Week 5: Dec. 1, 2022

Note Writer: Yu-Chieh Kuo† and **Collaborators**[1]: **Whiney Yu, Tzu-Yue Huang**⋆

†Department of Information Management, National Taiwan University
⋆Department of Economics, National Taiwan University

Summary of Consistent Estimators

Least Squares: $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2 \xrightarrow{p} Q_\infty(\theta) = \mathbb{E}[y_i - \hat{y}_i]^2$, where $\hat{\theta} \equiv \arg\min Q_n(\theta)$.

Maximum Likelihood: $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(y_i, x_i, \theta) \xrightarrow{p} Q_\infty(\theta) = \mathbb{E}[\log f(y_i, x_i, \theta)]$, where $\hat{\theta} \equiv \arg\max Q_n(\theta)$.

GMM, Minimum Distance Estimators: We have ℓ equations satisfying $\mathbb{E}[g(y_i, x_i, z_i, \theta)] = 0$ such that $\overline{g_n} \equiv \frac{1}{n}\sum_{i=1}^n g_i$ and

$$Q_n(\theta) = \overline{g_n}' \hat{W} \overline{g_n} \xrightarrow{p} Q_\infty(\theta) = \mathbb{E}[g_i]' W \, \mathbb{E}[g_i],$$

where $\hat{\theta} \equiv \arg\min Q_n(\theta)$.

Restricted Estimation

(This section refers to Hansen's textbook, CH8.)

Given $y_i = x_i'\beta + e_i$ and $Ex_ie_i = 0$, we have q linear constraints such that

$$\overbrace{R'}^{q\times k} \overbrace{\beta}^{k\times 1} = \overbrace{C}^{q\times 1}.$$

Note that the constraint is on the population (parameter space).

High-dimensional / regularized estimators

The objective function here might be

$$\min_{\hat{\beta}_i} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^k \hat{\beta}_j^2,$$

where the last term $\lambda \sum_{j=1}^k \hat{\beta}_j$ is the Lagrange multiplier corresponding to $\sum_{j=1}^k \hat{\beta}_j^2 \leq C$. It is called ridge regression.

In addition, the objective function can be also in the form

$$\min_{\hat{\beta}_i} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^k |\hat{\beta}_j|,$$

where the last term $\lambda \sum_{j=1}^k |\hat{\beta}_j|$ is the Lagrange multiplier corresponding to $\sum_{j=1}^k |\hat{\beta}_j| \leq C$. It is called LASSO.

[1]Yu-Chieh thanks their supports to take photo and provide notes.

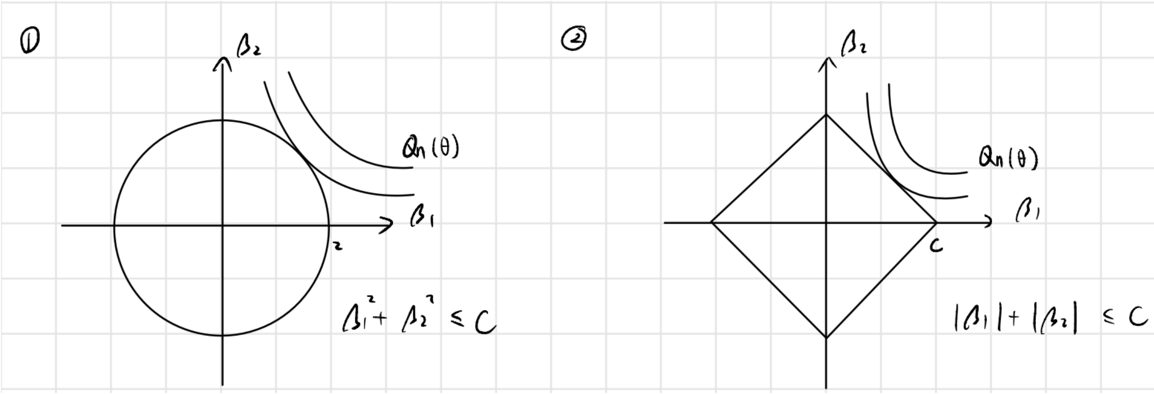

Figure 1: Visualization of the ridge regression and LASSO

Lagrange function

First, we define the sum of squared errors (SSE) as

$$\begin{aligned}
SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 \\
&= (Y - X\beta)'(Y - X\beta) \\
&= Y'Y - 2Y'X\beta + \beta'X'X\beta.
\end{aligned}$$

Be careful about the dimension issues of each matrix: Y is $n \times 1$, X is $n \times k$, and β is $k \times 1$.

