline{X}_i]$ and $Var_{\underline{\theta}_s}[Y_i^*|D_i = 0, \underline{X}_i]$, from (**), $\underline{\beta}_{s+1}$ can be computed as follows.

Let $\underline{Y}$ be the vector consisting of $Y_1,\ldots,Y_n$; assumed all positive WOLOG.

Let $\underline{\overline{Y}}_0 \equiv [\overline{Y}_{i0}]$ consist of $n-n_1$ elements.

Let $\underline{\underline{X}} \equiv [\underline{X}_1 \ \ldots \ \underline{X}_n]^\top$ as usual.

Then by (**),

$$E_{\underline{\theta}_s}\left[\frac{1}{n}\log f(\underline{Y}^*;\underline{\theta})\,\middle|\,\underline{\underline{Z}}\right]$$

is maximized w.r.t. $\underline{\beta}$ by OLS, i.e.

$$\underline{\beta}_{s+1} = (\underline{\underline{X}}^\top \underline{\underline{X}})^{-1}\underline{\underline{X}}^\top \begin{bmatrix}\underline{Y}\\ \underline{\overline{Y}}^0\end{bmatrix}, \qquad \cdots(\dagger)$$

Similarly, by (**),

$$E_{\underline{\theta}_s}\left[\frac{1}{n}\log f(\underline{Y}^*;\underline{\theta})\,\middle|\,\underline{\underline{Z}}\right]$$

is maximized w.r.t. σ^2 by OLS i.e.

$$\sigma_{s+1}^2 = \frac{1}{n}\left[\sum_{\{i:D_i=1\}}\left(Y_i - \underline{X}_i^\top\underline{\beta}_{s+1}\right)^2 + \sum_{\{i:D_i=0\}}\left(\overline{Y}_{i0} - \underline{X}_i^\top\underline{\beta}_{s+1}\right)^2 + \sum_{\{i:D_i=0\}} Var_{\underline{\theta}_s}[Y_i^*|D_i = 0, \underline{X}_i]\right], \qquad \cdots(\ddagger)$$

We can show that $\underline{\hat{\theta}}_{ML}$ is the equilibrium solution of the iterative progress given by (†), (‡). Partition

$$\underline{\underline{X}} = \begin{bmatrix}\underline{\underline{X}}_1\\ \underline{\underline{X}}_0\end{bmatrix}$$

where $\underline{\underline{X}}_1$ is $(n_1\times d)$ and $\underline{\underline{X}}_0$ is $[(n-n_1)\times d]$. Insert $\underline{\hat{\theta}}_{ML} \equiv \left[\underline{\hat{\beta}}_{ML}^\top \quad \hat{\sigma}_{ML}^2\right]^\top$ into both sides of (†) to get

$$\underline{\underline{X}}_1^\top\underline{\underline{X}}_1\underline{\hat{\beta}}_{ML} \equiv \underline{\underline{X}}_1^\top\underline{Y} - \underline{\underline{X}}_0^\top\underline{\hat{\gamma}}_{ML} \Leftrightarrow -\underline{\underline{X}}_0^\top\underline{\hat{\gamma}}_{ML} = -\underline{\underline{X}}_1^\top\left(\underline{Y} - \underline{\underline{X}}_1\underline{\hat{\beta}}_{ML}\right), \qquad \cdots(***)$$

Where $\underline{\hat{\gamma}}_{ML}$ is an $[(n-n_1)\times 1]$-vector whose j-th element is

$$\frac{\hat{\sigma}_{ML}\phi\left(\underline{X}_j^\top\underline{\hat{\beta}}_{ML}/\hat{\sigma}_{ML}\right)}{1-\Phi\left(\underline{X}_j^\top\underline{\hat{\beta}}_{ML}/\hat{\sigma}_{ML}\right)}$$

We know that if $\frac{\partial}{\partial\underline{\beta}^\top}\frac{1}{n}\log\hat{L}_n(\underline{\theta})\Big|_{\underline{\theta}=\underline{\hat{\theta}}_{ML}} = \underline{0}$, so

$$-\sigma\sum_{\{i:D_i=0\}}\frac{\phi_i}{1-\Phi_i}\underline{X}_i + \sum_{\{i:D_i=1\}}\left(Y_i - \underline{X}_i^\top\underline{\beta}\right)\underline{X}_i = \underline{0}$$

which is equivalent to (***). This shows that the $\underline{\beta}$-limit of the EM algorithm is the MLE of $\underline{\beta}_0$.

(Exercise: Show that the solution to $\frac{\partial}{\partial\sigma^2}\frac{1}{n}\log\hat{L}_n(\underline{\theta}) = 0$ is the σ^2-limit of the EM algorithm)

- Binary response models
- Multinomial response models
- Models with censoring/truncation

Suppose $[\underline{X}^\intercal\ Y]^\intercal$ is a random vector with

$$P[Y=1|\underline{X}] = g(\underline{X})$$
$$P[Y=0|\underline{X}] = 1-g(\underline{X})$$

where g(.) is a Borel measurable function into (0,1). Different models of binary response involve different specifications of the link function g(.).

Typically one assumes a parametrization, i.e. $g(\underline{x})=g(\underline{x};\underline{\theta}_0)$ for some parameter $\underline{\theta}_0 \in \Theta$.

Typically use ML to estimate $\underline{\theta}_0$.

Where do binary response models come from?

- Standard answer: Involves a latent variable model.

Let Y^* be a continuous random variable that is only partially observed (i.e. Y^* is latent). In particular, assume that we observe

$$Y = \begin{cases} 1, & Y^* > 0 \\ 0, & Y^* \le 0 \end{cases}$$

In addition, usually assume that Y^* is related to observable regressors $\underline{X}$ via an index model.

$$Y^* = \underline{X}^\intercal\underline{\beta}_0 - U, \qquad \cdots (*)$$

where $\underline{\beta}_0$ is an unknown vector of constants and U is an unobservable random variable.
(We do not observe Y^*, only observe $Y \equiv 1\{Y^*>0\}$.)

From (*) we have

$$\begin{aligned} P[Y=1|\underline{X}] &= P[Y^*>0] \\ &= P\left[\underline{X}^\intercal\underline{\beta}_0 - U > 0\right] \\ &= F_U\left(\underline{X}^\intercal\underline{\beta}_0\right) \end{aligned}$$

where F_U is the cdf of U. In the context of (*), we see that different models of binary response are characterized by different distributions of U.

e.g. Suppose U~logistic, i.e. U is continuous on $\mathbb{R}$ with density

$$f_U(u) = \frac{e^u}{(1+e^u)^2}$$

which implies that

$$P[Y=1|\underline{X}] = F_U\left(\underline{X}^\intercal\underline{\beta}_0\right) = \frac{e^{\underline{X}^\intercal\underline{\beta}_0}}{1+e^{\underline{X}^\intercal\underline{\beta}_0}}$$

i.e. a logit model.

e.g. Suppose $U \sim N(0,1)$. Then $P[Y=1|\underline{X}]=\Phi(\underline{X}^\top \underline{\beta}_0)$ i.e. a probit model.

Identification of $\underline{\beta}_0$ in a latent index model

$$Y^* = \underline{X}^\top \underline{\beta}_0 - U$$

What we typically impose are restrictions on the variance and mean of U_i.

Observe that $Y=1\{Y^*>0\}=1\{\underline{X}^\top \underline{\beta}_0 - U > 0\}$, but $1\{\underline{X}^\top \underline{\beta}_0 - U > 0\}=1\{c\underline{X}^\top \underline{\beta}_0 - cU > 0\}$ for any $c>0$. This means that observations generated by parameter $\underline{\beta}_0$ and latent error U are indistinguishable from observations generated by parameter $c\underline{\beta}_0$ and latent error cU for any $c>0$.

This problem is solved by fixing the variance of the latent error.

e.g. In a logit model, assume that the latent error has variance $\pi^2/3$, i.e. the variance of a logistic random variable. So, if $U \sim$ logistic and $Var[cU]=\pi^2/3$, then $c=1$.

In addition, note that $\forall$ constant d

$$1\{\underline{X}^\top \underline{\beta}_0 - U > 0\} = 1\{d + \underline{X}^\top \underline{\beta}_0 - U - d > 0\}$$

which shows that it is impossible to distinguish between observations generated by parameter $\underline{\beta}_0$ and latent error U and observations generated by parameter $[d \quad \underline{\beta}_0^\top]^\top$ and latent error U−d.

This problem can be solved by requiring the latent error to have mean zero.

e.g. If $U \sim N(0,1)$ and $E[U-d]=0$ then $d=0$.

An important, special case of a latent index model is an additive random-utility model (ARUM). Suppose a consumer has to pick between two alternatives (0 and 1). Observe

$$Y = \begin{cases} 0, & \text{if alternative 0 has higher utility} \\ 1, & \text{if alternative 1 has higher utility} \end{cases}$$

An ARUM specifies the utilities of alternatives 0,1, as

$$U_0 = V_0 + \varepsilon_0$$
$$U_1 = V_1 + \varepsilon_1$$

Call V_0, V_1 the deterministic components of utility; ε_0, ε_1 the random components of utility, then

$$\begin{aligned} P[Y=1] &= P[U_1 \geq U_0] \\ &= P[V_1 - V_0 \geq \varepsilon_0 - \varepsilon_1] \\ &= F_{\varepsilon_0-\varepsilon_1}(V_1 - V_0) \end{aligned}$$

where $F_{\varepsilon_0-\varepsilon_1}$ is the cdf of $\varepsilon_0-\varepsilon_1$. Frequently assume that $V_1 - V_0 = \underline{X}^\top \underline{\beta}_0$ for some observable random vector $\underline{X}$ and some unobservable constant vector $\underline{\beta}_0$. Identification of $\underline{\beta}_0$ is guaranteed by assuming a fixed value (e.g. 1 or $\pi^2/3$) for the difference in random utility components and also a mean of zero for the difference in random utility components.

In the context of binary choice modeling often encounter the assumption that ε_0, ε_1 have type-I extreme-value distributions, i.e. ε_j (j=0,1) is continuous on $\mathbb{R}$ with density

$$f_\varepsilon(\varepsilon') = e^{-\varepsilon'} e^{-e^{-\varepsilon'}}$$

which implies a cdf $F_\varepsilon(\varepsilon') = e^{-e^{-\varepsilon'}}$.

Can show that if ε_0, ε_1 are independent and type-I extreme-value, then $\varepsilon_0-\varepsilon_1\sim$Logistic, and the binary-choice model $P[Y=1]=F_{\varepsilon_0-\varepsilon_1}(V_1-V_0)$ is a logit.

Often specifies ARUMs where the deterministic component of utility for alternative $j\in\{0,1\}$ is

$$V_{ij}=\underline{Z}_{ij}^{\top}\underline{\alpha}_j+\underline{W}_i^{\top}\underline{\gamma}_j$$

for individual i.

In this case, $\underline{Z}_{ij}$ are regressors that vary across alternatives (e.g. interactions between individual characteristics and product attributes) and $\underline{W}_i$ are individual characteristics that are fixed across alternatives (e.g. income/gender/schooling). Then the binary-choice model has the form

$$P[Y_i=1|\underline{Z}_{i0},\underline{Z}_{i1},\underline{W}_i]=F_{\varepsilon_0-\varepsilon_1}\left(\underline{Z}_{i1}^{\top}\underline{\alpha}_1-\underline{Z}_{i0}^{\top}\underline{\alpha}_0+\underline{W}_i^{\top}\left(\underline{\gamma}_1-\underline{\gamma}_0\right)\right)$$

which show that $\underline{\gamma}_1$, $\underline{\gamma}_0$ cannot be separately identified.

General problem of MLE of a binary response model

Suppose we observe a random sample $[\underline{X}_i^{\top}\ Y_i]^{\top}$ (i=1,…,n) from a model with

$$P[Y_i=1|\underline{X}_i]=G\left(\underline{X}_i^{\top}\underline{\beta}_0\right)$$
$$P[Y_i=0|\underline{X}_i]=1-G\left(\underline{X}_i^{\top}\underline{\beta}_0\right)$$

for some constant vector $\underline{\beta}_0\in\mathbb{R}^d$ and where $G:\mathbb{R}\to(0,1)$ is a positive, non-decreasing and continuously differentiable function on $\mathbb{R}$.

