Lecture Stat 461-561
M-Estimation

AD

February 2008

Introduction & Motivation

- In most applications, we have $X_i \overset{\text{i.i.d.}}{\sim} g$ and we obtain an estimate $\widehat{a}$ by minimizing a suitable cost function; e.g.
  - the mean corresponds to $\sum_{i=1}^{n} (\theta - x_i)^2$.
  - the median corresponds to $\sum_{i=1}^{n} |\theta - x_i|$.
  - the MLE corresponds to negative log-likelihood $-\sum_{i=1}^{n} \log f(x_i | \theta)$.
- However, even the mean estimate of a location parameter is typically not robust. In contrast, the median could be too 'rough'.
- Example: Consider

$$\mathbf{x} = (-1.28, -0.96, -0.46, -0.44, -0.26, -0.21, -0.063, 0.39, 3, 6, 9)$$

where $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $i = 1, ..., 8$ but X_9, X_{10}, X_{11} are outliers.

- The mean is 1.33 and the median is -0.21.
- Huber (1964) introduced a general loss function which is a compromise between mean and median; i.e. we minimize

$$\sum_{i=1}^{n} \rho\left(x_i - \theta\right)$$

where

$$\rho\left(x\right) = \begin{cases} \frac{1}{2}x^2 & \text{if } \left|x\right| \leq k \\ k\left|x\right| - \frac{1}{2}k^2 & \text{if } \left|x\right| \geq k \end{cases}$$

- Large observations are not as heavily weighted as for $\sum_{i=1}^{n} \left(\theta - x_i\right)^2$.
- k is a tuning parameter which controls the mix between the mean and median-like estimators.

- Results

k	0	1	2	3	4	5	6	8	10
Estimate	-.21	.03	-.04	.29	.41	.52	.87	.97	1.33

Huber's estimator as a function of k

- When $k = 0$, Huber's estimator corresponds to the median and as k increases it gets closer to the mean; i.e. the robustness properties of the estimator are decreasing.
- *Remark*: Clearly minimizing

$$\sum_{i=1}^{n} \rho\left(x_i - \theta\right)$$

appears equivalent to maximizing a log-likelihood for which $\log f\left(x_i \mid \theta\right) = cste - \rho\left(x_i - \theta\right)$. We will describe later more general estimates which do not support such an interpretation.

Basic Approach

- Now we assume a more general case where $X_i \sim g$ and our estimate $\widehat{\theta}_n$ is solution of
$$\sum_{i=1}^{n} \psi(x_i, \theta) = 0.$$
- Under regularity conditions $\widehat{\theta}_n$ will converge towards the parameter θ^* satisfying
$$\mathbb{E}_g [\psi(X, \theta^*)] = \int \psi(x, \theta^*) g(x) dx = 0.$$
- If $g(x) = f(x|\theta_0)$ then θ^* is defined by
$$\mathbb{E}_{f(\cdot|\theta)} [\psi(X, \theta^*)] = \int \psi(x, \theta^*) f(x|\theta_0) dx = 0,$$
i.e. be careful: we do not have necessarily $\theta^* = \theta_0$! Also in practice, we would like it to be the case.
- For example, if $\psi(x, \theta) = x - \theta$ then $\theta^* = \mathbb{E}_g [X]$.

- To study the asymptotic properties of $\widehat{\theta}_n$, we use the (now standard) Taylor expansion method of $\sum_{i=1}^{n} \psi(\theta, x_i)$ around the value θ^* which yields

$$0 = \sum_{i=1}^{n} \psi(x_i, \theta^*) + \left(\widehat{\theta}_n - \theta^*\right) \sum_{i=1}^{n} \psi'(x_i, \theta^*) + R_n$$

- By ignoring the remainder term R_n, we obtain

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) = \frac{-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i, \theta^*)}{\frac{1}{n} \sum_{i=1}^{n} \psi'(x_i, \theta^*)}$$

- The CLT yields

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(x_i,\theta^*\right)\xrightarrow{\text{D}}\mathcal{N}\left(0,\mathbb{E}_g\left[\psi^2\left(X,\theta^*\right)\right]\right)$$

as $\mathbb{E}_g\left[\psi\left(X_i,\theta^*\right)\right]=0$ and $var_g\left[\psi\left(X_i,\theta^*\right)\right]=\mathbb{E}_g\left[\psi^2\left(X,\theta^*\right)\right]$.

