Another example: estimating equations. Suppose $Y_1, \ldots, Y_n$ independent and $x_1, \ldots, x_n$ constants.

Given: a function $g(y, x, \theta)$ such that

$$\mathrm{E}\left[g(Y_i, x_i, \theta_o)\right] = 0$$

for all i and some particular θ_o.

Includes linear and generalized linear model problems.

Define a random function of θ by

$$S_n(\theta) = \sum_{i=1}^{n} g(Y_i, x_i, \theta)$$

Plan: estimate θ by solving the equation

$$S_n(\theta) = 0 \qquad (1)$$

for θ to get $\widehat{\theta}_n$.

Study large sample behaviour of $\widehat{\theta}_n$.

Hoped for behaviour:

1. $\widehat{\theta}_n$ is *weakly consistent*: for each $\epsilon > 0$

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta_o| > \epsilon) = 0.$$

2. $\widehat{\theta}_n$ is *strongly consistent*:

$$P(\lim_{n\to\infty} \widehat{\theta}_n = \theta_o) = 1.$$

3. $\widehat{\theta}_n$ is asymptotically normal.

Do one example:

$$\epsilon_i = (Y_i - \alpha_o x_i)/\beta_o$$

are iid standard Cauchy. Log likelihood is

$$\ell_n(\theta) = -\sum_1^n \log\left\{1 + \beta^{-2}(Y_i - x_i\alpha)^2\right\} \\ - n\log\pi - n\log\beta$$

Score function has two components:

$$U_{n\alpha}(\theta) = \sum_{i=1}^n \frac{2\beta^{-2}x_i(Y_i - x_i\alpha)}{\left\{1 + \beta^{-2}(Y_i - x_i\alpha)^2\right\}^2}$$

and

$$U_{n\beta}(\theta) = \sum_{i=1}^n \frac{2\beta^{-3}(Y_i - x_i\alpha)^2}{\left\{1 + \beta^{-2}(Y_i - x_i\alpha)^2\right\}^2} - \frac{n}{\beta}$$

We will show the existence of a unique consistent root of the likelihood equations

$$U_n(\theta) = 0.$$

Notice: if all $x_i = 0$ (or all $x_i \approx 0$) then get no (or little) information about α.

If one x_i very large then estimate of α largely determined by that one data point.

Assumptions

A1: The constants x_i satisfy

$$0 < \liminf \frac{1}{n} \sum_1^n x_i^2$$

A2: There is $\delta > 0$ so that constants x_i satisfy

$$\limsup \frac{1}{n} \sum_1^n x_i^{2+\delta} < \infty$$

A3: There is $\delta > 0$ so that constants x_i satisfy

$$\limsup \frac{1}{n} \sum_1^n |x_i|^{3+\delta} < \infty$$

Tools for proving a set of equations has a root in some domain:

- If equations are derivative of scalar function a local max or min of the function interior to domain corresponds to a root.

- Brouwer fixed point theorem: if f is continuous then $f(x) = x$ must have a root in any set K which is compact and for which $f(K) \subset K$. See Atchison & Silvey, *Ann Math Statist*, 1958.

- Contraction mapping theorem: if $\exists \alpha < 1$ with

$$|f(y) - f(x)| \le \alpha |y - x|$$

for $f : \mathbb{R}^p \to \mathbb{R}^p$ then f has a fixed point.

Note: $g(x) = 0$ iff $f(x) = x$ where $f(x) = g(x) + x$.

In what follows a will be some positive real and

$$A_n = \{\ell_n \text{ is strictly concave on } B_a(\theta_o)\}$$

Theorem 1 *There is an $a > 0$ such that*

$$P(A_n) \to 1$$

as $n \to \infty$. In fact

$$P(\cup_{N=1}^{\infty} \cap_{n=N}^{\infty} A_n) = 1$$

In words ℓ_n is almost surely strictly concave for all large n on some neighbourhood of the true parameter values. A strictly concave function can have no more than 1 local maximum on the set in question.

Now let

$$\bar{\ell}_n(a) = \text{sup}\{\ell(\theta) : |\theta - \theta_o| = a\}$$

and

$$B_n = \left\{\ell_n(\theta_o) > \bar{\ell}_n(a)\right\}$$

Theorem 2 *There is an* $a > 0$ *such that*

$$P(B_n) \to 1$$

An almost sure version is true, too.

When B_n happens there must be at least one root in the ball in question.