Then (obviously) $Y_1,\ldots,Y_n$ are discrete with probability mass function

$$f\left(y|\underline{X}_i;\underline{\beta}_0\right)=\begin{cases}G\left(\underline{X}_i^{\top}\beta_0\right)^y\left(1-G\left(\underline{X}_i^{\top}\underline{\beta}_0\right)\right)^{1-y}, & y\in\{0,1\}\\ 0, & \text{otherwise}\end{cases}$$

Evaluate the pmf at an actual observation Y_i ($i\in\{1,\ldots,n\}$) replace $\underline{\beta}_0$ with an arbitrary $\underline{\beta}\in\mathbb{R}^d$ to get

$$\log f\left(Y_i|\underline{X}_i;\underline{\beta}\right)=Y_i\log G\left(\underline{X}_i^{\top}\underline{\beta}\right)+(1-Y_i)\log\left(1-G\left(\underline{X}_i^{\top}\underline{\beta}\right)\right)$$

which implies the log likelihood

$$\begin{aligned}\hat{L}_n\left(\underline{\beta}\right)&=\sum_{i=1}^{n}\log f\left(Y_i|\underline{X}_i;\underline{\beta}\right)\\&=\sum_{i=1}^{n}\left[Y_i\log G\left(\underline{X}_i^{\top}\underline{\beta}\right)+(1-Y_i)\log\left(1-G\left(\underline{X}_i^{\top}\underline{\beta}\right)\right)\right]\end{aligned}$$

By differentiating

$$\frac{1}{n}\frac{\partial}{\partial\underline{\beta}^{\top}}\hat{L}_n\left(\underline{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i}{G\left(\underline{X}_i^{\top}\underline{\beta}\right)}G'\left(\underline{X}_i^{\top}\underline{\beta}\right)X_i-\frac{1-Y_i}{1-G\left(\underline{X}_i^{\top}\underline{\beta}\right)}G'\left(\underline{X}_i^{\top}\underline{\beta}\right)\underline{X}_i\right)$$

The MLE $\hat{\underline{\beta}}_n$ is given by

$$\frac{1}{n}\frac{\partial}{\partial \underline{\beta}^\top}\hat{L}_n(\underline{\beta})\Big|_{\underline{\beta}=\hat{\underline{\beta}}_n} = \underline{0} \Leftrightarrow \sum_{i=1}^{n}\frac{Y_i - G(\underline{X}_i^\top \hat{\underline{\beta}}_n)}{G(\underline{X}_i^\top \hat{\underline{\beta}}_n)\left(1 - G(\underline{X}_i^\top \hat{\underline{\beta}}_n)\right)}G'(\underline{X}_i^\top \hat{\underline{\beta}}_n)\underline{X}_i = 0, \qquad \cdots (**)$$

We can see that in general, there is no closed form for $\hat{\underline{\beta}}_n$.

Note: If G is the logistic or N(0,1)-cdf then $\hat{L}_n(\underline{\beta})$ is globally (exercise) and Newton–Raphson method converges quickly.

When is the MLE $\hat{\underline{\beta}}_n$ consistent?

Need the "success probability" $G(\underline{X}_i^\top \underline{\beta}_0)$ to be correctly specified.

Since Y_i is Bernoulli, we have

$$E[Y_i|\underline{X}_i] = P[Y_i = 1|\underline{X}_i]$$

so we need $P[Y_i = 1|\underline{X}_i] = G(\underline{X}_i^\top \underline{\beta}_0)$ if

$$\frac{1}{n}E\left[\frac{\partial}{\partial \underline{\beta}^\top}\hat{L}_n(\underline{\beta})\Big|_{\underline{\beta}=\underline{\beta}_0}\right] = \underline{0} \Leftrightarrow E\left[\frac{E[Y_1|\underline{X}_1] - G(\underline{X}_1^\top \underline{\beta}_0)}{G(\underline{X}_1^\top \underline{\beta}_0)\left(1 - G(\underline{X}_1^\top \underline{\beta}_0)\right)}G'(\underline{X}_1^\top \underline{\beta}_0)\underline{X}_1\right] = \underline{0}$$

Observe a random sample

$$[\underline{X}_i^\top \quad Y_i]^\top, \qquad i = 1, \cdots, n$$

from a model with

$$P[Y_i = 1|\underline{X}_i] = G(\underline{X}_i^\top \underline{\beta}_0)$$
$$P[Y_i = 0|\underline{X}_i] = 1 - G(\underline{X}_i^\top \underline{\beta}_0)$$

for some constant vector $\underline{\beta}_0 \in \mathbb{R}^d$ and some non-decreasing continuously differentiable function $G: \mathbb{R} \to (0,1)$. Then the pmf of each Y_i is

$$f(y|\underline{X}_i; \underline{\beta}_0) = \begin{cases} G(\underline{X}_i^\top \underline{\beta}_0)^y \left(1 - G(\underline{X}_i^\top \underline{\beta}_0)\right)^{1-y}, & y \in \{0,1\} \\ 0, & \text{otherwise} \end{cases}$$

which implies the log-likelihood

$$\hat{L}_n(\underline{\beta}) \equiv \sum_{i=1}^n \log f(Y_i|\underline{X}_i; \underline{\beta})$$
$$= \sum_{i=1}^n \left[Y_i \log G(\underline{X}_i^\top \underline{\beta}) + (1 - Y_i) \log\left(1 - G(\underline{X}_i^\top \underline{\beta})\right)\right]$$
$$\Rightarrow \frac{1}{n} \frac{\partial}{\partial \underline{\beta}^\top} \hat{L}_n(\underline{\beta}) = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i}{G(\underline{X}_i^\top \underline{\beta})} - \frac{1 - Y_i}{1 - G(\underline{X}_i^\top \underline{\beta})}\right) G'(\underline{X}_i^\top \underline{\beta}) \underline{X}_i$$

The MLE $\hat{\underline{\beta}}_n$ is implicitly defined as

$$\frac{1}{n} \frac{\partial}{\partial \underline{\beta}^\top} \hat{L}_n(\underline{\beta})\Big|_{\underline{\beta} = \hat{\underline{\beta}}_n} = \underline{0} \Leftrightarrow \sum_{i=1}^n \frac{Y_i - G(\underline{X}_i^\top \hat{\underline{\beta}}_n)}{G(\underline{X}_i^\top \hat{\underline{\beta}}_n)\left(1 - G(\underline{X}_i^\top \hat{\underline{\beta}}_n)\right)} G'(\underline{X}_i^\top \hat{\underline{\beta}}_n) \underline{X}_i = \underline{0}$$

($\hat{\underline{\beta}}_n$ in general has no closed form)

Note: With logit or probit, G(.) is globally concave and Newton–Raphson will converge quickly.

Consistency requires that G(.) be correctly specified: This has to do with the necessary condition that

$$E\left[\frac{\partial}{\partial \underline{\beta}^\top} \hat{L}_n(\underline{\beta})\Big|_{\underline{\beta} = \underline{\beta}_0}\right] = \underline{0}$$

which implies that

$$E[Y_i|\underline{X}_i] = P[Y_i = 1|\underline{X}_i] = G(\underline{X}_i^\top \underline{\beta}_0)$$

Asymptotic Normality in the standard case requires that G(.) to be correctly specified and twice continuously differentiable.

Under these assumptions, we can make the standard argument that the MLE is $\sqrt{n}$-consistent and asymptotically Normal with

$$\sqrt{n}\left(\hat{\underline{\beta}}_n - \underline{\beta}_0\right) \xrightarrow{d} N_d\left(\underline{0}, \left(\operatorname{plim}\left(-\frac{1}{n}\frac{\partial^2}{\partial \underline{\beta}\,\partial \underline{\beta}^\top}\hat{L}_n\left(\underline{\beta}\right)\Big|_{\underline{\beta}=\underline{\beta}_0}\right)\right)^{-1}\right)$$

Note that

$$\begin{aligned}\frac{1}{n}\frac{\partial^2}{\partial \underline{\beta}\,\partial \underline{\beta}^\top}\hat{L}_n\left(\underline{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&-\frac{Y_i}{G^2\left(\underline{X}_i^\top\underline{\beta}\right)}G'\left(\underline{X}_i^\top\underline{\beta}\right)^2\underline{X}_i\underline{X}_i^\top + \frac{Y_i}{G\left(\underline{X}_i^\top\underline{\beta}\right)}G''\left(\underline{X}_i^\top\underline{\beta}\right)\underline{X}_i\underline{X}_i^\top \\ &+\frac{1-Y_i}{\left(1-G\left(\underline{X}_i^\top\underline{\beta}\right)\right)^2}G'\left(\underline{X}_i^\top\underline{\beta}\right)^2\underline{X}_i\underline{X}_i^\top - \frac{1-Y_i}{1-G\left(\underline{X}_i^\top\underline{\beta}\right)}G''\left(\underline{X}_i^\top\underline{\beta}\right)\underline{X}_i\underline{X}_i^\top\Bigg)\end{aligned}$$

This implies that

$$\begin{aligned}\frac{1}{n}\frac{\partial^2}{\partial \underline{\beta}\,\partial \underline{\beta}^\top}\hat{L}_n\left(\underline{\beta}\right)\Big|_{\beta=\underline{\beta}_0} &\xrightarrow{p} -E\left[\frac{G'\left(\underline{X}_1^\top\underline{\beta}_0\right)^2}{G\left(\underline{X}_1^\top\underline{\beta}_0\right)}\underline{X}_1\underline{X}_1^\top + G''\left(\underline{X}_1^\top\underline{\beta}_0\right)\underline{X}_1\underline{X}_1^\top\right. \\ &\quad\left.+\frac{G'\left(\underline{X}_1^\top\underline{\beta}_0\right)^2}{1-G\left(\underline{X}_1^\top\underline{\beta}_0\right)}\underline{X}_1\underline{X}_1^\top - G''\left(\underline{X}_1^\top\underline{\beta}_0\right)\underline{X}_1\underline{X}_1^\top\right] \\ &= -E\left[\frac{G'\left(\underline{X}_1^\top\underline{\beta}_0\right)^2}{G\left(\underline{X}_1^\top\underline{\beta}_0\right)\left(1-G\left(\underline{X}_1^\top\underline{\beta}_0\right)\right)}\underline{X}_1\underline{X}_1^\top\right]\end{aligned}$$

Therefore a consistent estimator of the asymptotic variance of $\hat{\underline{\beta}}_n$ is

$$\hat{\underline{\underline{V}}}\left[\hat{\underline{\beta}}_n\right] = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{G'\left(\underline{X}_i^\top\hat{\underline{\beta}}_n\right)^2}{G\left(\underline{X}_1^\top\hat{\underline{\beta}}_n\right)\left(1-G\left(\underline{X}_1^\top\hat{\underline{\beta}}_n\right)\right)}\underline{X}_i\underline{X}_i^\top\right]^{-1}$$

More on logit: Logit satisfies

$$G\left(\underline{X}^\top\underline{\beta}\right) = \frac{e^{\underline{X}^\top\underline{\beta}}}{1+e^{\underline{X}^\top\underline{\beta}}}$$

The corresponding FOC are

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{e^{\underline{X}^\top\underline{\beta}}}{1+e^{\underline{X}^\top\underline{\beta}}}\right)\underline{X}_i = \underline{0}$$

because $G'\left(\underline{X}^\top\underline{\beta}\right) = G\left(\underline{X}^\top\underline{\beta}\right)\left(1-G\left(\underline{X}^\top\underline{\beta}\right)\right)$

If $\underline{X}$ includes a constant term (i.e. $\underline{X}^\top\underline{\beta}$ includes an intercept), then the FOCs imply that

$$\sum_{i=1}^{n}\left(Y_i - \frac{e^{\underline{X}^\top\underline{\beta}}}{1+e^{\underline{X}^\top\underline{\beta}}}\right) = 0 \Leftrightarrow \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} G\left(\underline{X}_i^\top\underline{\hat{\beta}}_n\right)$$

i.e. the sample proportion $\bar{Y}$ is equal to the in-sample predicted probability.
Note also that

$$G\left(\underline{X}^\top\underline{\beta}\right) = \frac{e^{\underline{X}^\top\underline{\beta}}}{1+e^{\underline{X}^\top\underline{\beta}}} \Leftrightarrow e^{\underline{X}^\top\underline{\beta}} = \frac{G\left(\underline{X}^\top\underline{\beta}\right)}{1-G\left(\underline{X}^\top\underline{\beta}\right)} \Leftrightarrow \log\frac{G\left(\underline{X}^\top\underline{\beta}\right)}{1-G\left(\underline{X}^\top\underline{\beta}\right)} = \log\frac{P[Y=1|\underline{X}]}{P[Y=0|\underline{X}]} = \underline{X}^\top\underline{\beta}$$

Therefore if β_j is the jth component of $\underline{\beta}$ and X_j is the jth component of $\underline{X}$ then β_j is the change in the log-odds ratio. When X_j increases by one unit and all other components of $\underline{X}$ are constant
e.g. Suppose $\Delta X_j=1$. Then $e^{\underline{X}^\top\underline{\beta}}$ increases to $e^{\underline{X}^\top\underline{\beta}+\beta_j} = e^{\underline{X}^\top\underline{\beta}}e^{\beta_j}$, i.e. the odds ratio increases by e^{β_j}. So if $\beta_j=.10$, then $\Delta X_j=1$ multiplies the odds ratio by $e^{.10}\approx 1.105$, i.e. a 10.5% increase.

More on the Probit model
Probit specifies $G\left(\underline{X}^\top\underline{\beta}\right) = \Phi\left(\underline{X}^\top\underline{\beta}\right)$ where $\Phi(.)$ is the N(0,1)-cdf. The corresponding FOCs are

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\varphi\left(\underline{X}_i^\top\underline{\beta}\right)}{\Phi\left(\underline{X}_i^\top\underline{\beta}\right)\left(1-\Phi\left(\underline{X}_i^\top\underline{\beta}\right)\right)}\left(Y_i - \Phi\left(\underline{X}_i^\top\underline{\beta}\right)\right)\underline{X}_i = \underline{0}$$

(compare with the logit, OLS) Also,

$$\frac{\partial}{\partial X_j}G\left(\underline{X}^\top\underline{\beta}\right) = \varphi\left(\underline{X}^\top\underline{\beta}\right)\beta_j$$

So β_j is not intuitively interpretable.