- The law of large numbers provides

$$\frac{1}{n}\sum_{i=1}^{n}\psi'\left(x_i,\theta^*\right)\xrightarrow{\text{P}}\mathbb{E}_g\left[\psi'\left(X,\theta^*\right)\right]$$

- So by Slutsky's theorem, we get

$$\sqrt{n}\left(\widehat{\theta}_n-\theta^*\right)\xrightarrow{\text{D}}\mathcal{N}\left(0,\frac{\mathbb{E}_g\left[\psi^2\left(X,\theta^*\right)\right]}{\mathbb{E}_g\left[\psi'\left(X,\theta^*\right)\right]^2}\right).$$

- This is a generalization of the misspecified model we discussed before where ψ is arbitrary.

Study of the Huber's estimate

- For the Huber's estimate, we have $\psi(x,\theta) = \psi(x-\theta)$ where

$$\psi(x) = \begin{cases} x & \text{if } |x| \leq k \\ k & \text{if } x \geq k \\ -k & \text{if } x < -k \end{cases}.$$

- Assume we have $X_i \overset{\text{i.i.d.}}{\sim} f(x-\theta)$ where f is symmetric around 0 and we want to estimate θ then indeed

$$\begin{aligned} &\mathbb{E}[\psi(X-\theta)] \\ &= \int_{\theta-k}^{\theta+k} (x-\theta) f(x-\theta)\,dx - k\int_{-\infty}^{\theta-k} f(x-\theta)\,dx \\ &+k\int_{\theta+k}^{+\infty} f(x-\theta)\,dx \\ &= \int_{-k}^{k} uf(u)\,du - k\int_{-\infty}^{-k} f(u)\,du + k\int_{k}^{+\infty} f(u)\,du \\ &= 0. \end{aligned}$$

- In this case we have indeed that $\theta^* = \theta$!

- We also have

$$\mathbb{E}\left[\psi'\left(X-\theta\right)\right] = \int_{\theta-k}^{\theta+k} f\left(x-\theta\right) dx = P_{\theta}\left(\left|X\right| \leq k\right),$$

$$\begin{aligned}\mathbb{E}\left[\psi^{2}\left(X-\theta\right)\right] &= \int_{\theta-k}^{\theta+k} \left(x-\theta\right)^{2} f\left(x-\theta\right) dx + k^{2} \int_{-\infty}^{\theta-k} f\left(x-\theta\right) dx \\ &\quad + k^{2} \int_{\theta+k}^{+\infty} f\left(x-\theta\right) dx \\ &= \int_{-k}^{k} u^{2} f\left(u\right) du + 2k^{2} \int_{k}^{+\infty} f\left(u\right) du.\end{aligned}$$

- It follows that the Huber's estimate satisfies

$$\sqrt{n}\left(\widehat{\theta}_{n}-\theta^{*}\right) \xrightarrow{\text{D}} \mathcal{N}\left(0, \frac{\int_{-k}^{k} u^{2} f\left(u\right) du + 2k^{2} P_{\theta^{*}}\left(\left|X\right| > k\right).}{\left[P_{\theta^{*}}\left(\left|X\right| \leq k\right)\right]^{2}}\right)$$

- We compare the asymptotic relative efficiencies of Huber's estimate for $k = 1.5$ to mean and median

	Normal	Double Exponential
vs. mean	.96	1.37
vs. median	1.51	.68

that is $\sigma^2_{\text{Huber}}/\sigma^2_{\text{mean}}$ and $\sigma^2_{\text{Huber}}/\sigma^2_{\text{median}}$.

- Remember that mean is the MLE of normal and median is the MLE of double exponential so ARE are <1 as expected.
- Huber's estimator performs however reasonably well compared to the MLE.

- An M-estimator is a tradeoff between robustness and efficiency.
- To see how much we are losing, we study in more details the asymptotic variance given by $\mathbb{E}\left[\psi'\left(X,\theta^*\right)\right]^{-2}\mathbb{E}\left[\psi^2\left(X,\theta^*\right)\right]$.
- We have

$$\mathbb{E}\left[\psi'\left(X,\theta\right)\right] = -\int \frac{d\psi\left(x,\theta\right)}{d\theta} f\left(x|\,\theta\right) dx$$

where

$$\begin{aligned}\frac{d}{d\theta}\int \psi\left(x,\theta\right) f\left(x|\,\theta\right) dx &= \int \frac{d\psi\left(x,\theta\right)}{d\theta} f\left(x|\,\theta\right) dx \\ &\quad + \int \psi\left(x,\theta\right) \frac{df\left(x|\,\theta\right)}{d\theta} dx\end{aligned}$$

