Empirical Likelihood (EL)

Art Owen (1988, 1990) introduced.

This is nonparametric methodology for creating likelihood-type inference without specifying a joint distributional form for the data.

$\Rightarrow$ we can't misspecify!

EL is going to use the fact that the empirical cdf is a nonparametric MLE to assess how plausible a value of a parameter is to perform inference.

↳ without making distributional assumptions!

1 Mean Case

Suppose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ $\in \mathbb{R}^q$ are iid with mean $\boldsymbol{\mu}$ $\in \mathbb{R}^q$ and covariance-variance Σ ($q \times q$). For simplicity, say we are interested in estimating $\boldsymbol{\mu}$.

Imagine $\underbrace{\text{assigning probabilities}}_{p_i \mapsto \underline{Y}_i}$ $p_1, \ldots, p_n$ to the data $\underline{Y}_1, \ldots, \underline{Y}_n$ where $\underbrace{0 \le p_i \le 1 \text{ and } \sum_{i=1}^{n} p_i = 1}_{(*)}$.

Unlike parametric likelihood, where we assume a functional form for p_i's, only constraints $(*)$.

Define a multinomial likelihood $\prod_{i=1}^{n} p_i$ (likelihood for $\underline{Y}_1, \ldots, \underline{Y}_n$ using $p_1, \ldots, p_n$).

Recall from class (likelihood notes pg 10) if you maximize $\prod_{i=1}^{n} p_i$ for $p_1, \ldots, p_n$ the maximizer $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$ ← 1 observation in each "class" and we have also seen the empirical cdf

$$F_n(\underline{y}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\underline{Y}_i \le \underline{y}) \qquad \underline{y} \in \mathbb{R}^q$$ is the MLE (pg 23 likelihood notes).

in other words, given the data the empirical cdf maximizes $\prod_{i=1}^{n} p_i$.

To perform *nonparametric* likelihood inference on $\boldsymbol{\mu}$, we can consider a constrained multinomial likelihood, known as the **Empirical Likelihood function of $\boldsymbol{\mu}$:**

$$\underbrace{L_n(\overset{\text{function of } \boldsymbol{\mu}}{\boldsymbol{\mu}}|\boldsymbol{Y})}_{\text{EL function}} = \sup\left\{\underbrace{\prod_{i=1}^n p_i}_{\text{multinomial likelihood}} : p_i \mapsto \boldsymbol{Y}_i, \overset{p_i \geq 0}{\vphantom{x}} \sum_{i=1}^n p_i = 1, \underbrace{\overbrace{\sum_{i=1}^n \boldsymbol{Y}_i p_i}^{\text{mean of a dsn } (p_1,\dots,p_n) \text{ on } (Y_1,\dots,Y_n)} = \boldsymbol{\mu}}_{\text{mean constraint on } (p_1,\dots,p_n)}\right\}.$$

Given a parameter value $\underline{\mu}$ and $\underline{Y}$, $L_n(\underline{\mu}|\underline{Y})$ assesses how plausible the value of $\underline{\mu}$ is.

$L_n(\underline{\mu}|\underline{Y})$ is the largest multinomial likelihood possible for a probability assignment to the data having mean $\underline{\mu}$.

The largest possible value of $L_n(\boldsymbol{\mu}|\boldsymbol{Y})$ is

$$\underbrace{\prod_{i=1}^n \frac{1}{n}}_{\substack{\| \\ L_n(\bar{\underline{Y}}, \underline{Y})}} \Rightarrow \underline{\mu} \text{ would be } \sum_{i=1}^n Y_i \cdot \frac{1}{n} \Rightarrow \underline{\mu} \text{ would be } \bar{Y}.$$

So $\bar{Y} = \frac{1}{n}\sum_{i=1}^n \underline{Y}_i$ is a nonparametric ML estimator of $\underline{\mu}$, i.e. the EL estimator $\hat{\underline{\mu}} = \bar{Y}$ of $\underline{\mu}$.