What about OLS? OLS in this context assumes a linear probability model, i.e. the model

$$P[Y=1|\underline{X}] = \underline{X}^\top\underline{\beta}_0$$
$$P[Y=0|\underline{X}] = 1-\underline{X}^\top\underline{\beta}_0$$

which implies $E[Y|\underline{X}] = \underline{X}^\top\underline{\beta}_0$.

OLS will work as expected, but generally yield predicted probabilities $\underline{X}^\top\underline{\hat{\beta}}_{OLS}$ that are not in the interval [0,1].

If OLS is to be used, then heteroskedasticity-consistent standard errors should be used, since

$$Var[Y|\underline{X}] = \underline{X}^\top\underline{\beta}_0\left(1-\underline{X}^\top\underline{\beta}_0\right)$$

which varies with $\underline{X}$.

How to choose between logit or probit?

In the first instance, we should choose depending on how confident we are in the assumed functional form of G(.), i.e.

$$\text{if } P[Y = 1|\underline{X}] = F\left(\underline{X}^{\top}\underline{\beta}_0\right)$$

where F(.) is the logistic or N(0,1)-cdf, then logit or probit should be used as appropriately.

In the second instance, misspecification of G(.) may not be serious.

Ruud (1983, EMA): Sufficient condition for scale consistency in this context is that for each j

$$E\left[X_j\middle|\underline{X}^{\top}\underline{\beta}_0 = z\right] \text{ is linear in } z$$

Scale consistency: i.e. $\hat{\underline{\beta}}_n \xrightarrow{p} c\underline{\beta}_0$ for some constant c

Unfortunately, the condition that $E\left[X_j\middle|\underline{X}^{\top}\underline{\beta}_0 = z\right]$ is linear in z is hard to verify.

In practice: Usually little difference between predicted probabilities coming from logit or probit, with the exception of predicted probabilities near 0 or 1.

Diagnostic measures

a) Pseudo-R^2

A popular measure is the McFadden (1974, *Frontiers in Econometrics*) Pseudo-R^2, i.e.

$$R_M^2 \equiv 1 - \frac{\hat{L}_n\left(\hat{\underline{\beta}}_n\right)}{\hat{L}_n(\overline{Y})} = 1 - \frac{\sum_{i=1}^{n}\left[Y_i \log G\left(\underline{X}_i^{\top}\hat{\underline{\beta}}_n\right) + (1 - Y_i)\log\left(1 - G\left(\underline{X}_i^{\top}\hat{\underline{\beta}}_n\right)\right)\right]}{n\overline{Y}\log\overline{Y} + (1 - \overline{Y})\log(1 - \overline{Y})}$$

Suggestion: Compare logit/probit/OLS with a "distribution-free" method that allows for misspecification of G(.). Refer Klein and Spady (1993, EMA), "Semiparametric MLE" and Ichimura (1993, JEM), "Semiparametric Least Squares" for "distribution-free" cases.

Values of R_M^2 close to one indicate a fitted binary response model that predicts Y=1 with probability close to $G\left(\underline{X}_i^{\top}\hat{\underline{\beta}}_n\right)$ and Y=0 with probability close to $1 - G\left(\underline{X}_i^{\top}\hat{\underline{\beta}}_n\right)$.

b) Quality of predicted outcomes

Let $\hat{Y} \in \{0,1\}$ be a predicted value of Y. Can plot a receiver operating characteristics (ROC) curve by setting $\hat{Y}=1$ if $G\left(\underline{X}_i^{\top}\hat{\underline{\beta}}_n\right) > \tau$ for some $\tau \in (0,1)$; $\hat{Y}=0$ otherwise.

The ROC curve plots

$$\frac{\#\{Y = 1 \text{ values correctly predicted}\}}{\#\{Y = 0 \text{ values correctly predicted}\}} \text{ as a function of } \tau.$$

So if $\tau=1$, then all Y=0 values are correctly predicted, and all Y=1 values are incorrectly predicted, i.e. the ROC curve is equal to 0 at $\tau=1$; similarly, the ROC curve is ∞ at $\tau=0$. Ideally want a ROC curve close to 1 for all $\tau \in (0,1)$.

Example

- Sample of n=630 recreational fishers
- For each $i \in \{1,\ldots,630\}$

$$\text{Let } Y_i = \begin{cases} 1, & \text{if fishing from a charter boat} \\ 0, & \text{if fishing from a pier} \end{cases}, \qquad X_i \equiv \log \text{RELP}_i \equiv \log\left(\frac{P_{\text{Charter},i}}{P_{\text{Pier},i}}\right)$$

- Obviously: Expect the incidence of charter boat fishing to be decreasing in X_i
- We observe $\overline{Y} \equiv \frac{1}{n}\sum Y_i = .717$
- For observations i with $Y_i=1$, the average RELP was 75/121≈.6198. So price apparently has the expected effect
- We also find that for observations i with $Y_i=0$, the average RELP was 110/31≈3.5484

How to model the relationship between Y_i and X_i?

a) OLS. Regress Y_i on a constant and X_i (cannot constrain predicted choice probabilities to be between 0 and 1): Intercept=.784, Slope=.243
b) Logit (the assumption is that

$$P[Y_i = 1|X_i] = \frac{e^{\beta_{01}+\beta_{02}X_i}}{1 + e^{\beta_{01}+\beta_{02}X_i}}$$

for some constants β_{01} and β_{02}): $\hat{\beta}_{01} = 2.035, \hat{\beta}_{02} = -1.823.$

The implie estimate of the marginal effect of X_i on the choice of probability is

$$\frac{\hat{\beta}_{02}e^{\hat{\beta}_{01}+\hat{\beta}_{02}X_i}}{\left(1 + e^{\hat{\beta}_{01}+\hat{\beta}_{02}X_i}\right)^2} < 0$$

for all i as expected. The estimated average marginal effect

$$\overline{Y}(1 - \overline{Y})\hat{\beta}_{02} = -.370$$

c) Probit (the assumption is that

$$P[Y_i = 1|X_i] = \Phi(\beta_{01} + \beta_{02}X_i)$$

for constants β_{01}, β_{02}): $\hat{\beta}_{01} = 1.194, \hat{\beta}_{02} = -1.056$. The estimated marginal effect is

$$\phi\left(\hat{\beta}_{01} + \hat{\beta}_{02}X_i\right)\hat{\beta}_{02} < 0$$

for all i as expected.

In this example, the logit and probit predicted choice probabilities are similar for all X_i, which is often.

Multinomial response models

Setup: m mutually exclusive alternatives. The dependent variable for each observation i is conditionally multinomial given regressors $\underline{X}_i$, i.e.

$$Y_i = j, \qquad \text{if i chooses alternative } j \in \{1, \dots, m\}$$

e.g. Mode of commuting

$$Y_i = \begin{cases} 1, & \text{if drives own car} \\ 2, & \text{if bus} \\ 3, & \text{if walk} \\ 4, & \text{if bike} \end{cases}$$

Now define binary indicators

$$Y_{ij} = 1\{Y_i = j\}, \qquad \text{for } j \in \{1, \dots, m\}$$

Therefore, the (multinomial) density for a single observation is

$$f_Y(y) \equiv \prod_{j=1}^{m} P_{ij}^{y_{ij}}$$

where $P_{ij}=P[Y_i=j|\underline{X}_i]=P[Y_{ij}=1|\underline{X}_i]$. Typically model each P_{ij} as

$$P_{ij} = G_j(\underline{X}_i, \underline{\beta}_0)$$

for some unknown constant $\underline{\beta}_0$. Call $G_j(\underline{x},\underline{\beta})$ (j=1,…,m) the multinomial link functions. Different specifications of link function correspond to different models of multinomial response.

Maximum likelihood estimation

Suppose for individual i and alternative j

$$P_{ij} = P[Y_{ij} = 1 | \underline{X}_i] = G(\underline{X}_i, \underline{\beta}_0)$$

Then the log-likelihood is

$$\hat{L}_n(\underline{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \log G_j(\underline{X}_i, \underline{\beta})$$

(assuming a sample of n individuals)

In many applications, individual i faces a mean of m_i alternatives, so

$$\hat{L}_n(\underline{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} Y_{ij} \log G_j(\underline{X}_i, \underline{\beta})$$

FOCs given by

$$\frac{1}{n} \frac{\partial}{\partial \underline{\beta}^\top} \hat{L}_n(\underline{\beta}) \Big|_{\beta=\hat{\beta}_n} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{Y_{ij}}{G_j(\underline{X}_i, \hat{\beta}_n)} \frac{\partial}{\partial \underline{\beta}^\top} G_j(\underline{X}_i, \beta) \Big|_{\beta=\hat{\beta}_n} = \underline{0}$$

Sufficient for consistency is the score identity, which holds if $P_{ij}=G_j(\underline{X}_i,\underline{\beta}_0)$, i.e. correct specification of each link function (Exercise: Under correct specification $\sum_{j=1}^{m_i} G_j(\underline{X}_i,\underline{\beta}_0)=1$ for each i).

Note: The standard consistency argument for MLE on the score identity relies on $G_j(\underline{X}_i,\underline{\beta}_0)$ being correctly specified and $G_j(\underline{X}_i,\underline{\beta}_0)$ being continuously differentiable at $\underline{\beta}=\underline{\beta}_0$.

If link functions are correctly specified and twice continuously differentiable at $\underline{\beta}=\underline{\beta}_0$, then the standard argument shows that the MLE $\hat{\underline{\beta}}_n$ is $\sqrt{n}$-consistent and asymptotically Normal with

$$\sqrt{n}\left(\hat{\underline{\beta}}_n-\underline{\beta}_0\right)\xrightarrow{d}N\left(0,\operatorname{plim}\left[-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\frac{1}{G_j(\underline{X}_i,\underline{\beta}_0)}\frac{\partial G_j(\underline{X}_i,\underline{\beta})}{\partial\underline{\beta}^\top}\bigg|_{\underline{\beta}=\underline{\beta}_0}\frac{\partial G_j(\underline{X}_i,\underline{\beta})}{\partial\underline{\beta}}\bigg|_{\underline{\beta}=\underline{\beta}_0}-\frac{\partial^2 G_j(\underline{X}_i,\underline{\beta})}{\partial\underline{\beta}\,\partial\underline{\beta}^\top}\bigg|_{\underline{\beta}=\underline{\beta}_0}\right)\right]^{-1}\right)$$

Exercise: Fill in the details.

Next: MLE is convenient for iid observations. When this assumption is implausible (say independence without identical marginals), can adopt moment-based estimator of the form

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(Y_{ij}-G_j(\underline{X}_i,\hat{\underline{\beta}}_n)\right)\underline{w}_i=\underline{0},\qquad(*)$$

where $\underline{w}_i$ has the same dimension as $\underline{\beta}_0$, but does not depend on Y_{ij}.

Under further (but standard) conditions, $\hat{\underline{\beta}}_n$ is $\sqrt{n}$-consistent and asymptotically Normal with an asymptotic variance that depends on $\underline{W}_i$ (Exercise: Derive the asymptotic distribution of $\sqrt{n}\left(\hat{\underline{\beta}}_n-\underline{\beta}_0\right)$ under these conditions). GMM theory can help select the efficient setting of $\underline{W}_i$.

We've seen that different specifications of link function imply different models of multinomial response.

Can also get different multinomial response models for a given link function by allowing regressors or parameters to vary across alternatives.

1) Alternative-varying regressors

For individual i, suppose the regressors are

$$\underline{X}_i=[\underline{X}_{i1}^\top\quad\cdots\quad\underline{X}_{im}^\top]^\top$$

The link function for alternative j has a multi-index form given by

$$G_j(\underline{X}_i,\underline{\beta}_0)=G_j(\underline{X}_{i1}^\top\underline{\beta}_0,\ldots,\underline{X}_{im}^\top\underline{\beta}_0)$$

assuming $\underline{\beta}_0$ is invariant across alternatives.

2) Alternative-invariant regressors

This is typically coupled with alternative-varying parameters. Then the link function also has a multi-index form

$$G_j(\underline{X}_i,\underline{\beta}_0)=G_j(\underline{X}_i^\top\underline{\beta}_{01},\ldots,\underline{X}_i^\top\underline{\beta}_{0m})$$

where $\underline{\beta}_0=[\underline{\beta}_{01}^\top\quad\cdots\quad\underline{\beta}_{0m}^\top]^\top$

In this case, typically normalize

$$\underline{\beta}_{01} = \underline{0}$$

(Important: Software for multinomial response models requires one to specify alternative-varying regressors with alternative-invariant parameters or vice versa)
Recommendation: Use software for alternative varying regressors. Then if alternative-invariant regressors seem more plausible, then write

$$X_{ij} \equiv \left[\underbrace{\underline{0}^\top \quad \cdots \quad \underline{0}^\top}_{j-1\times d} \quad \underbrace{\underline{X}_i^\top}_{1\times d} \quad \underbrace{\underline{0}^\top \quad \cdots \quad \underline{0}^\top}_{n-j}\right]^\top$$

Where if $\underline{X}_I$ is a d-variate, $\underline{0}$ is a (d×1)-vector of zeroes.
In addition, need to set

$$\underline{\beta} \equiv \left[\underline{0}^\top \quad \underline{\beta}_2^\top \quad \cdots \quad \underline{\beta}_m^\top\right]^\top$$

where each $\underline{\beta}_j$ (j=1,…,m) is d×1, and $\underline{\beta}_1 = \underline{0}$ is a normalization.
The simples category of multinomial response model involves logistic link function (Luce, 1959).