so **if** $\int \psi\left(x,\theta\right) f\left(x|\,\theta\right) dx = 0$ for all θ then

$$\begin{aligned}\mathbb{E}\left[\psi'\left(X,\theta\right)\right] &= \int \psi\left(x,\theta\right) \frac{df\left(x|\,\theta\right)}{d\theta} dx \\ &= \int \psi\left(x,\theta\right) \frac{d\log f\left(x|\,\theta\right)}{d\theta} f\left(x|\,\theta\right) dx\end{aligned}$$

- Recall that the asymptotic variance of the MLE is in $\mathbb{E}_{\theta}\left[\frac{d \log f(X|\theta)}{d\theta}^2\right]^{-1}$ thus

$$ARE = \frac{var\,(\text{MLE})}{var\,(\text{M})} = \frac{\mathbb{E}\left[\psi\,(X,\theta)\,\frac{d \log f(x|\theta)}{d\theta}\right]^2}{\mathbb{E}\left[\psi^2\,(X,\theta^*)\right]\mathbb{E}_{\theta}\left[\frac{d \log f(X|\theta)}{d\theta}^2\right]} \leq 1$$

  follows from the Cauchy-Schwartz inequality.
- An M-estimate is always less efficient than the MLE and matches its efficiency only if $\psi\,(x,\theta)$ is proportional to $\frac{d \log f(x|\theta)}{d\theta}$.
- This result does not say much; if one uses an M-estimate it is because it it not believed that the model $f\,(x|\,\theta)$ is reliable...

General multivariate case

- We want to estimate the multidimensional parameter θ^* which satisfies
$$\mathbb{E}\left[\psi\left(X,\theta^*\right)\right]=\int\psi\left(x,\theta^*\right)f\left(x\right)dx=0.$$
- This extension is trivial theoretically but will allow us to study numerous interesting estimates; e.g. consider the estimate
$$\widehat{\theta}_{1,n}=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\overline{x}\right|.$$
- At first glance, this is not an M-estimate as there is no single equation of the form
$$\sum_{i=1}^{n}\psi\left(x_i,\theta_1\right)=0$$
that yields $\widehat{\theta}_{1,n}$.
- Moreover there is no family of densities $f\left(x\middle|\theta\right)$ such that $\widehat{\theta}_{1,n}$ is a component of the MLE of θ.

- However, we can write

$$\psi(x,\theta) = \begin{pmatrix} \psi_1(x,\theta_1,\theta_2) \\ \psi_2(x,\theta_1,\theta_2) \end{pmatrix} = \begin{pmatrix} |x-\theta_2| - \theta_1 \\ (x-\theta_2) \end{pmatrix}$$

- We find out that

$$\sum_{i=1}^{n} \psi(x_i,\theta_1,\theta_2) = 0$$

implies $\widehat{\theta}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\widehat{\theta}_{1,n} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|$.

- Asymptotic results can be easily establish using a straightforward generalization of the scalar case and, under regularity assumptions, we obtain

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) \xrightarrow{\text{D}} \mathcal{N}\left(0, V\left(\theta^*\right)\right)$$

  where

$$V\left(\theta^*\right) = A^{-1}\left(\theta^*\right) B\left(\theta^*\right) \left\{A^{-1}\left(\theta^*\right)\right\}^{\mathsf{T}}$$

  with

$$A\left(\theta^*\right) = \mathbb{E}\left[-\frac{\partial \psi\left(X,\theta\right)}{\partial \theta^{\mathsf{T}}}\right]\Bigg|_{\theta=\theta^*}, \; B\left(\theta^*\right) = \mathbb{E}\left[\psi\left(X,\theta^*\right)\psi\left(X,\theta^*\right)^{\mathsf{T}}\right].$$

- Clearly if $\psi\left(x,\theta\right) = \frac{\partial \log f\left(x|\theta\right)}{\partial \theta}$ and if the data truly come from the assumed parametric family $f\left(x|\,\theta\right)$ then

$$A\left(\theta^*\right) = B\left(\theta^*\right) = I\left(\theta^*\right) \Rightarrow V\left(\theta^*\right) = I\left(\theta^*\right)^{-1}.$$

- However, in many cases the data do not come from the assumed family and valid inference should be carried out using the correct limiting covariance matrix.