Let $C_n(a)$ denote the event that there is a unique root of U_n inside the ball $B_a(\theta_o)$ Note that $C_n(a) \supset A_n \cap B_n$.

WARNING: I won't try to prove C_n is really an event.

Let $D_{n\epsilon}$ be the event $C_n(a) \cap C_n(\epsilon)$ (unique root in big ball and the root is in a small ball).

From the theorems so far: for each fixed $\epsilon > 0$

$$P(D_{n\epsilon}) \to 1$$

Useful lemma: if $a_{n\epsilon}$ is a sequence of numbers for each $\epsilon > 0$ and for each fixed $\epsilon > 0$ we have

$$\lim_{n\to\infty} a_{n\epsilon} = 0$$

then there is a sequence $\epsilon_n \to 0$ such that

$$\lim_{n\to\infty} a_{n\epsilon_n} = 0$$

Hence there is a sequence $\epsilon_n \to 0$ such that

$$P(D_{n\epsilon_n}) \to 1$$

Define $\widehat{\theta}_n$ as follows: on the event $A_n \cap B_n$ let $\widehat{\theta}_n$ be the unique root of U_n inside $B_a(\theta_o)$. On the complement of this event put

$$\widehat{\theta}_n = (\widehat{\alpha}_n, \widehat{\beta}_n) = (0, 1)$$

I also won't prove $\widehat{\theta}_n$ is a random variable but it is.

Theorem 3 *The sequence $\widehat{\theta}_n$ is weakly consistent.*

This is the content of the assertion on the previous slide.

WARNING: $\widehat{\theta}_n$ is not an estimator; definition depends on θ_o.

These things are all proved by studying derivatives. Define

$$E_i = (Y_i - x_i \alpha)/\beta$$

Write the log likelihood as

$$\sum \ell_i(\theta)$$

where

$$\ell_i(\theta) = -\log(1 + E_i^2) - \log(\beta)$$

I claim there is are bounded functions g_{rs} such that

$$\frac{\partial^{r+s}}{\partial \alpha^r \partial \beta^s} \ell_i(\theta) = \frac{x_i^r}{\beta^{r+s}} g_{rs}(E_i)$$

(I only need this for $r + s \leq 3$.)

Proof of Theorem 1: Suppress n. We let $\bar{U}$, $\bar{V}$ and $\bar{W}$ denote the arrays of first, second and third derivatives of ℓ/n.

Each entry in $\bar{W}$ has the form

$$\beta^{-3}\sum_i x_i^r g_{rs}(E_i)/n$$

where $r+s=3$. Every such quantity is no more than

$$\beta^{-3}\sum_i |x_i|^r/n \le \beta^{-3}(1+\sum_i |x_i|^3/n).$$

Under assumption **A3** this quantity is bounded over all pairs α,β with $\beta > c > 0$ for any fixed positive c. Fix some $a < \beta_o$. For each fixed θ we have

$$\bar{V}(\theta) - E(\bar{V}(\theta)) \to 0 \qquad (2)$$

in probability by an application of a weak law of large numbers. (One I know applies is in a 1988 Annals paper I wrote; a variance calculation doesn't quite work.)

Now consider the maps

$$\mathbf{I}(\theta) = -\mathsf{E}(\bar{V}(\theta))$$

We find

$$\mathbf{I}(\theta_o) = \begin{matrix} \sum x_i^2/(2n) & 0 \\ 0 & 1/2 \end{matrix}$$

which is positive definite. Let M_3 denote the upper bound on the third derivatives from the last slide.

If

$$||\theta - \theta_o|| < \delta \equiv \min\{\sum x_i^2/n, 1/2\}/(4M_3)$$

then $\mathbf{I}(\theta)$ must be positive definite. In fact we must have

$$\begin{aligned} \mathbf{I}_{11}(\theta) &\geq 3\delta/4 \\ \mathbf{I}_{22}(\theta) &\geq 3\delta/4 \\ \mathbf{I}_{12}(\theta) &\leq \delta/4 \\ \det \mathbf{I}(\theta) &\geq \delta/2 \end{aligned}$$

We now take $a = \delta$ and prove Theorem 7.