Several different variants

a) Alternative-varying regressors, i.e. conditional logit (CL), we have

$$G_j(\underline{X}_i, \underline{\beta}_0) = \frac{e^{\underline{X}_{ij}^\top \underline{\beta}_0}}{1 + e^{\underline{X}_{ij}^\top \underline{\beta}_0}}, \qquad j = 1, \dots, m$$

b) Alternative-invariant regressors, i.e. multi-nomial logit (MNL)

$$G_j(\underline{X}_i, \underline{\beta}_0) = \frac{e^{\underline{X}_i^\top \underline{\beta}_{0j}}}{1 + e^{\underline{X}_i^\top \underline{\beta}_{0j}}}, \qquad \text{for } j \in \{1, \dots, m\} \text{ and where } \underline{\beta}_{0i} = \underline{0}$$

c) A combination of alternative-varying and alternative-invariant regressors, i.e. mixed logit: Let $\underline{X}_i \equiv [\underline{X}_{i1}^\top \quad \cdots \quad \underline{X}_{im}^\top]^\top$ be the vector of alternative-varying regressors; let $\underline{W}_i$ be the vector of alternative-invariant regressors. Let $\underline{\gamma}_0 \equiv \left[\underline{\gamma}_{01}^\top \quad \cdots \quad \underline{\gamma}_{0m}^\top\right]^\top$, then

$$P_{ij} = G_j(\underline{X}_i, \underline{W}_i, \underline{\beta}_0, \underline{\gamma}_0) = \frac{e^{\underline{X}_{ij}^\top \underline{\beta}_0 + \underline{w}_i^\top \underline{\gamma}_{0j}}}{\sum_{k=1}^m e^{\underline{X}_{ik}^\top \underline{\beta}_0 + \underline{w}_i^\top \underline{\gamma}_{0k}}}$$

Luce (1959): Conditional Logit (CL)

$$P[Y_i = j|\underline{X}_i] = G_j(\underline{X}_i; \underline{\beta}_0) \equiv \frac{e^{\underline{X}_{ij}^\top \underline{\beta}_0}}{\sum_{k=1}^m e^{\underline{X}_{ik}^\top \underline{\beta}_0}}, \qquad \text{individual } i, \qquad j \in \{1, \cdots, m\}$$

Multinomial Logit (MNL)

$$P[Y_i = j|\underline{X}_i] = G_j(\underline{X}_i; \underline{\beta}_0) \equiv \frac{e^{\underline{X}_i^\top \underline{\beta}_{0j}}}{\sum_{k=1}^m e^{\underline{X}_i^\top \underline{\beta}_{0k}}}, \qquad \text{individual } i, \qquad j \in \{1, \cdots, m\}$$

MNL requires normalization, usually $\beta_{01} \equiv 0$.

Maximum-likelihood Estimation of CL and MNL; for CL

$$\frac{\partial}{\partial \underline{\beta}^\top} G_j(\underline{X}_i, \underline{\beta}) = G_j(\underline{X}_i, \underline{\beta})(\underline{X}_{ij} - \overline{\underline{X}}_i), \qquad \text{where } \overline{\underline{X}}_i = \sum_{j=1}^m G_j(\underline{X}_i, \underline{\beta})\underline{X}_{ij}$$

i.e. conditional probability weighted average of regressors. This implies that the FOCs are

$$\sum_{i=1}^n \sum_{j=1}^m Y_{ij}(\underline{X}_{ij} - \overline{\underline{X}}_i) = \underline{0}, \qquad \text{(Exercise)}$$

In addition, the MLE of a CL satisfies

$$\sqrt{n}(\hat{\underline{\beta}}_{CL} - \underline{\beta}_0) \xrightarrow{d} N\left(\underline{0}, \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^m G_j(\underline{X}_i, \underline{\beta}_0)(\underline{X}_{ij} - \overline{\underline{X}}_i)(\underline{X}_{ij} - \overline{\underline{X}}_i)^\top\right)^{-1}\right)$$

(Exercise: Fill in details) For MNL, we have for each $j \in \{1,\ldots,m\}$

$$\frac{\partial}{\partial \underline{\beta}_k^\top} G_j(\underline{X}_i, \underline{\beta}) = G_j(\underline{X}_i, \underline{\beta})\left(\delta_{ijk} - G_k(\underline{X}_i, \underline{\beta})\right)\underline{X}_i$$

Where $\delta_{ijk} = 1\{j=k\}$. So the FOCs have the form

$$\sum_{i=1}^n \left(Y_{ij} - G_j(\underline{X}_i, \underline{\beta})\right)\underline{X}_i = \underline{0}, \qquad \text{for each } j \in \{1, \cdots, m\}$$

Also, the MNL MLE satisfies

$$\sqrt{n}\left(\hat{\underline{\beta}}_{MNL} - \underline{\beta}_0\right) \xrightarrow{d} N\left(\underline{0}, \operatorname{plim}\left(-\frac{1}{n}\frac{\partial^2 \hat{L}_n(\underline{\beta})}{\partial \underline{\beta}\, \partial \underline{\beta}^\top}\bigg|_{\underline{\beta}=\underline{\beta}_0}\right)^{-1}\right)$$

where the (j,k)-block of $\operatorname{plim}\left(-\frac{1}{n}\frac{\partial^2 \hat{L}_n(\underline{\beta})}{\partial \underline{\beta}\, \partial \underline{\beta}^\top}\Big|_{\underline{\beta}=\underline{\beta}_0}\right)$ is

$$\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n} G_j\left(\underline{X}_i, \underline{\beta}_0\right)\left(\delta_{ijk} - G_k\left(\underline{X}_i, \underline{\beta}_0\right)\right)\underline{X}_i\underline{X}_i^\top, \qquad \text{where } j,k \in \{1,\cdots,m\}$$

(Exercise: Fill in details)

Next: Consider the interpretation of CL and MNL model parameters

CL model: What is the effect on $G_j(\underline{X}_i,\underline{\beta}_0)$ of an infinitesimal change in $\underline{X}_{ik}$?

$$\frac{\partial G_j\left(\underline{X}_i, \underline{\beta}_0\right)}{\partial \underline{X}_{ik}^\top} = G_j\left(\underline{X}_i, \underline{\beta}_0\right)\left(\delta_{ijk} - G_k\left(\underline{X}_i, \underline{\beta}_0\right)\right)\underline{\beta}_0$$

So, if the e-th component of $\underline{\beta}_0$ is positive, an infinitesimal change in the e-th component of $\underline{X}_{ik}$ increases the probability of alternative k and decreases the probability of every other alternative.

MNL model: What is the effect on $G_j(\underline{X}_i,\underline{\beta}_0)$ of an infinitesimal change in $\underline{X}_i$?
The answer is summarized by

$$\frac{\partial G_j\left(\underline{X}_i, \underline{\beta}_0\right)}{\partial \underline{X}_i^\top} = G_j\left(\underline{X}_i, \underline{\beta}_0\right)\left(\underline{\beta}_{0j} - \bar{\underline{\beta}}_{0i}\right), \qquad \text{where } \bar{\underline{\beta}}_{0i} = \sum_{j=1}^{m} G_j\left(\underline{X}_i, \underline{\beta}_0\right)\underline{\beta}_{0j}$$

(Exercise: Verify this)
Note: The sign of this effect is ambiguous.

An analytically convenient feature of CL or MNL is the possibility of rewriting either model in terms of a logit.
Take MNL: MNL is equivalent to a series of pairwise comparisons of all but one alternative to a "baseline category," i.e. an alternative whose coefficient vector, say $\underline{\beta}_{01}$, is zero. To see this, note that

$$P[Y_i = j | Y_i \in \{j,k\}, \underline{X}_i] = \frac{G_j\left(\underline{X}_i, \underline{\beta}_0\right)}{G_j\left(\underline{X}_i, \underline{\beta}_0\right) + G_k\left(\underline{X}_i, \underline{\beta}_0\right)} = \frac{e^{\underline{X}_i^\top \underline{\beta}_{0j}}}{e^{\underline{X}_i^\top \underline{\beta}_{0j}} + e^{\underline{X}_i^\top \underline{\beta}_{0k}}} = \frac{e^{\underline{X}_i^\top (\underline{\beta}_{0j} - \underline{\beta}_{0k})}}{1 + e^{\underline{X}_i^\top (\underline{\beta}_{0j} - \underline{\beta}_{0k})}}$$

i.e. a logit model with coefficient vector $\underline{\beta}_{0j} - \underline{\beta}_{0k}$.

Usually, set up alternative 1 as the baseline, so $\underline{\beta}_{01} = \underline{0}$. In this case, for j∈{2,…,m}

$$P[Y_i = j | Y_i \in \{1, j\}, \underline{X}_i] = \frac{e^{\underline{X}_i^T \underline{\beta}_{0j}}}{1 + e^{\underline{X}_i^T \underline{\beta}_{0j}}}$$

i.e. a logit with coefficient vector $\underline{\beta}_{0j}$. Thus, the relative risk of picking alternative j over alternative 1 is

$$\frac{P[Y_i = j | Y_i \in \{1, j\}, \underline{X}_i]}{P[Y_i = 1 | Y_i \in \{1, j\}, \underline{X}_i]} = e^{\underline{X}_i^T \underline{\beta}_{0j}}$$

which implies that $e^{\beta_{0j\ell}}$ (where $\beta_{0j\ell}$ is the ℓ-th component of $\underline{\beta}_{0j}$) is the instantaneous change in the relative risk when the ℓ-th component of $X_{i\ell}$ of $\underline{X}_i$ increases by one unit.

It follows that MNL coefficients only have a natural interpretation if there exists a logical base alternative.

e.g. Let Y_i=j according to which of m modes of commuting individual I uses to get to work.

Depending on the audience/geography might make sense to see $Y_i \equiv 1$ if individual I chooses to drive his/her own car.

Next: Consider CL. Can show that

$$P[Y_i = j | Y_i \in \{j, k\}, \underline{X}_i] = \frac{e^{(\underline{X}_{ij} - \underline{X}_{ik})^T \underline{\beta}_0}}{1 + e^{(\underline{X}_{ij} - \underline{X}_{ik})^T \underline{\beta}_0}}$$

i.e. a logit with regressors $\underline{X}_{ij} - \underline{X}_{ik}$. Therefore, should normalize w.r.t. the regressor values (say $\underline{X}_{i1}$) for the base alternative.

Important: The reducibility of CL or MNL to a series of m−1 logits is a serious limitation because it is implausible that individual choice over m alternatives can be reduced to a series of m−1 (pairwise comparisons).

e.g. "Red-bus, blue-bus" problem: Suppose Y_i=1,2,3 as commuter I chooses to drive his own car to work (Y_i=1), ride the red bus (Y_i=2) or ride the blue bus (Y_i=3).

Assume that alternatives 2 and 3 do not differ a part from the color of the bus, which is assumed to be irrelevant.

MNL or CL imply that conditional on $\underline{X}_i$, the probability of observing Y_i=3 (red bus) is the same regardless of whether the choice set is {car, red bus} (m=2) or {car, red bus, blue bus} (m=3). This is unrealistic.

- Introducing on equivalent blue bus when a red bus is available should have the use of the red bus and have no effect on the popularity of driving

Essentially: CL and MNL both imply independence of irrelevant alternatives (IIA)

- The implausibility of IIA has led to an emphasis on multinomial response models that don't have this feature

e.g. nested logit, random parameters logit (RPL), multinomial probit (MNP)

- IIA is testable via the Hausman test: Hausman and McFadden (1984, EMA)

Consider the fundamental question of how to derive a multinomial response model for consumer choice: Easiest to start with an additive random utility (ARUM) model, where alternative $j \in \{1,\ldots,m\}$ has utility

$$u_j = v_j + \varepsilon_j$$

where v_j is deterministic and ε_j is random. The alternative chosen is the one with the highest u_j. Suppose that

$$u_{ij} = v_{ij} + \varepsilon_{ij}, \qquad \text{for individual } i, \qquad j \in \{1, \cdots, m\}$$

Then,

$$\begin{aligned} P[i \text{ chooses } j] &= P[Y_i = j] \\ &= P\left[u_{ij} + u_{ik}\right], \qquad \forall j \neq k \\ &= P\left[\varepsilon_{ik} - \varepsilon_{ij} \leq v_{ij} - v_{ik}\right], \qquad \forall j \neq k \\ &= F_{\varepsilon_{ik}-\varepsilon_{ij}}\left(v_{ij} - v_{ik}\right) \end{aligned}$$

where $F_{\varepsilon_{ik}-\varepsilon_{ij}}$ is the cdf of $\varepsilon_{ik}-\varepsilon_{ij}$.