- Let us define $G_n(\theta) := \sum_{i=1}^{n} \psi(x_i, \theta)$. The idea of the proof is always the same

$$G_n\left(\widehat{\theta}_n\right) = 0 = G_n(\theta^*) + G_n'(\theta^*)\left(\widehat{\theta}_n - \theta^*\right) + R_n$$

where $G_n'(\theta^*) = \left[\frac{\partial G_n(\theta)}{\partial \theta^{\mathsf{T}}}\right]\Big|_{\theta=\theta^*}$ so

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^*\right) = -\left[G_n'(\theta^*)\right]^{-1}\sqrt{n}G_n(\theta^*) + \sqrt{n}R_n^*$$

- Moreover under regularity conditions

$$-G_n'(\theta^*) \xrightarrow{\mathsf{P}} A(\theta^*),$$
$$\sqrt{n}G_n(\theta^*) \xrightarrow{\mathsf{D}} \mathcal{N}(0, B(\theta^*)),$$
$$\sqrt{n}R_n^* \xrightarrow{\mathsf{P}} 0.$$

- We can estimate $A(\theta^*)$, $B(\theta^*)$ and $V(\theta^*)$ using the data samples via

$$\begin{aligned} A_n\left(\widehat{\theta}_n, \mathbf{x}\right) &= \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial\psi\left(x_i, \widehat{\theta}_n\right)}{\partial\theta^{\mathsf{T}}}, \\ B\left(\widehat{\theta}_n, \mathbf{x}\right) &= \frac{1}{n}\sum_{i=1}^{n} \psi\left(x_i, \widehat{\theta}_n\right)\psi\left(x_i, \widehat{\theta}_n\right)^{\mathsf{T}}, \\ V\left(\widehat{\theta}_n, \mathbf{x}\right) &= A^{-1}\left(\widehat{\theta}_n, \mathbf{x}\right)B\left(\widehat{\theta}_n, \mathbf{x}\right)\left\{A^{-1}\left(\widehat{\theta}_n, \mathbf{x}\right)\right\}^{\mathsf{T}} \end{aligned}$$

- Under mild regularity assumptions, we have

$$V\left(\widehat{\theta}_n, \mathbf{x}\right) \xrightarrow{\mathsf{P}} V(\theta^*)$$

- An interesting extension consists of considering

$$\sum_{i=1}^{n} \psi\left(x_i, \theta\right) = c_n$$

where $c_n / \sqrt{n} \xrightarrow{\text{P}} 0$ as $n \longrightarrow \infty$.

- In these cases, the asymptotic results still hold as we can simply write

$$\begin{aligned} G_n\left(\widehat{\theta}_n\right) - c_n &= 0 = G_n\left(\theta^*\right) - c_n + G_n'\left(\theta^*\right)\left(\widehat{\theta}_n - \theta^*\right) + R_n \\ &= G_n\left(\theta^*\right) + G_n'\left(\theta^*\right)\left(\widehat{\theta}_n - \theta^*\right) + R_n - c_n \end{aligned}$$

and c_n is absorbed in the remainder.

- *Example*. **Posterior mode**. In this case, assume we are interested in maximizing the posterior distribution which is proportional to

$$\pi(\theta) \prod_{i=1}^{n} f(x_i | \theta).$$

- Then it can be written as

$$\sum_{i=1}^{n} \frac{\partial \log f(x_i | \theta)}{\partial \theta} = -\frac{\partial \log \pi(\theta)}{\partial \theta}.$$

- It follows that as long as

$$c_n(\theta) = -\frac{\partial \log \pi(\theta)}{\partial \theta}$$

is such that $c_n(\theta) / \sqrt{n} \xrightarrow{\text{P}} 0$ then the Bayesian MAP estimator has the same asymptotic covariance as the MLE.

- Let $\widehat{\theta}_n = (\overline{x}_n, s_n) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 \right)$. This estimate is an M-estimate for

$$\psi(x, \theta) = \begin{pmatrix} \psi_1(x, \theta_1, \theta_2) \\ \psi_2(x, \theta_1, \theta_2) \end{pmatrix} = \begin{pmatrix} x - \theta_1 \\ (x - \theta_1)^2 - \theta_2 \end{pmatrix}.$$