Fix $\epsilon > 0$. Choose N so large that for $n \geq N$ we have

$$P(|V_{ij}(\theta_o) + \mathbf{I}_{ij}(\theta_0)| \geq \delta/4) \leq \epsilon$$

Find points $\theta_1, \ldots, \theta_N$ all in $B_a(\theta_o)$ with the property that

$$B_a(\theta_o) \cup B_\delta$$

Fix some θ. We need some notation for the entries in H and S/n. Let

$$
\begin{aligned}
E_i &= (Y_i - x_i\alpha)/\beta \\
U_{i\alpha} &= 2\beta^{-1} x_i E_i / \left\{1 + E_i^2\right\} \\
U_{i\beta} &= \frac{2\beta^{-1} E_i^2}{\left\{1 + E_i^2\right\}} - \frac{1}{\beta} \\
H_{i\alpha\alpha} &= \beta^{-2} \left(\frac{4E_i^2 x_i^2}{\left\{1 + E_i^2\right\}^2} - \frac{2x_i^2}{1 + E_i^2} \right) \\
H_{i\alpha\beta} &= \beta^{-2} \left(\frac{4E_i^3 x_i}{\left\{1 + E_i^2\right\}^2} - \frac{4x_i E_i}{1 + E_i^2} \right) \\
H_{i\beta\beta} &= \beta^{-2} \left(\frac{4E_i^4}{\left\{1 + E_i^2\right\}^2} - \frac{6E_i^2}{1 + E_i^2} + 1 \right)
\end{aligned}
$$

Theorem 4 *The score function at the true parameter value is asymptotically normal.*

As function of Y_i each term in the score is bounded so

$$\begin{aligned} \mathsf{E}\left[S_n(\theta_o)\right] &= 0 \\ \mathsf{Var}\left[S_n(\theta_o)\right] &\equiv \Sigma_n < \infty \end{aligned}$$

We try to apply the Lyapunov CLT to the score function.

Terms in S_{n1} are not iid.

To state the result we will fix an $a > 0$ and a sequence ϵ_n of constants converging to 0 and define events

$$A_n = \{\ell \text{ is concave on } B_a(\theta_0)\}$$
$$B_n = \{\ell(\theta_0) > \sup\{\ell(\theta) : |\theta - \theta_0 = \epsilon_n\}$$

Define a random variable $\tilde{\theta}$ as follows. On the event $A_n \cap B_n$ we define $\tilde{\theta}$ to be the uniqe root of $U(\theta) = 0$ in $B_a(\theta_0)$. On the complement of this event put $\tilde{\theta}_n = \theta_0$ (for definiteness). It is true, but I will not prove that $\tilde{\theta}$ so defined is a random variable.

Put

$$\mathcal{I} = \begin{bmatrix} \frac{\sum x_i^2}{2} & 0 \\ 0 & \frac{n}{2} \end{bmatrix}$$

By $\mathcal{I}^{1/2}$ and $\mathcal{I}^{-1/2}$ we mean the non-negative definite diagonal square roots of $\mathcal{I}$ and $\mathcal{I}^{-1}$.

Theorem 5 *Under conditions A1 and A3 we have:*

1. *There is a constant* $a > 0$ *such*

$$P(A_n) \to 1$$

2. *There is a sequence* $\epsilon_n \to 0$ *such that*

$$P(B_n) \to 1$$

3. *The sequence of random variables* $\tilde{\theta}$ *is consistent for* θ_0 *in the sense that* $\tilde{\theta} \to \theta_0$ *in probability.*

4. *The random variables* $\tilde{\theta}$ *are asymptotically normal in the sense*

$$\mathcal{I}^{1/2}(\tilde{\theta} - \theta_0) \Rightarrow MVN(\mathbf{0}, \mathbf{I})$$

where $\mathbf{I}$ *is the two by two identity.*

Proof of 4: This conclusion is based on Taylor expansion of the identity (on $A_n \cap B_n$)

$$U(\tilde{\theta}) = 0$$

We write that expansion in the form

$$0 = U(\theta_0) + V(\theta_0)(\tilde{\theta} - \theta) + R \qquad (3)$$

where the remainder term R is given by

$$R_i = \sum_{jk} \int_0^1 (1-t) W_{ijk}(\theta_0 + t(\tilde{\theta} - \theta_))dt \\ \times (\tilde{\theta}_j - \theta_{0j})(\tilde{\theta}_k - \theta_{0k})$$

Step 1: There is a constant M such that for all n:

$$|W_{ijk}| \le Mn.$$

It follows (using say Cauchy Schwarz) that on $A_n \cap B_n$

$$|R_i| \le Mn|\tilde{\theta} - \theta_0|^2 \le Mn\epsilon_n|\tilde{\theta} - \theta_0|$$

and

$$|R| \leq 3Mn|\tilde{\theta} - \theta_0|^2 \leq 3Mn\epsilon_n|\tilde{\theta} - \theta_0|$$