- Different specifications of $F_{\varepsilon_{ik}-\varepsilon_{ij}}$ imply different models of multinomial response
- Suppose j=1 is the base category. Let $\bar{\varepsilon}_{ij} \equiv \varepsilon_{ij} - \varepsilon_{i1}$ and $\bar{v}_{ij} \equiv v_{ij} - v_{i1}$. Then

$$\begin{aligned} P[Y_i = j] &= P[\bar{\varepsilon}_{i2} \leq -\bar{v}_{i2}, \cdots, \bar{\varepsilon}_{im} \leq -\bar{v}_{im}] \\ &= \int_{-\infty}^{-\bar{v}_{i2}} \cdots \int_{-\infty}^{-\bar{v}_{im}} f_{\bar{\varepsilon}_{i2}\cdots\bar{\varepsilon}_{im}}(t_2, \cdots, t_m) dt_m \cdots dt_2 \end{aligned}$$

which generally has no closed form solution.

In addition, expect that $\varepsilon_{i1},\ldots,\varepsilon_{im}$ to be correlated in general across alternatives, particularly amongst complements or substitutes.

On the other hand, easiest to start with the case where $\varepsilon_{i1},\ldots,\varepsilon_{im}$ are iid. In particular, suppose $\varepsilon_{i1},\ldots,\varepsilon_{im}$ are iid with a common type I estreme value distribution, i.e. continuous with common density

$$f_{\varepsilon_{ij}}(t) = e^{-t}e^{-e^{-t}}, \qquad \text{for each } j$$

(In the case m=2, this implies a logit model) For general m, if $\varepsilon_{i1},\ldots,\varepsilon_{im}$ are iid type I extreme value then

$$P[Y_i = j | V_{i1}, \cdots, V_{im}] = \frac{v_{ij}}{\sum_{k=1}^{m} e^{v_{ik}}}$$

which is CL if $v_{ij} = \underline{X}_{ij}^{\top}\underline{\beta}_0$
or MNL if $v_{ij} = \underline{X}_{i}^{\top}\underline{\beta}_{0j}$

Suppose consumer i picks one of m alternatives. Let

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \qquad \text{for consumer i,} \qquad \text{alternative } j \in \{1, \dots, m\}$$

(i.e. an ARUM) Let Y_i be the corresponding multinomial response variable. Then

$$\begin{aligned} P[\text{i chooses j}] &= P[Y_i = j] \\ &= P\left[U_{ij} \geq U_{ik}, \qquad \forall k \neq j\right] \\ &= P\left[V_{ij} + \varepsilon_{ij} \geq V_{ik} + \varepsilon_{ik}, \qquad \forall k \neq j\right] \\ &= P\left[\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - V_{ik}, \qquad \forall k \neq j\right] \end{aligned}$$

Let $F_{\varepsilon_{ik}-\varepsilon_{ij}}$ denote the cdf of $\varepsilon_{ik}-\varepsilon_{ij}$. Different specifications of $F_{\varepsilon_{ik}-\varepsilon_{ij}}$ imply different models of multinomial choice. Let j=1 be the base category. Let

$$\begin{aligned} \bar{\varepsilon}_{ij} &\equiv \varepsilon_{ij} - \varepsilon_{i1}, \qquad j \in \{2, \dots, m\} \\ \bar{V}_{ij} &\equiv V_{ij} - V_{i1}, \qquad j \in \{2, \dots, m\} \end{aligned}$$

Then

$$P[Y_i = 1] = P[\bar{\varepsilon}_{i2} \leq -\bar{V}_{i2}, \dots, \bar{\varepsilon}_{im} \leq -\bar{V}_{im}] = \int_{-\infty}^{-\bar{V}_{i2}} \dots \int_{-\infty}^{-\bar{V}_{im}} f_{\bar{\varepsilon}_{i2},\dots,\bar{\varepsilon}_{im}}(t_2, \dots, t_m) dt_m \dots dt_2$$

Where $f_{\bar{\varepsilon}_{i2},\dots,\bar{\varepsilon}_{im}}(.)$ is the joint density of $\bar{\varepsilon}_{i2}, \dots, \bar{\varepsilon}_{im}$ (assuming absolute continuity of the corresponding joint distribution).

Note: $P[Y_i=j]$ has generally no closed form solution.

Easiest special case: Suppose $\varepsilon_{i1},\dots,\varepsilon_{im}$ are iid. In particular, suppose $\varepsilon_{i1},\dots,\varepsilon_{im}$ are iid type-I Extreme-value, with common density

$$f_{\varepsilon_{ij}}(t) = e^{-t}e^{-e^{-t}}, \qquad \text{for each j}$$

(for m=2, this is a logit model) For general m, we get

$$\begin{aligned} P[Y_i = j | V_{i1}, \dots, V_{im}] &= \frac{V_{ij}}{\sum_{k=1}^{m} e^{V_{ik}}} \\ \text{i.e. CL if } V_{ij} &= \underline{X}_{ij}^{\top}\underline{\beta}_0 \\ \text{i.e. MNL if } V_{ij} &= \underline{X}_{i}^{\top}\underline{\beta}_{0j} \end{aligned}$$

Naturally, assuming independence of $\varepsilon_{i1},\dots,\varepsilon_{im}$ is implausible if any two alternatives are similar/complementary/substitutes.

e.g. Suppose alternatives 2, 3 are similar to most consumers. Then if for a representative consumer i, ε_{i2} is large and negative, then U_{i2} will tend to be overpredicted. So U_{i3} will also tend to be overpredicted. i.e. ε_{i3} is also large and negative (this shows that ε_{i2}, ε_{i3} are positively correlated).

Note: The red-bus blue-bus problem exists because in many cases it is unrealistic to assume independence of $\varepsilon_{i1},\dots,\varepsilon_{im}$.

Next: It is therefore natural to assume that

$$\underline{\varepsilon}_i \equiv [\varepsilon_{i1} \quad \cdots \quad \varepsilon_{im}]^\top \sim N_m(\underline{0}, \underline{\underline{\Sigma}})$$

with $\underline{\underline{\Sigma}}$ an arbitrary positive definite matrix.

This implies a multinomial probit (MNP) model (we will return to this shortly).

Welfare Analysis with an ARUM

- The idea is to place dollar values on the effect of changing one or more components of V_{ij}
- Suppose for consumer i the deterministic utility component is

$$V_{ij} = V(M_i - P_j, \underline{X}_{ij})$$

where M_i is income, P_j is the price of alternative j and $\underline{X}_{ij}$ are other observables. Then

$$U_{ij} = V(M_i - P_j, \underline{X}_{ij}) + \varepsilon_{ij}$$

Suppose $\underline{X}_{ij}$ changes from $\underline{X}_{ij}^{(1)}$ to $\underline{X}_{ij}^{(2)}$. The compensating variation (CV) is the resulting change in income required to U_{ij} at its initial level. So the highest utility possible with income M_i and characteristic $\underline{X}_{ij}^{(1)}$ is the same as that possible with income $M_i - CV_i$ and characteristics $\underline{X}_{ij}^{(2)}$, i.e.

$$\max_{j\in\{1,\ldots,m\}} V\left(M_i - P_j, \underline{X}_{ij}^{(1)}\right) + \varepsilon_{ij} = \max_{j\in\{1,\ldots,m\}} V\left(M_i - CV_i - P_j, \underline{X}_{ij}^{(2)}\right) + \varepsilon_{ij}$$

e.g. Suppose m=2 and

$$U_{ij} = M_i + X_{ij} + \varepsilon_{ij}, \qquad j \in \{1,2\}$$

Then if X_{ij} goes from $X_{ij}^{(1)}$ to $X_{ij}^{(2)}$, we have

- If alternative 1 is chosen before and after, i.e. if $U_{i1}^{(2)} \equiv M_i - CV_i + X_{i1}^{(2)} + \varepsilon_{i1} = M_i + X_{i1}^{(1)} + \varepsilon_{i1}$, then

$$CV_i = X_{i1}^{(2)} - X_{i1}^{(1)}$$

- Similarly, if alternative 2 is chosen before and after, then $CV_i = X_{i2}^{(2)} - X_{i1}^{(1)}$

$$CV_i = X_{i2}^{(2)} - X_{i1}^{(1)}$$

- If alternative 1 is chosen before and alternative 2 after, then

$$CV_i = X_{i2}^{(2)} - X_{i1}^{(1)} + \varepsilon_{i2} - \varepsilon_{i1}, \qquad \text{i.e. } CV_i \text{ is random}$$

- If alternative 2 is chosen before and alternative 1 after, then

$$CV_i = X_{i1}^{(2)} - X_{i2}^{(1)} + \varepsilon_{i1} - \varepsilon_{i2}, \qquad \text{i.e. } CV_i \text{ is also random}$$

Of interest is $E[CV_i]$, which involves integrating ε_{i1}, ε_{i2}. Now, go back to the problem of replacing the assumption that $\varepsilon_{i1},\ldots,\varepsilon_{im}$ are independent. The easiest generalization involves nested logit. Start by assuming that $\underline{\varepsilon}_i \equiv [\varepsilon_{i1} \quad \cdots \quad \varepsilon_{im}]^\top$ has a generalized extreme value (GEV) distribution (McFadden, 1978) with joint pdf

$$F_{\underline{\varepsilon}}(\varepsilon_1, \ldots, \varepsilon_m) = e^{-G(e^{-\varepsilon_1},\ldots,e^{-\varepsilon_m})}$$

where the function $G(t_1,\ldots,t_m)$ is specified to make $F_{\underline{\varepsilon}}$ a proper cdf and the marginal distributions of $\varepsilon_1,\ldots,\varepsilon_m$ to be well defined.

Note: Assuming an ARUM $U_{ij}=V_{ij}+\varepsilon_{ij}$. We have

$$P[Y_i = j|\underline{X}_i] = e^{V_{ij}} \frac{G_j\left(e^{-V_{i1}}, \ldots, e^{-V_{im}}\right)}{G(e^{-V_{i1}}, \ldots, e^{-V_{im}})}$$

where

$$G_j(t_1, \ldots, t_m) = \frac{\partial G(t_1, \ldots, t_m)}{\partial t_j}$$

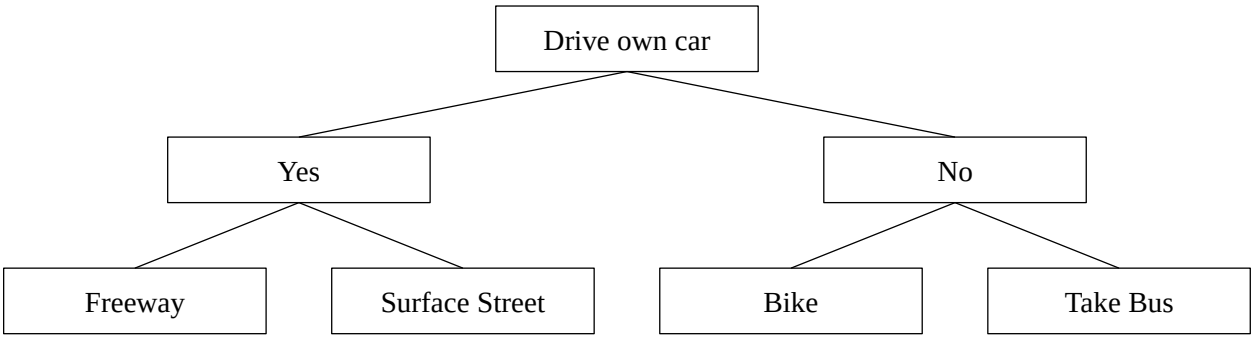
(Different specifications of $G(t_1,\ldots,t_m)$ imply different models of multinomial response, e.g.

$$G(t_1, \ldots, t_m) = \sum_{j=1}^{m} t_j$$

implies MNL)

The nested logit model involves a somewhat complicated specification of G(.). Nested logit assumes a multi-stage decision process.

e.g. Choice of mode for a morning commute



More generally, assume: The consumer has initially a mean of J choices (i.e. limbs). The j-th limb (j=1,…,J) has K_j branches labelled (j1,…,jk,…,jK_j). The utility of the alternative attained by picking limb j and branch k (i.e. address (j,k)) is

$$U_{jk} = V_{jk} + \varepsilon_{jk}, \qquad j \in \{1, \ldots, J\}, \qquad k \in \{1, \ldots, K_j\}, \qquad \text{where } \sum_{j=1}^{J} K_j = m$$

(this is a two-level decision tree; an arbitrary number of levels can be accommodated)

In a two-level model, we have

$$\begin{aligned}\text{P[individual i chooses (j,k)]} &= \text{P[individual i chooses limb j]} \\ &\quad \times \text{P[individual i chooses branch k|individual i chooses limb j]} \\ \Leftrightarrow P_{ijk} &= P_{ij} \times P_{ik|j}\end{aligned}$$

Nested logit involves specifying the random utility component $\underline{\varepsilon}$ in an ARUM as a generalized extreme value (GEV) random variable with cdf

$$F_{\underline{\varepsilon}}(\underline{\varepsilon}') = e^{-G\left(e^{-\varepsilon'_{i1}},\ldots,e^{-\varepsilon'_{iK_1}},\ldots,e^{-\varepsilon'_{J1}},\ldots,e^{-\varepsilon'_{JK_J}}\right)}$$

where

$$G\left(t_{11},\ldots,t_{1K_1},\ldots,t_{J1},\ldots,t_{JK_J}\right) = \sum_{j=1}^{J}\left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j}\right)^{\rho_j}, \qquad \cdots(*)$$

where the <u>scale parameters</u> are given by

$$\rho_j = \sqrt{1 - \text{Corr}\left(\varepsilon_{ijk}, \varepsilon_{ij\ell}\right)}, \qquad \text{for any } k, \ell \in \{1,\ldots,K_j\}$$