- We can calculate

$$\begin{aligned} A(\theta^*) &= \mathbb{E}\left[-\frac{\partial \psi(X, \theta)}{\partial \theta^{\mathsf{T}}} \right] \Big|_{\theta = \theta^*} \\ &= \mathbb{E} \begin{pmatrix} 1 & 0 \\ 2(x - \theta_1^*) & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \end{aligned}$$

$$B(\theta^*) = \mathbb{E}\left[\psi(X, \theta^*) \psi(X, \theta^*)^{\mathsf{T}} \right]$$

$$= \mathbb{E} \begin{pmatrix} (x - \theta_1^*)^2 & (x - \theta_1^*)\left((x - \theta_1^*)^2 - \theta_2^* \right) \\ (x - \theta_1^*)\left((x - \theta_1^*)^2 - \theta_2^* \right) & \left((x - \theta_1^*)^2 - \theta_2^* \right)^2 \end{pmatrix}$$

$$= \begin{pmatrix} \theta_2^* & \mu_3 \\ \mu_3 & \mu_4 - \theta_2^{*2} \end{pmatrix} = \begin{pmatrix} \theta_2^* & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}$$

where μ_3 is the 3th central moment of X and we have use the more familiar notation $\sigma^2 = \theta_2^*$.

- We can estimate $B\left(\theta^{*}\right)$ by

$$B\left(\widehat{\theta}_{n},\mathbf{x}\right)=\left(\begin{array}{cc} s_{n}^{2} & m_{3} \\ m_{3} & m_{4}-s_{n}^{4} \end{array}\right).$$

- $\widehat{\theta}_{n}$ is a MLE estimate associated to $f\left(x\right|\theta)=\left(2\pi\theta_{2}\right)^{-1/2}\exp\left(-\left(x-\theta_{1}\right)^{2}/2\theta_{2}\right)$ but $\psi_{1}\left(x,\theta_{1},\theta_{2}\right)=x-\theta_{1}$ and $\psi_{2}\left(x,\theta_{1},\theta_{2}\right)=\left(x-\theta_{1}\right)^{2}-\theta_{2}$ are not the score functions which are equal to $\frac{\partial\log f\left(x|\theta\right)}{\partial\theta_{1}}=\left(x-\theta_{1}\right)/\theta_{2}$ and $\frac{\partial\log f\left(x|\theta\right)}{\partial\theta_{2}}=\left(x-\theta_{1}\right)^{2}/2\theta_{2}^{2}-1/2\theta_{2}$.
- It follows that clearly the ψ functions are not unique - many different functions lead to the same estimator. They also yield different $A\left(\theta^{*}\right)$ and $B\left(\theta^{*}\right)$ but the same $V\left(\theta^{*}\right)$.

- If we pick $\psi_{\text{MLE}}(x,\theta) = \frac{\partial \log f(x|\theta)}{\partial\theta}$ then

$$
\begin{aligned}
A(\theta^*) &= \mathbb{E}\left[-\frac{\partial\psi(X,\theta)}{\partial\theta^{\mathsf{T}}}\right]\bigg|_{\theta=\theta^*} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/\sigma^4 \end{pmatrix}, \\
B(\theta^*) &= \mathbb{E}\left[\psi(X,\theta^*)\,\psi(X,\theta^*)^{\mathsf{T}}\right] = \begin{pmatrix} 1/\sigma^2 & \frac{\mu_3}{2\sigma^3} \\ \frac{\mu_3}{2\sigma^3} & \frac{\mu_4-\sigma^4}{4\sigma^8} \end{pmatrix}.
\end{aligned}
$$

- If the data are distributed according to $f(x|\theta)$ then $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$ and it follows that

$$A(\theta^*) = B(\theta^*) = Diag\left(1/\sigma^2, 1/\sigma^4\right).$$

- Note that the likelihood score functions ψ_{MLE} are related to the original ψ by

$$\psi_{\text{MLE}} = C\psi$$

where $C = Diag\left(1/\sigma^2, 1/\sigma^4\right)$. Generally speaking all functions $\psi' = C\psi$ where C is non singular (but possibly dependent on $\theta^*$and $\mathbf{x}$) leads to the same estimator and the same asymptotic matrix.

- *Example* **Ratio Estimator**: Let

$$\widehat{\theta}_n = \frac{\overline{y}}{\overline{x}}$$

where $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $\overline{y} = n^{-1} \sum_{i=1}^{n} y_i$ with $\mathbb{E}(X) = \mu_X$, $\mathbb{E}(Y) = \mu_Y$, $var(X) = \sigma_X^2$, $var(Y) = \sigma_Y^2$ and $cov(X, Y) = \sigma_{XY}$.