Step 2: Multiply 3 by $\mathcal{I}^{-1/2}$ and get

$$0 = \mathcal{I}^{-1/2}U(\theta_0) + \mathcal{I}^{-1/2}V(\theta_0)\mathcal{I}^{-1/2}\mathcal{I}^{1/2}(\tilde{\theta} - \theta_0) + \mathcal{I}^{-1}$$
$$\equiv T_1 + T_2 + T_3$$

say.

Step 3: We show that T_1 converges in distribution to standard bivariate normal by applying the Lindeberg central limit theorem and the Cramér-Wold device. It suffices to show that for each unit vector $\mathbf{a}$ is $\mathbb{R}^2$ we have

$$\mathbf{a}'\mathcal{I}^{-1/2}U(\theta_0) \Rightarrow N(0,1).$$

The vector $\mathcal{I}^{-1/2}U(\theta_0)$ has two components which we label temporarily X_n and Y_n. Define

$$A_i = 2\frac{\epsilon_i}{1+\epsilon_i^2}$$

$$B_i = 2\frac{\epsilon_i^2}{1+\epsilon_i^2} - 1$$

$$X_{in} = \frac{x_i A_i}{\sqrt{\sum x_i^2/2}}$$

$$Y_{in} = \frac{B_i}{\sqrt{n/2}}$$

Then

$$X_n = \sum_i X_{in}$$

and

$$Y_n = \sum_i Y_{in}$$

Direct computation shows that

$$\begin{aligned} \mathrm{E}(X_n) = \mathrm{E}(Y_n) &= 0 \\ \mathrm{Var}(X_n) = \mathrm{Var}(Y_n) &= 1 \\ \mathrm{Cov}(X_n, Y_n) &= 0 \end{aligned}$$

If we put $Z_{in} = a_1 X_{in} + a_2 Y_{in}$ and $Z_n = \sum_i Z_{in}$ then

$$\mathbf{a}' \mathcal{I}^{-1/2} U(\theta_0) = Z_n$$

and the triangular array $\{Z_{in}\}$ is independent within rows. It thus remains to check Lindeberg's condition. To this end we note

If two triangular arrays $\{X_{in}\}$ and $\{Y_{in}\}$ each satisfy Lindeberg's condition then for each a_1

and a_2 the triangular array $\{Z_{in} = a_1 X_{in} + a_2 Y_{in}\}$ satisfies Lindeberg's condition.

Proof: The assertion is trivial if either $a_1 = 0$ or $a_2 = 0$ so we assume both are non-zero. Fix $\epsilon > 0$ and note that

$$Z_{in}^2 \le 2a_1^2 X_{in}^2 + 2a_2 Y_{in}^2$$

and that

$$1(|Z_{in}| > \epsilon) \le 1(|a_1 X_{in}| > \epsilon/2) + 1(|a_2 Y_{in}| > \epsilon/2)$$

Hence

$$Z_{in}^2 1(|Z_{in}| > \epsilon) \le 2a_1^2 X_{in}^2 1(|a_1 X_{in}| > \epsilon/2) + 2a_1^2 Y_{in}^2 1(|a_1 X$$

Moreover by considering each case of the indicators involved we may check that

$$a_1^2 Y_{in}^2 1(|a_1 X_{in}| > \epsilon/2) \le a_1^2 Y_{in}^2 1(|a_1 Y_{in}| > \epsilon/2) + a_1^2 X_{in}^2 1($$

and the analogous inequality reversing the roles of X and Y. Combining these gives

$$Z_{in}^2 1(|Z_{in}| > \epsilon) \le 4a_1^2 X_{in}^2 1(|a_1 X_{in}| > \epsilon/2) + 2a_2^2 X_{in}^2 1(|a_2$$