<u>Note</u>: $\rho_j \equiv 1$ for each $j \in \{1,\ldots,J\}$ iff the random utility components are independent (i.e. MNL)
Now: Let $Y_{ijk} \equiv 1\{\text{individual I chooses (j,k)}\}$. Then

$$P\left[Y_{ijk} = 1\right] = P\left[U_{ijk} \geq U_{i\ell m}, \qquad \forall(\ell, m)\right]$$

Also: Assume that the deterministic utility component for alternative (j,k) is

$$V_{ijk} = \underline{Z}_{ij}^{\top}\underline{\alpha}_0 + \underline{X}_{ijk}^{\top}\underline{\beta}_{0j}, \qquad j \in \{1,\ldots,J\}, \qquad k \in \{1,\ldots,K_j\}$$

where $\underline{Z}_{ij}$ varies over limbs but not branches. $\underline{X}_{ijk}$ varies over limbs and branches. Then

$$P_{ijk} = P_{ij} \times P_{ik|j} = \frac{e^{\underline{Z}_{ij}^{\top}\underline{\alpha}_0 + \rho_j I_{ij}}}{\sum_{m=1}^{J} e^{\underline{Z}_{im}^{\top}\underline{\alpha}_0 + \rho_m I_{im}}} \times \frac{e^{\frac{1}{\rho_j}\underline{X}_{ijk}^{\top}\underline{\beta}_j}}{\sum_{\ell=1}^{K_j} e^{\frac{1}{\rho_j}\underline{X}_{ij\ell}^{\top}\underline{\beta}_j}}$$

where

$$I_{ij} = \log\left(\sum_{\ell=1}^{K_j} e^{\frac{1}{\rho_j}\underline{X}_{ij\ell}^{\top}\underline{\beta}_j}\right)$$

is the <u>inclusive value</u> or <u>log sum</u>.
Warning: $\text{Corr}(\varepsilon_{ij1},\varepsilon_{ij2}) = \text{Corr}(\varepsilon_{ij1},\varepsilon_{ij3}) = \ldots = \text{Corr}(\varepsilon_{ijK_j-1}, \varepsilon_{ijK_j}) = \rho_j \; \forall j$ (branches share identical correlation)

Nested Logit

- Two level decision process
- J limbs, j-th limb has K_j branches

$$\text{P[individual i chooses limb j branch k]} = \text{P[i chooses limb j]P[i chooses branch k|i chooses limb j]}$$
$$\Leftrightarrow P_{ijk} = P_{ij}P_{ik|j}$$

Assume that the deterministic utility component for alternative (j,k) is

$$V_{ijk} = \underline{Z}_{ij}^{\top}\underline{\alpha}_0 + \underline{X}_{ijk}^{\top}\underline{\beta}_{0j}, \qquad j \in \{1, \dots, J\}, \qquad k \in \{1, \dots, K_j\}$$

The nested logit model specifies

$$P_{ijk} = P_{ij}P_{ik|j} = \frac{e^{\underline{Z}_{ij}^{\top}\underline{\alpha}_0 + \rho_{0j}I_{0ij}}}{\sum_{m=1}^{J} e^{\underline{Z}_{im}^{\top}\underline{\alpha}_0 + \rho_{0m}I_{0im}}} \times \frac{e^{\frac{1}{\rho_{0j}}\underline{X}_{ijk}^{\top}\underline{\beta}_{0j}}}{\sum_{\ell=1}^{K_j} e^{\frac{1}{\rho_{0j}}\underline{X}_{ij\ell}^{\top}\underline{\beta}_{0j}}}$$

where

$$\rho_{0j} = \sqrt{1 - \underbrace{\text{Corr}[\varepsilon_{ijk}, \varepsilon_{ij\ell}]}_{\text{assumed to be the same } \forall k \neq \ell}}, \qquad \text{for any } k, \ell \in \{1, \dots, K_j\}$$

$$\text{and } I_{0ij} = \log\left(\sum_{\ell=1}^{K_j} e^{\frac{1}{\rho_{0j}}\underline{X}_{ij\ell}^{\top}\underline{\beta}_j}\right)$$

Let

$$\underline{\rho}_0 \equiv [\rho_{01} \quad \cdots \quad \rho_{0J}]^{\top}$$

How to estimate $\underline{\alpha}_0$, $\underline{\beta}_0$ and $\underline{\rho}_0$?

For individual I, we observe $m=K_1+\dots+K_J$ outcomes Y_{ijk}, where $Y_{ijk}=1\{\text{individual i chooses (j,k)}\}$.

$$\text{Let } P_{ijk} \equiv P[Y_{ijk} = 1], \qquad Y_{ij} \equiv \sum_{k=1}^{K_j} Y_{ijk}, \qquad P_{ij} \equiv P[Y_{ij} = 1], \qquad P_{ik|j} \equiv P[Y_{ijk} = 1 | Y_{ij} = 1]$$

Then $P_{ijk}=P_{ik|j}P_{ij}$ and the density of

$$\underline{Y}_i \equiv [Y_{i11} \quad \cdots \quad Y_{i1K_2} \quad \cdots \quad Y_{iJ1} \quad \cdots \quad Y_{iJK_J}]$$

is given by

$$f_{\underline{Y}}(\underline{y}_i; \underline{\alpha}_0, \underline{\beta}_0, \underline{\rho}_0) = \prod_{j=1}^{J}\prod_{k=1}^{K_j}(P_{ik|j}P_{ij})^{y_{ijk}} = \prod_{j=2}^{J} p_{ij}^{y_{ij}} \prod_{k=1}^{K_j} P_{ik|j}^{y_{jk}}$$

Therefore the joint density of the sample $\underline{Y}_1,\ldots,\underline{Y}_n$ is

$$\prod_{i=1}^{n} f_{\underline{Y}}\left(\underline{y}_i; \underline{\alpha}_0, \underline{\beta}_0, \underline{\rho}_0\right)$$

and the full information maximum-likelihood (FIML) estimator maximizes

$$\frac{1}{n}\log \hat{L}_n\left(\underline{\alpha}, \underline{\beta}, \underline{\rho}\right) \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij} \log P_{ij} + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K_j} Y_{ijk} \log P_{ik|j}\,, \qquad \cdots (*)$$

with respect to $\underline{\alpha}$, $\underline{\beta}$, $\underline{\rho}$.

Important: In general, $\frac{1}{n}\log \hat{L}_n\left(\underline{\alpha}, \underline{\beta}, \underline{\rho}\right)$ is not globally concave → May be useful to use limited information maximum-likelihood (LIML) to generate starting values.

LIML in this context first estimates the parameters in the second term of the RHS of (*), which is the log-likelihood of a CL model with parameter $\frac{1}{\rho_j}\underline{\beta}_j$ for each limb. LIML then estimates the parameters in the first term of the RHS of (*), which is the log-likelihood of another CL model with added regressors $\hat{I}_{ij}$ which embed the estimates from the first-stage estimation of the second term of (*).

Some discussion regarding the nested logit model

- (Obvious) Not all choice problems have a natural nesting structure. In practice, may need to treat the nesting structure as another parameter to estimate, perhaps via AIC/BIC
- Can occasionally get ML estimates of the scale parameters ρ_j that lie outside the range [0,1], which implies a violation of the ARUM specification
- Often see decision trees with more than two levels

e.g. Consumer behavior in a car dealership

1. Buy a car (Y/N)
2. If Y, buy a new car (Y/N)
3. If Y, which class of new car to buy (e.g. sedan, minivan, SUV, etc.)
4. Given the choice at level 3, foreign/domestic
5. Given the choice at level 4, which model to buy

Next: Consider random parameters logit (RPL)

- Another generalization of CL or MNL in which the random utility components can exhibit correlation across alternatives
- For individual i, suppose the utility of alternative j is

$$U_{ij} = \underline{X}_{ij}^{\top}\underline{\beta}_i + \varepsilon_{ij}, \qquad j \in \{1, \ldots, m\}$$

where $\varepsilon_{i1},\ldots,\varepsilon_{im}$ are iid type-I extreme value (as in CL), but where $\underline{\beta}_i \sim N\left(\underline{\beta}_0, \underline{\underline{\Sigma}}_{\underline{\beta}_0}\right)$, i.e. random coefficients. Assume that $\underline{\beta}_i$ is independent of ε_{ij} for all j.

Therefore

$$U_{ij} = \underline{X}_{ij}^{\top}\underline{\beta}_0 + \underline{X}_{ij}^{\top}\underline{u}_i + \varepsilon_{ij}, \qquad \cdots (\dagger)$$
$$\text{where } \underline{u}_i \sim N\left(\underline{0}, \underline{\underline{\Sigma}}_{\underline{\beta}_0}\right)$$

Let

$$\zeta_{ij} \equiv \underline{X}_{ij}^{\top}\underline{u}_i + \varepsilon_{ij}, \qquad \text{then } \mathrm{Cov}[\zeta_{ij}, \zeta_{ik}] = \underline{X}_{ij}^{\top}\underline{\underline{\Sigma}}_{\underline{\beta}_0}\underline{X}_{ij} \neq 0, \qquad \text{for } j \neq k \text{ in general}$$

In most RPL applications, $\underline{\underline{\Sigma}}_{\underline{\beta}_0}$ is diagonal with some diagonal elements equal to zero, i.e. $\underline{\beta}_i$ is singular Normal (often see (†) aggregated over I to model market demand)

How to estimate RPL models?

- Cannot apply CL to estimate $\underline{\beta}_0$ in (†); inconsistent
- However, can integrate out $\underline{\beta}_i$ to get

$$P_{ij} \equiv P[Y_i = j] = \int \frac{e^{\underline{X}_{ij}^{\top}\underline{\beta}_i}}{\sum_{k=1}^{m} e^{\underline{X}_{ik}^{\top}\underline{\beta}_i}} \varphi\left(\underline{\beta}_i; \underline{\beta}_0, \underline{\underline{\Sigma}}_{\underline{\beta}_0}\right) d\underline{\beta}_i$$

a multivariate integral, where $\varphi\left(\underline{\beta}_i; \underline{\beta}_0, \underline{\underline{\Sigma}}_{\underline{\beta}_0}\right)$ is the $N\left(\underline{\beta}_0, \underline{\underline{\Sigma}}_{\underline{\beta}_0}\right)$-density.

The MLE of $\underline{\beta}_0$, $\underline{\underline{\Sigma}}_{\underline{\beta}_0}$ maximizes

$$\frac{1}{n}\log \hat{L}_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}\log P_{ij}, \qquad \text{wrt } \underline{\beta}, \underline{\underline{\Sigma}}_{\underline{\beta}}$$

Major problem: P_{ij} has no closed form, so P_{ij} needs to be approximated, usually by simulation.
A popular method of approximating P_{ij}: Take the average of S evaluations of the integrand at each S random draws of $\underline{\beta}_i$ from a $N\left(\underline{\beta}_0, \underline{\underline{\Sigma}}_{\underline{\beta}_0}\right)$-distribution ($\underline{\beta}_0, \underline{\underline{\Sigma}}_{\underline{\beta}_0}$ are unknown, so the simulation is part of an iterative procedure tha produces $\underline{\beta}_0^{(r)}$ and $\underline{\underline{\Sigma}}_{\underline{\beta}_0}^{(r)}$ at the iteration)

The maximum simulated likelihood (MSL) estimator maximizes

$$\frac{1}{n}\log \hat{L}_n\left(\underline{\beta}, \underline{\underline{\Sigma}}_{\underline{\beta}}\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}\log\left(\frac{1}{S}\sum_{s=1}^{S}\frac{e^{\underline{X}_{ij}^{\top}\underline{\beta}_i^{(s)}}}{\sum_{k=1}^{m} e^{\underline{X}_{ik}^{\top}\underline{\beta}_i^{(s)}}}\right)$$

where $\underline{\beta}_i^{(s)}$ (s=1,…,S) are iid $N\left(\underline{\beta}_0^{(r)}, \underline{\underline{\Sigma}}_{\underline{\beta}_0}^{(r)}\right)$.

Consistency requires $n\to\infty$, $S\to\infty$ with $\frac{\sqrt{n}}{S} \to \infty$: Gouriéroux and Monfort (1991, Annales d'Économie et de Statistique).

Christian Gouriéxroux and Alain Monfort (1991), Simulation based inference in models with heterogeneity, *Annales d'Économie et de Statistique* (20/21), 69–107

Extension: Can impose random parameters on nested logit models.

Consider a generalized random utility model, where individual i and alternative j

$$U_{ij} = V_{ij}(\underline{X}_i, \underline{\beta}_0) + \varepsilon_{ij}$$

where $\underline{X}_i$ is a vector of observed regressors, $\underline{\beta}_0$ is an unknown vector of constants and $\{\varepsilon_{ij}\}$ are independent over i but (in general) correlated over j.