- We have

$$\psi(X, Y, \theta) = Y - \theta X$$

thus

$$\begin{aligned} A(\theta^*) &= \mathbb{E}\left[-\frac{\partial\psi(X,\theta)}{\partial\theta^{\mathsf{T}}}\right]\bigg|_{\theta=\theta^*} = \mu_X, \\ B(\theta^*) &= \mathbb{E}\left[\psi(X,\theta^*)^2\right] = \mathbb{E}\left[(Y-\theta X)^2\right], \\ V(\theta^*) &= \mathbb{E}\left[(Y-\theta^* X)^2\right] / \mu_X^2, \end{aligned}$$

- These matrices can be estimated through

$$\begin{aligned} A\left(\widehat{\theta}_n, \mathbf{x}, \mathbf{y}\right) &= \overline{x}, \\ B\left(\widehat{\theta}_n, \mathbf{x}, \mathbf{y}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{\overline{y}}{\overline{x}}x_i\right)^2, \\ V\left(\widehat{\theta}_n, \mathbf{x}, \mathbf{y}\right) &= \frac{1}{n\overline{x}^2}\sum_{i=1}^{n}\left(y_i - \frac{\overline{y}}{\overline{x}}x_i\right)^2. \end{aligned}$$

- If we are interested in the joint distribution of $\left(\overline{x}, \overline{y}, \frac{\overline{y}}{\overline{x}}\right)$, we only need to define

$$\psi(X, Y, \theta) = \left(\begin{array}{l} Y - \theta_1 \\ X - \theta_2 \\ \theta_1 - \theta_3 \theta_2 \end{array} \right).$$

- We obtain

$$A(\theta^*) = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & \theta_3^* & \theta_2^* \end{array} \right), \; B(\theta^*) = \left(\begin{array}{ccc} \sigma_Y^2 & \sigma_{XY} & 0 \\ \sigma_{XY} & \sigma_X^2 & 0 \\ 0 & 0 & 0 \end{array} \right).$$

- We can check that the $(3,3)$th element of $V(\theta^*) = A^{-1}(\theta^*) B(\theta^*) \left\{ A^{-1}(\theta^*) \right\}^{\mathrm{T}}$ is

$$\begin{aligned} v_{33} &= \frac{1}{\theta_2^{*2}} \left[\sigma_Y^2 - 2\theta_3^* \sigma_{XY} + \theta_3^{*2} \sigma_X^2 \right] \\ &= \mathbb{E}\left[(Y - \theta^* X)^2 \right] / \mu_X^2 \end{aligned}$$

- *Example* **Instrumental Variable Estimation**:

$$\begin{aligned} Y_i &= \beta X_i + \sigma_\varepsilon \varepsilon_{1,i} \\ W_i &= X_i + \sigma_U \varepsilon_{2,i} \\ T_i &= \gamma + \delta X_i + \sigma_\tau \varepsilon_{3,i} \end{aligned}$$

where $\varepsilon_{j,i}$ are mutually independent erros with zero mean and unit variance. We also assume that $X_1, \ldots, X_n$ are unobserved, independent of $\{\varepsilon_{j,i}\}$ and have finite variance σ_X^2.

- W_i is a measurement of X_i and T_i is an instrumental variable for X_i (for estimating β) provided that $\delta \neq 0$.
- The OLS estimator of slope obtained by regressing Y on W is

$$\begin{aligned} \widehat{\beta}_{Y|W} &= \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2} = \frac{\sum_{i=1}^n W_i(\beta(W_i - \sigma_U \varepsilon_{2,i}) + \sigma_\varepsilon \varepsilon_{1,i})}{\sum_{i=1}^n W_i^2} \\ &= \beta - \beta \underbrace{\frac{\sigma_U \sum_{i=1}^n W_i \varepsilon_{2,i}}{\sum_{i=1}^n W_i^2}}_{\to \frac{\sigma_U^2}{\sigma_U^2 + \sigma_X^2}} + \underbrace{\frac{\sigma_\varepsilon \sum_{i=1}^n W_i \varepsilon_{1,i}}{\sum_{i=1}^n W_i^2}}_{\to 0} \\ &\xrightarrow{\text{P}} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} \beta. \end{aligned}$$