2 Statistical Inference

We can form an EL ratio for $\boldsymbol{\mu}$

$$R_n(\boldsymbol{\mu}) = \frac{L_n(\boldsymbol{\mu}|\boldsymbol{Y})}{L_n(\hat{\boldsymbol{\mu}}|\boldsymbol{Y})}$$

$$= \frac{L_n(\underline{\mu}|\underline{y})}{\prod_{i=1}^{n} \frac{1}{n}}$$

$\hat{\underline{\mu}} = \bar{\underline{y}} \Rightarrow p_i = \frac{1}{n}$

$$= n^n L_n(\underline{\mu}|\underline{y})$$

$$= \sup\left\{ \prod_{i=1}^{n} n p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \underbrace{\sum_{i=1}^{n} \underline{y}_i p_i = \underline{\mu}}_{\underbrace{\sum_{i=1}^{n} (\underline{y}_i - \underline{\mu}) p_i = 0}_{\text{look familiar??}}} \right\}$$

Theorem (Wilk's Theorem): If $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^q$ are iid with mean $\boldsymbol{\mu}_0$ and covariance-variance Σ where $\text{rank}(\Sigma) = q$, then

$$-2 \log R_n(\boldsymbol{\mu}_0) \xrightarrow{d} \chi^2_q \text{ as } n \to \infty.$$

In other words, for $H_0: \underline{\mu} = \underline{\mu}_0 \in \mathbb{R}^q$, if H_0 is true then $-2 \log R_n(\underline{\mu}_0) \xrightarrow{d} \chi^2_q$ as $n \to \infty$.

☆ EL behaves exactly like parametric likelihood for log ratios! ☆

So if $\chi^2_{1-\alpha,q}$ denotes the $1-\alpha$ quantile of χ^2_q, then an approximate $100(1-\alpha)\%$ confidence region for $\underline{\mu}$:

$$CR = \left\{ \underline{\mu} \in \mathbb{R}^q : -2 \log R_n(\underline{\mu}) \leq \chi^2_{1-\alpha,q} \right\}.$$

by inverting the EL test

$$P(\underline{\mu}_0 \in CR) = P\left(-2 \log R_n(\underline{\mu}_0) \leq \chi^2_{1-\alpha,q}\right) \xrightarrow{\text{as } n \to \infty} P\left(\chi^2_q \leq \chi^2_{1-\alpha,q}\right) = 1-\alpha //.$$

For proof of this theorem, see Owen (1988).

3 EL with Estimating Equations

(Qin and Lawless, 1994).

Recall:

For $\underline{Y}_1, \dots, \underline{Y}_n$ ($\in \mathbb{R}^k$) iid and $\underline{\theta} \in \mathbb{R}^b$ a parameter of interest

Estimating equations link a data point $\underline{Y}_i$ to parameters through $r \geq b$ functions.

$$\underline{\Psi}(\underline{Y}_i, \underline{\theta}) \text{ which satisfy } E\underline{\Psi}(\underline{Y}_i, \underline{\theta}) = \underline{0}_r.$$

extends mean example to any estimating equation!

For EL inference on $\boldsymbol{\theta} \in \mathbb{R}^b$, we make an EL function

$$L_n(\underline{\theta}) = \sup\left\{\prod_{i=1}^n p_i : p_i \geq 0, \sum_{i=1}^n p_i = 1, \underbrace{\sum_{i=1}^n p_i \underline{\Psi}(\underline{Y}_i, \underline{\theta}) = \underline{0}_r}_{p_i\text{'s are placed on } \underline{\Psi}(\underline{Y}_i, \underline{\theta}) \text{ to have expectation zero.}}\right\}$$

given value of $\underline{\theta}$

The EL function evaluates the plausibility of a given value of $\underline{\theta}$ based on the data.