Assume that the distribution of ε_{ij} is such that a closed form for the choice probabilities exists, with

$$P_{ij} \equiv P[\text{individual i chooses alternative j}] = G_j(\underline{V}_i(\underline{X}_i, \underline{\beta}_0), \underline{\theta}_{\underline{\varepsilon}})$$

where $\underline{V}_i(\underline{X}_i, \underline{\beta}_0) = [V_{i1}(\underline{X}_i, \underline{\beta}_0) \quad \cdots \quad V_{im}(\underline{X}_i, \underline{\beta}_0)]^\top$ and where $\underline{\theta}_{\underline{\varepsilon}}$ is a vector of unknown parameters of the distribution of

$$\underline{\varepsilon}_i \equiv [\varepsilon_{i1} \quad \cdots \quad \varepsilon_{im}]^\top$$

(e.g. $\underline{\varepsilon}_i$ could have a GEV distribution)

Can introduce additional randomness to this model by specifying

$$V_{ij} = V_{ij}(\underline{X}_i, \underline{\xi}_i, \underline{\beta}_0)$$

where $\underline{\xi}_i$ is a random vector.

In this case,

$$P_{ij} = \int G_j(\underline{V}_i(\underline{X}_i, \underline{\xi}_i, \underline{\beta}_0), \underline{\theta}_{\underline{\varepsilon}}) f\left(\underline{\xi}_i; \underline{\theta}_{\underline{\xi}}\right) d\underline{\xi}_i$$

where $f\left(\underline{\xi}; \underline{\theta}_{\underline{\xi}}\right)$ is the density of $\underline{\xi}$.

Can motivate the introduction of additional randomness in this case by assuming that individuals belong to one of C latent classes; suppose that $\underline{\beta}$, $\underline{\theta}_{\underline{\varepsilon}}$ vary by class. Then

$$P_{ij} = \sum_{c=1}^{C} G_j(\underline{V}_i(\underline{X}_i, \underline{\beta}^c), \underline{\theta}_{\underline{\varepsilon}}^c)\pi_c$$

where $\pi_c \in [0,1]$, and $\sum_c \pi_c = 1$: Walker and Ben-Akiva (2002, Mathematical Social Sciences)

Joan Walker and Moshe Ben-Akiva (2002), Generalized random utility model, *Mathematical Social Sciences* 43 (3), 303–343

Generalized random utility models

Suppose that individuals belong to one of C latent classes. If the utility parameters vary by class

$$P_{ij} = \sum_{c=1}^{C} \int G_j\left(\underline{V}_i\left(\underline{X}_i, \underline{\xi}_i, \underline{\beta}^c\right), \underline{\theta}_{\underline{\xi}}^c\right) f\left(\underline{\xi}_i; \underline{\theta}_{\underline{\xi}}\right) d\underline{\xi}_i \, \pi_c$$

where $\pi_c \in [0,1]$ with $\sum_c \pi_c = 1$ (usually assume $c \in \{2,3\}$).

The corresponding MSL estimator maximizes

$$\frac{1}{n} \log \hat{L}_n\left(\underline{\beta}, \underline{\underline{\Sigma}}_{\underline{\beta}}\right) \equiv \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \log\left(\frac{1}{S} \sum_{s=1}^{S} \sum_{c=1}^{C} G_j\left(\underline{V}_i\left(\underline{X}_i, \underline{\xi}_i^s, \underline{\beta}^c\right), \underline{\theta}_{\underline{\xi}}^c\right) \pi_c\right)$$

where $\underline{\xi}_i^s$ ($s \in \{1,\ldots,S\}$) are iid with density $f\left(. \middle| \underline{\theta}_{\underline{\xi}}\right)$ (Walker and Ben-Akiva, 2002, *Mathematical Social Sciences*)

Lastly: Consider an obvious way of modelling correlation amongst the elements of

$$\underline{\varepsilon}_i = [\varepsilon_{i1} \quad \cdots \quad \varepsilon_{im}]^\top$$

The multinomial probit (MNP) model assumes for individual I and alternative j that

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \qquad \text{where } \underline{\varepsilon}_i = [\varepsilon_{i1} \quad \cdots \quad \varepsilon_{im}]^\top \sim N_m(\underline{0}, \underline{\underline{\Sigma}})$$

where usually $V_{ij} = \underline{X}_{ij}^\top \underline{\beta}$ or $V_{ij} = \underline{X}_i^\top \underline{\beta}_j$. Need to impose restrictions on $\underline{\underline{\Sigma}}$ to guarantee identification.

Recall

$$P_{ij} \equiv P[Y_i = j] = P\left[\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - V_{ik}, \qquad \forall k \neq j\right]$$

Normalize the utility differences w.r.t. the utility of alternative 1; we can then impose $\varepsilon_{i1} \equiv 0$ $\forall i$ and restrict the (1,1)-element of $\underline{\underline{\Sigma}}$ to zero as well as several other elements of $\underline{\underline{\Sigma}}$.

e.g. If m=2, and we impose $\varepsilon_{i1} \equiv 0$, then $\sigma_{11} = \sigma_{12} = \sigma_{21} = 0$. If also impose $\sigma_{22} = 1$, then

$$\varepsilon_{i2} - \varepsilon_{i1} = \varepsilon_{i2} \sim N(0,1)$$

which implies a probit model of binary response. One often in the context of MNP models sees factor models of the form

$$\varepsilon_{ij} = \zeta_{ij} + \sum_{\ell=1}^{L} c_{ij\ell} \xi_{i\ell}, \qquad j \in \{1, \ldots, m\}$$

where ζ_{ij}, ξ_{i1}, …, ξ_{iL} are iid N(0,1) and $c_{ij\ell}$ are factor loadings that can be estimated.

In this case the number of distinct parameters in $\underline{\underline{\Sigma}}$ to estimate goes from $\frac{m(m+1)}{2}$ to L, which is invariably smaller.

How to estimate MNP models?

No closed form for the choice probabilities.

e.g. If m=3 with $\varepsilon_{i1} \equiv 0$, then

$$P_{ij} = P[Y_i = j] \int_{-\infty}^{-\bar{V}_{i31}} \int_{-\infty}^{-\bar{V}_{i21}} f(\bar{\varepsilon}_{i21}, \bar{\varepsilon}_{i31}) d\bar{\varepsilon}_{i21} d\bar{\varepsilon}_{i31}$$

where

$$\bar{\varepsilon}_{i21} \equiv \varepsilon_{i2} - \varepsilon_{i1} = \varepsilon_{i2}, \qquad \bar{\varepsilon}_{i31} \equiv \varepsilon_{i3} - \varepsilon_{i1} = \varepsilon_{i3}$$

and where $f(\bar{\varepsilon}_{i21}, \bar{\varepsilon}_{i31})$ is a (singular) bivariate Normal density with up to two free parameters in the covariance matrix.

In general, a scenario with m choices requires the evaluation of an (m-1)-fold integral, which is usually impossible when m>4.

When m is large, usually use simulation to approximate P_{ij} and estimate $\underline{\beta}_0, \underline{\underline{\Sigma}}_0$ by MSL, i.e. by maximizing

$$\frac{1}{n} \log \hat{L}_n(\underline{\beta}, \underline{\underline{\Sigma}}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} \log \hat{P}_{ij}$$

where $\hat{P}_{ij}$ is obtained by simulation.

A logical alternative to MSL (especially in the context of MNP) is to use the method of simulated moments (MSM), i.e. to solve

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (Y_{ij} - \hat{P}_{ij}) \underline{Z}_i = \underline{0}$$

where $\hat{P}_{ij}$ is obtained by simulation such that

$$(Y_{ij} - \hat{P}_{ij}) \underline{Z}_i \text{ is unbiased for } (Y_{ij} - P_{ij}) \underline{Z}_i$$

Some discussion

- MNP generally involves a more challenging approximation to P_{ij} (via simulation) than is the case with MSL applied to RPL
- This likely explains why nested logit and RPL tend to be more popular than MNP

Example: Choice of fishing mode (based on a sample of n=1,182 recreational fishers; Thompson and Crooke (1991))

For individual $i \in \{1,\ldots,1,182\}$, observe $Y_i \in \{1,2,3,4\}$ according to whether i chooses to fish at a beach (j=1), off a pier (j=2), off a private boat (j=3) or off a charter boat (j=4). Expect that choice of mode depends on relative prices of each mode and the catch rate for each mode.

Cynthia J. Thompson and Stephen J. Crooke (1991), Results of the southern California sportfish economic survey, Southwest Fisheries Center

	In-sample averages			
Variable	Y=1 (beach)	Y=2 (pier)	Y=3 (private)	Y=4 (charter)
P_{beach} ($)	**36**	**31**	138	121
P_{pier} ($)	**36**	**31**	138	121
$P_{private}$ ($)	98	82	**42**	**45**
$P_{charter}$ ($)	125	110	71	75
CR_{beach} (%)	**28**	**26**	**21**	**25**
CR_{pier} (%)	22	20	13	16
$CR_{private}$ (%)	16	15	18	18
$CR_{charter}$ (%)	**52**	**50**	**65**	**69**
Sample proportions (%)	11.3	15.1	35.4	38.2
Observations	134	178	418	452

So

- Fishers tend to pick the mode where it is cheapest for them to fish
- Beach and pier fishers find it much cheaper to fish from shore than from a boat
- Relationship between catch rates and mode choice is not obvious; charter boats consistently have the highest catch rate

Clearly have alternative-varying regressors, so a conditional logit (CL) is natural

$$P_{ij} = P[Y_i = j] = \frac{e^{\beta_p P_{ij} + \beta_c CR_{ij}}}{\sum_{k=1}^{4} e^{\beta_p P_{ik} + \beta_c CR_{ik}}}$$

We find $\hat{\beta}_p = -.021$ and $\hat{\beta}_c = .953$. So decrease in the price of one alternative increases the probability of choosing that alternative and decreases the probability of every other alternative.

Similar story if catch rates change. Customary to report estimated average marginal effects, i.e. estimates of

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial P_{ij}}{\partial X_{ikr}}$$

i.e. the average marginal response of the probability of choosing j when the r-th regressor for alternative k increases by a tiny amount, all else equal.

Given the CL specification, the average response is estimated as

$$\frac{1}{n}\sum_{i=1}^{n} \hat{P}_{ij}\left(\delta_{ijk} - \hat{P}_{ik}\right)\hat{\beta}_r = \bar{P}_j\left(\delta_{jk} - \bar{P}_k\right)\hat{\beta}_r$$

(where $\hat{P}_{ij}$ are predicted choice probabilities, $\bar{P}_j$ are in-sample proportions of each mode) Therefore: Can report the average effect of a $100 increase in the price of each mode, all else constant.

Average marginal response

Variable	Y=1 (beach)	Y=2 (pier)	Y=3 (private)	Y=4 (charter)
ΔP[beach] (%)	−27.2	11.9	8.5	6.8
ΔP[pier] (%)	11.9	−26.3	8.0	6.4
ΔP[private] (%)	8.0	8.0	−39.1	22.5
ΔP[charter] (%)	6.8	6.4	22.5	−35.7

We see that shore-based and boat-based modes have a high degree of substitutability. Now suppose that individual choice of fishing mode also depends on monthly income (in thousands of dollars).

Average incomes across the four modes

Variable	Y=1 (beach)	Y=2 (pier)	Y=3 (private)	Y=4 (charter)	(overall)
Income ($1K/month)	4.052	3.387	4.654	3.881	4.099

Income is obviously invariant across alternatives, so a multinomial logit (MNL) model can be specified

$$P_{ij} = P[Y_i = j] = \frac{e^{\alpha_j + \beta_{Ij} I_i}}{\sum_{k=1}^{4} e^{\alpha_k + \beta_{Ik} I_i}}, \qquad j \in \{1, \ldots, 4\}$$

Can normalize $\alpha_1 = \beta_{I_1} = 0$ to ensure that probabilities sum to one (this imposes beach fishing as the baseline category).

Example: Choice of fishing mode. Sample of n=1,182 recreational fishers (Thompson and Crooke, 1991).

For fisher i we have $Y_i \in \{1,\ldots,4\}$ according to which of four modes is chosen. Last week we looked at the effect of relative prices on choice of mode.

Now: Suppose that mode choice is also depends on monthly income (in thousands of dollars).

		Subsample averages:			
	Y=1	Y=2	Y=3	Y=4	
Monthly income	(beach)	(pier)	(private)	(charter)	(overall)
(thousands of dollars)	4.052	3.387	4.654	3.881	4.099

Income doesn't vary across alternatives, so consider an MNL specification.

$$P_{ij} = \frac{e^{\alpha_j + \beta_{Ij} I_i}}{\sum_{k=1}^{4} e^{\alpha_k + \beta_{Ik} I_i}}, \qquad j \in \{1, \ldots, 4\}$$

Normalize $\alpha_1=0$, $\beta_{I1}=0$. So:

$\hat{\alpha}_1$(beach) = 0.000 $\hat{\beta}_{I1}$(beach) = 0.000

$\hat{\alpha}_2$(pier) = 0.814 $\hat{\beta}_{I2}$(pier) = −0.143

$\hat{\alpha}_3$(private) = 0.739 $\hat{\beta}_{I3}$(private) = 0.092

$\hat{\alpha}_4$(charter) = 1.341 $\hat{\beta}_{I4}$(charter) = −0.032

The average marginal effect of an income change is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial P_{ij}}{\partial I_i}, \qquad \text{which is estimated as } \frac{1}{n}\sum_{i=1}^{n} \hat{P}_{ij}\left(\hat{\beta}_{Ij} - \hat{\bar{\beta}}_i\right), \qquad \text{where } \hat{\bar{\beta}}_i = \sum_{j=1}^{4} \hat{P}_{ij}\hat{\beta}_{Ij}$$

(Exercise: show this) Therefore the average effect of a \$1,000 increase in monthly income is

ΔP[beach]=0.000 ΔP[pier]=−0.021 ΔP[private]=0.033 ΔP[charter]=−0.012

So: A \$1,000 increase in monthly income tends on average to cause fishers to move out of pier and charter boat fishing into private boat fishing.