- For sake of simplicity, lets take here $\gamma = 0$. Let $\widehat{\beta}_{Y|W}$ and $\widehat{\beta}_{W|T}$ be the slopes from the LS regressions of Y on T and W. We have

$$\begin{aligned}
\widehat{\beta}_{Y|T} &= \frac{\sum_{i=1}^{n} Y_i T_i}{\sum_{i=1}^{n} T_i^2} = \frac{\sum_{i=1}^{n}\left(\beta\delta^{-1}(T_i - \sigma_\tau \varepsilon_{3,i}) + \sigma_\varepsilon \varepsilon_{1,i}\right) T_i}{\sum_{i=1}^{n} T_i^2} \\
&= \beta\delta^{-1} - \beta\delta^{-1}\sigma_\tau \underbrace{\frac{\sum_{i=1}^{n} \varepsilon_{3,i} T_i}{\sum_{i=1}^{n} T_i^2}}_{\to \frac{\sigma_\tau^2}{\sigma_\tau^2 + \delta^2 \sigma_X^2}} + \beta\sigma_\varepsilon \underbrace{\frac{\sum_{i=1}^{n} \varepsilon_{1,i} T_i}{\sum_{i=1}^{n} T_i^2}}_{\to 0} \\
&\to \beta\delta^{-1} \frac{\delta^2 \sigma_X^2}{\sigma_\tau^2 + \sigma_X^2}
\end{aligned}$$

$$\begin{aligned}
\widehat{\beta}_{W|T} &= \frac{\sum_{i=1}^{n} W_i T_i}{\sum_{i=1}^{n} T_i^2} = \frac{\sum_{i=1}^{n}\left(\delta^{-1}(T_i - \sigma_\tau \varepsilon_{3,i}) + \sigma_U \varepsilon_{2,i}\right) T_i}{\sum_{i=1}^{n} T_i^2} \\
&= \delta^{-1} - \delta^{-1} \underbrace{\frac{\sum_{i=1}^{n} \sigma_\tau \varepsilon_{3,i} T_i}{\sum_{i=1}^{n} T_i^2}}_{\to \frac{\sigma_\tau^2}{\sigma_\tau^2 + \delta^2 \sigma_X^2}} + \underbrace{\frac{\sigma_U \sum_{i=1}^{n} \varepsilon_{2,i} T_i}{\sum_{i=1}^{n} T_i^2}}_{\to 0} \\
&\to \delta^{-1} \frac{\delta^2 \sigma_X^2}{\sigma_\tau^2 + \sigma_X^2}
\end{aligned}$$

- The intrumental variable estimator is defined by

$$\widehat{\beta}_{\text{IV}} = \frac{\widehat{\beta}_{Y|T}}{\widehat{\beta}_{W|T}} = \frac{\sum_{i=1}^{n} Y_i T_i}{\sum_{i=1}^{n} W_i T_i} \rightarrow \beta.$$

- This estimate is an M-estimator. A choice for ψ consists of using

$$\psi(Y, W, T, \theta) = \begin{pmatrix} \theta_1 - T \\ (Y - \theta_2 W)(\theta_1 - T) \end{pmatrix}.$$

- Indeed

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(\theta_1 - t_i) &= 0 \Rightarrow \widehat{\theta}_1 = \overline{t} = \frac{1}{n}\sum_{i=1}^{n} t_i, \\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_2 w_i)(\theta_1 - t_i) &= 0 \Rightarrow \widehat{\theta}_2 = \frac{\sum_{i=1}^{n} y_i (t_i - \overline{t})}{\sum_{i=1}^{n} w_i (t_i - \overline{t})} \end{aligned}$$

with

$$\widehat{\theta}_1 = \overline{T},\ \widehat{\theta}_2 = \widehat{\beta}_{\text{IV}}.$$

- We obtain

$$A(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_{X,T} \end{pmatrix}, \; B(\theta) = \begin{pmatrix} \sigma_T^2 & 0 \\ 0 & \sigma_T^2 \left(\sigma_\varepsilon^2 + \beta^2 \sigma_U^2\right) \end{pmatrix}$$

- This yields the asymptotic covariance matrix

$$A(\theta)^{-1} B(\theta) \left(A(\theta)^{-1}\right)^{\mathsf{T}} = \begin{pmatrix} \sigma_T^2 & 0 \\ 0 & \sigma_T^2 \left(\sigma_\varepsilon^2 + \beta^2 \sigma_U^2\right) / \sigma_{X,T}^2 \end{pmatrix}$$

- When there is doubt about the magnitude of σ_U^2, then we might want to estimate the joint asymptotic distribution of $\widehat{\beta}_{\mathrm{IV}}$ and $\widehat{\beta}_{Y|W}$.