Then we can get a point estimate, EL ratio, and corresponding CIs, as well as "profile" (regions) EL:

point estimate: maximize $L_n(\underline{\theta})$ to obtain maximum EL estimator $\hat{\underline{\theta}}$

EL ratio : $R_n(\underline{\theta}) = \dfrac{L_n(\underline{\theta})}{L_n(\hat{\underline{\theta}})}$ (just like parametric likelihood)

Credible region: $CR = \{\underline{\theta} \in \mathbb{R}^b : -2\log R_n(\underline{\theta}) \leq \chi^2_{1-\alpha, b}\}$ (invert EL ratio).

profile EL: suppose $\underline{\theta} = (\underline{\theta}_1, \underline{\theta}_2)$, $\underline{\theta}_1 \in \mathbb{R}^s$, $\underline{\theta}_2 \in \mathbb{R}^{b-s}$. Given $\underline{\theta}_1$ define $\hat{\underline{\theta}}_{2,\theta_1}$ where

$$L_n(\underline{\theta}_1, \hat{\underline{\theta}}_{2,\theta_1}) = \sup_{\theta_2} L_n(\underline{\theta}_1, \underline{\theta}_2)$$

Then the profile EL ratio for $\underline{\theta}_1$ is $R_n(\underline{\theta}_1) = \dfrac{L_n(\underline{\theta}_1, \hat{\underline{\theta}}_{2,\theta_1})}{L_n(\hat{\underline{\theta}})}$.

Main EL result

Theorem: Suppose $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots \in \mathbb{R}^q$ are iid with $\mathrm{E}\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0) = \mathbf{0}_r$ and $\mathrm{Var}[\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)] = \mathrm{E}\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)^\top$ is positive definite, where $\boldsymbol{\theta}_0$ denotes the true parameter value.

Suppose also that $\partial\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\partial^2\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ are continuous in a neighborhood of $\boldsymbol{\theta}_0$ and that, in this neighborhood, $||\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta})||^3$, $||\partial\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}||$ and $||\partial^2\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top||$ are bounded by an integrable function $\Psi(\boldsymbol{Y}_1)$.

Finally, suppose the $r \times b$ matrix $D_{\boldsymbol{\psi}} \equiv \mathrm{E}\partial\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$ has full column rank b.

Then, as $n \to \infty$,

i. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}_b, V)$, where $V = (D_{\boldsymbol{\psi}}^\top \mathrm{Var}[\boldsymbol{\psi}(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)] D_{\boldsymbol{\psi}})^{-1}$. EL point estimates are asymptotically Normal

ii. If $r > b$, the asymptotic variance V cannot increase if an estimating function is added. or decrease if an estimating function is dropped.

iii. To test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, we may use $-2\log R_n(\boldsymbol{\theta}_0)$ and when H_0 is true,

$$-2\log R_n(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_{b} \quad \text{\# parameters}$$

$R_n(\underline{\theta}_0) = \dfrac{L_n(\underline{\theta}_0)}{L_n(\hat{\underline{\theta}})}$

$\Rightarrow$ confidence regions: $CR = \{\underline{\theta} \in \mathbb{R}^b : -2\log R_n(\underline{\theta}) \le \chi^2_{b,1-\alpha}\}$.

iv. If $\underbrace{r > b}_{\text{more functions than parameters.}}$, to test $H_0 : \underbrace{\mathrm{E}\psi(\boldsymbol{Y}_1, \boldsymbol{\theta}) = \mathbf{0}_r}_{\text{moment condition}}$ holds for some $\boldsymbol{\theta}$, we may use

$$-2\log \frac{L_n(\hat{\boldsymbol{\theta}})}{\underbrace{\prod_{i=1}^{n}(1/n)}_{\text{biggest possible value could ever have for and EL function w/ no moment constraints}}} = -2\log(n^n L_n(\hat{\theta})).$$

and when H_0 is true this quantity converges in distribution to $\chi^2_{\underbrace{r-b}_{\text{\# excess estimating functions.}}}$.

Asymptotically, $-2\log R_n(\theta_0)$ and $-2\log n^n L_n(\hat{\theta}))$ are independent.

v. To test the profile assumption $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0 \in \mathbb{R}^s$, we can use the profile EL ratio $-2\log R_n(\boldsymbol{\theta}_1^0)$ and , when H_0 is true, $-2\log R_n(\boldsymbol{\theta}_1^0) \xrightarrow{d} \chi^2_s$.

\# parameters in EL functions after profiling.