Goodness of fit:

The CL model has a pseudo-R^2 of 0.162

The MNL model has a pseudo-R^2 of 0.099

Can do better than CL or MNL by adopting mixed logit (i.e. use prices and monthly income as regressors)

$$P_{ij} = \frac{e^{\beta_P P_{ij} + \beta_C C_{ij} + \alpha_j + \beta_{Ij} I_i}}{\sum_{k=1}^{4} e^{\beta_P P_{ik} + \beta_C C_{ik} + \alpha_k + \beta_{Ik} I_i}}$$

which is equivalent to the CL model

$$P_{ij} = \frac{e^{\beta_P P_{ij} + \beta_C C_{ij} + \sum_{\ell=1}^{4} \alpha_\ell d_{ij\ell} + \beta_{I\ell} dI_{ij\ell}}}{\sum_{k=1}^{4} e^{\beta_P P_{ik} + \beta_C C_{ik} + \sum_{\ell=1}^{4} \alpha_\ell d_{ik\ell} + \beta_{I\ell} dI_{ik\ell}}}$$

where $d_{ij\ell} = 1\{j = \ell\}$ and $dI_{ij\ell} = d_{ij\ell} I_i = I_i 1\{j = \ell\}$ (exercise: verify this).

We find that $\hat{\beta}_P = -0.025$, $\hat{\beta}_C = 0.358$.

$\hat{\alpha}_1$(beach) = 0.000	$\hat{\beta}_{I1}$(beach) = 0.000
$\hat{\alpha}_2$(pier) = 0.778	$\hat{\beta}_{I2}$(pier) = −0.128
$\hat{\alpha}_3$(private) = 0.527	$\hat{\beta}_{I3}$(private) = 0.089
$\hat{\alpha}_4$(charter) = 1.694	$\hat{\beta}_{I4}$(charter) = −0.033

- Compared to the CL fit, $\hat{\beta}_C$ in the mixed logit is much smaller
- While the others are not much different from their counterparts in the CL or MNL models
- However: The pseudo-R^2 is 0.258, so much higher than either the CL pseudo-R^2 or the MNL pseudo-R^2

Models with censored or truncated dependent variables

Let Y^* denote a latent dependent variable, i.e. Y^* is not completely observed; only observe a function Y of Y^*. For example,

Censoring If censoring is from below we observe

$$Y = \max\{Y^*, L\} = \begin{cases} Y^*, & Y^* > L \\ U, & Y^* \leq L \end{cases}$$

If censoring is from above we observe

$$Y = \min\{Y^*, U\} = \begin{cases} Y^*, & Y^* < U \\ U, & Y^* \geq U \end{cases}$$

e.g. Only observed offered wages Y^* if Y^* exceeds a reservation wage L (censoring from below)
e.g. Incomes above U dollars per annum are reported as U, i.e. incomes above U are top coded (censoring from above)

Truncation If truncation is from below we observe $Y=Y^*$ iff $Y^*>L$; nothing is reported if $Y^*\leq L$; i.e. there is information loss. Similarly, if truncation is from above we observe $Y=Y^*$ if $Y^*<U$; nothing is reported when $Y^*\geq U$.

MLE: Censored MLE: Suppose censoring is from below: If Y>L, then the conditional density of Y given $\underline{X}$ is the same as that of Y^*, i.e.

$$f_Y(y|\underline{X}) = f_{Y^*}(y|\underline{X})$$

If Y=L then

$$P[Y = L|\underline{X}] = P[Y^* \leq L|\underline{X}] = F_Y(L|\underline{X})$$

(Therefore the conditional distribution of Y given $\underline{X}$ is neither absolutely continuous nor discrete)

We have

$$f_Y(y|\underline{X}) = f_{Y^*}(y|\underline{X})^d F_{Y^*}(L|\underline{X})^{1-d}$$

where d=1{y>L}.

Now: Suppose Y^* has a conditional distribution given $\underline{X}$ parameterized with $\underline{\theta} \in \Theta$. Then the log-likelihood for a sample $[\underline{X}_i^\top \quad Y_i \quad D_i]$ (i=1,…,n) is

$$\hat{L}_n(\underline{\theta}) \equiv \sum_{i=1}^{n} [D_i \log f_{Y^*}(y_i|\underline{X}_i;\underline{\theta}) + (1 - D_i) \log F_{Y^*}(L|\underline{X}_i;\underline{\theta})], \qquad \cdots (*)$$

So: We see that ignoring censoring (i.e. using only observations for which Y_i>L) is equivalent to ignoring the second term of (*). This leads to an inconsistent estimate of $\underline{\theta}$.

Truncated MLE: Suppose truncation is from below. Then the conditional density off the observed Y given $\underline{X}$ is

$$\begin{aligned} f_Y(y|\underline{X}) &= f_{Y^*}(y|Y^* > L, \underline{X}) \\ &= \frac{f_{Y^*}(y|\underline{X})}{P[Y^* > L|\underline{X}]} \\ &= \frac{f_{Y^*}(y|\underline{X})}{1 - F_{Y^*}(L|\underline{X})} \end{aligned}$$

Suppose Y^* has a conditional distribution given $\underline{X}$ parameterized by $\underline{\theta} \in \Theta$. Then the log-likelihood for a sample $[\underline{X}_i^\top \quad Y_i]$ (i=1,…,n) is

$$\hat{L}_n(\theta) = \sum_{i=1}^{n} \left(\log f_{Y^*}(Y_i|\underline{X}_i;\underline{\theta}) - \log\left(1 - F_{Y^*}(L|\underline{X}_i;\underline{\theta})\right)\right), \qquad \cdots (\dagger)$$

So ignoring truncation is equivalent to ignoring the second term of (†), which leads to an inconsistent estimate of $\underline{\theta}$.

The most famous example of a regression with a censored dependent variable is the Tobit model (Tobin, 1958).

James Tobin (1958), Estimation of relationships for limited dependent variables, *Econometrica* 26 (1), 24–36

Assume that

$$Y^* = \underline{X}^\top \underline{\beta}_0 + \varepsilon, \qquad \text{where } \varepsilon|\underline{X} \sim N(0, \sigma_0^2)$$

Therefore

$$Y^*|\underline{X} \sim N\left(\underline{X}^\top \underline{\beta}_0, \sigma_0^2\right)$$

Suppose we observe

$$Y = \begin{cases} Y^*, & Y^* > 0 \\ 0, & Y^* \leq 0 \end{cases}$$

We have

$$\begin{aligned} F_{Y^*}(0|\underline{X}) = P[Y^* \leq 0|\underline{X}] &= P\left[\underline{X}^\top \underline{\beta}_0 + \varepsilon \leq 0 \middle| \underline{X}\right] \\ &= \Phi\left(-\frac{\underline{X}^\top \underline{\beta}_0}{\sigma_0}\right) \\ &= 1 - \Phi\left(\frac{\underline{X}^\top \underline{\beta}_0}{\sigma_0}\right) \end{aligned}$$

So the conditional density of Y given $\underline{X}$ is

$$f_Y(y|\underline{X}) = \left[\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}\left(y - \underline{X}^\top \underline{\beta}_0\right)^2}\right]^d \left[1 - \Phi\left(\frac{\underline{X}^\top \underline{\beta}_0}{\sigma_0}\right)\right]^{1-d}, \qquad \text{where } d = 1\{y > 0\}$$

The Tobit MLE $\hat{\underline{\theta}}_n \equiv \left[\hat{\underline{\beta}}_n^\top \quad \hat{\sigma}_n^2\right]^\top$ maximizes

$$\hat{L}_n(\underline{\theta}) = \sum_{i=1}^n \left\{ D_i \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\left(Y_i - \underline{X}_i^\top \underline{\beta}\right)^2 \right] + (1 - D_i)\log\left(1 - \Phi\left(\frac{\underline{X}_i^\top \underline{\beta}}{\sigma}\right)\right) \right\}$$

which has partial derivatives (verify)

$$\frac{\partial \hat{L}_n(\underline{\theta})}{\partial \underline{\beta}^\top} = \sum_{i=1}^n \left\{ \frac{1}{\sigma^2} \left[D_i \frac{Y_i - \underline{X}_i^\top \underline{\beta}}{\sigma^2} - (1 - D_i) \frac{\varphi\left(\frac{\underline{X}_i^\top \underline{\beta}}{\sigma}\right)}{1 - \Phi\left(\frac{\underline{X}_i^\top \underline{\beta}}{\sigma}\right)} \right] \underline{X}_i \right\}$$

$$\frac{\partial \hat{L}_n(\underline{\theta})}{\partial \sigma^2} = \sum_{i=1}^n \left\{ D_i \left[-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\left(Y_i - \underline{X}_i^\top \underline{\beta}\right)^2 \right] + (1 - D_i) \frac{\varphi\left(\frac{\underline{X}_i^\top \underline{\beta}}{\sigma}\right)}{1 - \Phi\left(\frac{\underline{X}_i^\top \underline{\beta}}{\sigma}\right)} \frac{\underline{X}_i^\top \underline{\beta}}{2\sigma^3} \right\}$$

Exercise: Derive the limiting distribution as $n \to \infty$ of the Tobit MLE.

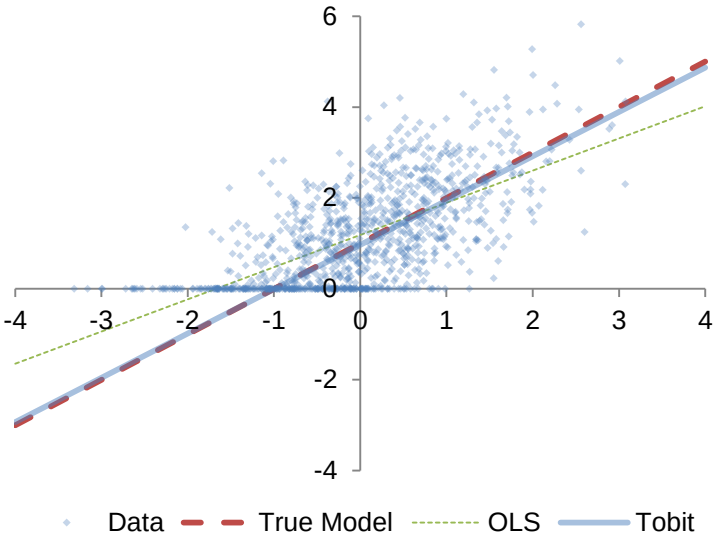
Fragility of the Tobit MLE: If the latent error ε is heteroskedastic or non-Normal then the Tobit MLE is inconsistent. To see this, observe that if the score identity holds, i.e. if

$$\mathrm{E}\left[\left.\frac{\partial \hat{\mathrm{L}}_n(\underline{\theta})}{\partial \underline{\beta}^{\mathrm{T}}}\right|_{\underline{\beta}=\underline{\beta}_0}\right] = \underline{0}$$

then $\mathrm{E}[\mathrm{D}_i] = \Phi\left(\frac{\underline{X}_i^{\mathrm{T}}\underline{\beta}_0}{\sigma_0}\right)$ and $\mathrm{E}[\mathrm{D}_i\mathrm{Y}_i] = \Phi\left(\frac{\underline{X}_i^{\mathrm{T}}\underline{\beta}_0}{\sigma_0}\right)\underline{X}_i^{\mathrm{T}}\underline{\beta}_0 + \sigma_0\varphi\left(\frac{\underline{X}_i^{\mathrm{T}}\underline{\beta}_0}{\sigma_0}\right)$. Under heteroskedasticity

$$\mathrm{E}[\mathrm{D}_i] = \Phi\left(\frac{\underline{X}_i^{\mathrm{T}}\underline{\beta}_0}{\sigma_i}\right) \neq \Phi\left(\frac{\underline{X}_i^{\mathrm{T}}\underline{\beta}_0}{\sigma_0}\right), \qquad \text{unless } \sigma_i \equiv \sigma_0 \text{ for all i}$$

i.e. heteroskedasticity implies failure of the score identity.



True DGP (n=1000, $\alpha=\beta=\sigma^2=1$)
OLS ($\hat{\alpha} = 1.19, \hat{\beta} = 0.71, \hat{\sigma}^2 = 0.73$)
Tobit ($\hat{\alpha} = 0.97, \hat{\beta} = 0.98, \hat{\sigma}^2 = 1.07$)

Final Exam

- Cumulative
- Wednesday, December 21, 2016
- 10–12 AM
- In Bolton B80