Put

It remains to check Lindeberg's condition for the specific arrays $X_{in}\}$ and $\{Y_{in}\}$ given above. Define

$$G_X(M) = \mathsf{E}(A_i^2 1(|A_i| > M))$$
$$G_Y(M) = \mathsf{E}(B_i^2 1(|B_i| > M))$$

and note that since the As and Bs have a finite variance we have

$$\lim_{M\to\infty} G_X(M) = \lim_{M\to\infty} G_Y(M) = 0.$$

Then

$$\sum_i \mathsf{E}(Y_{in}^2 1(|Y_{in}| > \epsilon)) = 2G_Y(\epsilon\sqrt{n/2})$$

which converges to 0. Second

$$\sum_i \mathsf{E}(X_{in}^2 1(|X_{in}| > \epsilon)) = \sum_i x_i^2 G_X(\epsilon\sqrt{\sum x_i^2/2}/|x_i|)/\sum_j x$$
$$\leq G_X(\epsilon\sqrt{\sum x_i^2/2}/\max\{|x_i|\})$$

Now in view of assumption 3 we must have

$$|x_i|^3 \le Mn$$

for all i so by assumption 1

$$\frac{x_i^2}{\sum_j x_j^2} \le \frac{(Mn)^{2/3}}{n\delta}$$

which shows

$$\frac{\max\{x_i^2\}}{\sum_1^n x_j^2} \to 0.$$

This ends Step 3.

Step 4: Define

$$\mathbf{M} \equiv -\mathcal{I}^{-1/2} V(\theta_0) \mathcal{I}^{-1/2}$$

I claim that

$$\mathbf{M} \to \mathbf{I}$$

in probability.

We now need a few pieces of elementary linear algebra. In the following $\mathbf{A}$ is a real $k \times k$ matrix and $\mathbf{x}$ a k-vector.

1. A symmetric matrix $\mathbf{A}$ can be written in the form

   $$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$$

   where the matrix $\mathbf{\Lambda}$ is diagonal and the matrix $\mathbf{P}$ is orthogonal, that is,

   $$\mathbf{P}\mathbf{P}' = \mathbf{I}$$

   The columns of $\mathbf{P}$, say $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are eigenvectors of $\mathbf{A}$ with corresponding eigenvalue

λ_i being the ith diagonal entry in $\mathbf{\Lambda}$. That is

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

2. We define the Euclidean (2) norm of matrix $\mathbf{A}$ by

$$\begin{aligned} |\mathbf{A}| &= \sqrt{\text{trace}(\mathbf{A}\mathbf{A}')} \\ &= \sqrt{\sum_{ij} A_{ij}^2} \end{aligned}$$

3. We have by Cauchy-Schwarz:

$$|\mathbf{A}\mathbf{x}| \leq |\mathbf{A}||\mathbf{x}|$$

4. If λ_{min} and λ_{max} are the smallest and largest eigenvalue of a symmetric matrix $\mathbf{A}$ then for all $\mathbf{x}$ we have

$$\lambda_{\text{min}}|\mathbf{x}| \leq |\mathbf{A}\mathbf{x}| \leq \lambda_{\text{max}}|\mathbf{x}|$$

We will also show

$$\mathbf{M} \equiv -\mathcal{I}^{-1/2} V(\theta_0) \mathcal{I}^{-1/2} \to \mathbf{I}$$

in probability.

This will show:

1. As $n \to \infty$

$$P(\mathbf{M} \text{ is invertible}) \to 1$$

2. If C_n is the event that $\mathbf{M}$ is invertible then on $A_n \cap B_n \cap C_n$ we may write

$$\mathcal{I}^{1/2}(\tilde{\theta} - \theta_0) = \mathbf{M}^{-1} T_1 + \mathbf{M}^{-1} T_3$$

3. The smallest eigenvalue of the matrix $\mathcal{I}^{1/2}$ is at least

$$\lambda_{\text{min}} \sqrt{n} \equiv \sqrt{n \text{min}\{\delta, 1\}/4}$$

where δ is the quantity in assumption A1. Hence

$$|\mathcal{I}^{1/2}(\tilde{\theta} - \theta_0)| \geq \lambda_{\text{min}} \sqrt{n} |\tilde{\theta} - \theta_0|$$

At the same time

$$|\mathcal{I}^{1/2}(\tilde{\theta} - \theta_0)| \leq |\mathbf{M}^{-1} T_1| + |\mathbf{M}^{-1} T_3|$$

Since

$$|T_3| \leq 3Mn\epsilon_n||\tilde{\theta} - \theta_0|$$

and

$$\sum_{ij} \mathbf{M}_{ij}^2 \to 2$$

we find

$$\lambda_{\text{min}} \sqrt{n} |\tilde{\theta} - \theta_0| \leq |\mathbf{M}^{-1} T_1| + 6Mn$$