- *Example*. The sample pth quantile $\widehat{\theta}_n = F_n^{-1}(p)$ satisfies

$$\psi(x, \theta) = p - \mathbb{I}(x \leq \theta)$$

- We have

$$\sum_{i=1}^{n} \psi(x_i, \theta) = c_n = n\left(p - F_n\left(\widehat{\theta}_n\right)\right) \leq 1$$

- This function is discontinuous at θ^* but we can have

$$\begin{aligned} A(\theta^*) &= -\frac{\partial}{\partial \theta^{\mathsf{T}}} \mathbb{E}\left[\psi(X, \theta)\right]\Big|_{\theta=\theta^*} . = -\frac{\partial}{\partial \theta^{\mathsf{T}}} \left[p - F(\theta)\right]\Big|_{\theta=\theta^*} \\ &= f(\theta^*) . \end{aligned}$$

- We also have

$$B(\theta^*) = \mathbb{E}\left[p - \mathbb{I}(X \leq \theta^*)\right]^2 = p(1-p).$$

thus we have

$$V(\theta^*) = \frac{p(1-p)}{f(\theta^*)^2}.$$

- We could also stack any finite number of quantiles ψ functions together to get the joint asymptotic distribution of $\left(F_n^{-1}(p_1), \ldots, F_n^{-1}(p_k)\right)$.
- However we cannot use $A\left(\widehat{\theta}_n, \mathbf{x}\right)$ to estimate $A(\theta^*)$: in fact, the derivative of the pth quantile ψ function is zero everywhere except at the location of the jump discontinuity!
- To estimate f, we can use a kernel density estimator. An alternative consists of approximating ψ by a smooth ψ function.

- *Example*. The positive mean deviation from the median is defined to be

$$\widehat{\theta}_{1,n} = \frac{2}{n} \sum_{i=1}^{n} \left(x_i - \widehat{\theta}_{2,n} \right) \mathbb{I} \left(x_i \geq \widehat{\theta}_{2,n} \right)$$

  where $\widehat{\theta}_{2,n}$ is the sample median.
- The ψ function is

$$\psi(x, \theta) = \begin{pmatrix} 2(x - \theta_2)\,\mathbb{I}(x \geq \theta_2) - \theta_1 \\ \frac{1}{2} - \mathbb{I}(x \leq \theta_2) \end{pmatrix}.$$

- The 1st component of ψ is continuous everywhere but not differentiable at $\theta_2 = x$. The 2nd component has a jump discontinuity at $\theta_2 = x$. To get $A(\theta^*)$, we calculate

$$\mathbb{E}[\psi(X, \theta)] = \begin{pmatrix} 2\int_{\theta_2}^{\infty} (x - \theta_2) f(x)\, dx - \theta_1 \\ \frac{1}{2} - F(\theta_2) \end{pmatrix}.$$

- We write

$$2\int_{\theta_2}^{\infty}(x-\theta_2)\,f(x)\,dx-\theta_1=2\int_{\theta_2}^{\infty}xf(x)\,dx-2\theta_2\left[1-F(\theta_2)\right]-\theta_1.$$

The derivative of this expression with respect to θ_1 is -1, the derivative with θ_2 is

$$-2\theta_2 f(\theta_2)-2\left[1-F(\theta_2)\right]+2\theta_2 f(\theta_2).$$

- It follows that

$$A(\theta^*)=\begin{pmatrix}1 & 1\\ 0 & f(\theta_2^*)\end{pmatrix},\; B(\theta^*)=\begin{pmatrix}b_{11} & \frac{\theta_1^*}{2}\\ \frac{\theta_1^*}{2} & \frac{1}{4}\end{pmatrix}$$

where $b_{11}=4\int_{\theta_2}^{\infty}(x-\theta_2^*)^2 f(x)\,dx-\theta_1^{*2}$.

- Finally we obtain

$$V(\theta^*)=\begin{pmatrix}b_{11}-\frac{\theta_1^*}{f(\theta_2^*)}+\frac{1}{4f(\theta_2^*)^2} & \frac{\theta_1^*}{2f(\theta_2^*)}-\frac{1}{4f(\theta_2^*)^2}\\ \frac{\theta_1^*}{2f(\theta_2^*)}-\frac{1}{4f(\theta_2^*)^2} & \frac{1}{4f(\theta_2^*)^2}\end{pmatrix}.$$