Combining SSE, restricted and regularized estimations, we can define the Lagrange function as

$$\mathcal{L} = \frac{1}{2}(Y'Y - 2Y'X\beta + \beta'X'X\beta) + \overbrace{\lambda'(R'\beta - C)}^{1\times g \text{ and } g\times 1},$$

where the fraction $\frac{1}{2}$ in the first term is used to cancel the left coefficient 2 after the derivation, and the second term is the Lagrangem multiplier.

Hence, the first partial derivative of the Lagrange function w.r.t. β and λ is

$$\frac{\partial \mathcal{L}}{\partial \beta} = -X'Y + X'X\tilde{\beta} + R\tilde{\lambda} = 0 \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = R'\tilde{\beta} - C = 0. \tag{2}$$

By solving the system to obtain $\tilde{\beta}$ and $\tilde{\lambda}$, we can use $\tilde{\beta}$ and $\tilde{\lambda}$ to denote the solutions to restricted estimation problem.

To solve the system, we first pre-multiply (1) by $R'(X'X)^{-1}$:

$$\begin{aligned}
& -R' \overbrace{(X'X)^{-1} X'Y}^{\hat{\beta}} + R' (X'X)^{-1} X'X\tilde{\beta} + R' (X'X)^{-1} R\tilde{\lambda} = 0 \\
\iff\ & -R'\hat{\beta} + R'\tilde{\beta} + R' (X'X)^{-1} R\tilde{\lambda} = 0 \\
\iff\ & R'\tilde{\beta} = R'\hat{\beta} + R' (X'X)^{-1} R\tilde{\lambda}.
\end{aligned}$$

Next, we substitute $R'\hat{\beta} + R'(X'X)^{-1}R\tilde{\lambda}$ for $R'\tilde{\beta}$ in (2) to solve $\tilde{\lambda}$:

$$\begin{aligned} & R'\tilde{\beta} = C \\ \iff & R'\hat{\beta} + R'(X'X)^{-1}R\tilde{\lambda} = C \\ \iff & \tilde{\lambda} = \left(R'(X'X)^{-1}R\right)^{-1}\left(R'\hat{\beta} - C\right). \end{aligned}$$

Lastly, we substitute $\left(R'(X'X)^{-1}R\right)^{-1}\left(R'\hat{\beta} - C\right)$ for $\tilde{\lambda}$ in (1) to solve $\tilde{\beta}$:

$$\begin{aligned} & -X'Y + X'X\tilde{\beta} + R\left(R'(X'X)^{-1}R\right)^{-1}\left(R'\hat{\beta} - C\right) = 0 \\ \iff & \tilde{\beta} = (X'X)^{-1}X'Y - (X'X)^{-1}R\left(R'(X'X)^{-1}R\right)^{-1}\left(R'\hat{\beta} - C\right) \\ \iff & \tilde{\beta} = \hat{\beta} - R\left(R'(X'X)^{-1}R\right)^{-1}\left(R'\hat{\beta} - C\right). \end{aligned}$$

Note that R is $k \times q$.

Remark.

1. If $R'\hat{\beta} - C = 0$, then $\tilde{\beta} = \hat{\beta}$.
2. $R'(X'X)^{-1}R$ is invertible only if $rank(R) = q$.

□

Consistency

Now we discuss the consistency of the restricted estimation. If it is given $R'B = C$ and $\hat{\beta} \xrightarrow{p} \beta_0$ (true β), then we have

$$R'\hat{\beta} - C \xrightarrow{p} 0 \text{ and } \hat{\beta} \xrightarrow{p} \beta_0 \quad \Longrightarrow \quad \tilde{\beta} \xrightarrow{p} \beta.$$

Asymptotic normality

$$\begin{aligned} \sqrt{n}\left(\tilde{\beta} - \beta\right) &= \overbrace{\sqrt{n}\left(\hat{\beta} - \beta\right)}^{\xrightarrow{d} \mathcal{N}\left(0, \overbrace{\left(Ex_ix_i'\right)^{-1}}^{\equiv Q_{XX}^{-1}} \overbrace{\mathbb{E}\left[x_ix_i'e_i^2\right]}^{\equiv\Omega} \overbrace{\left(Ex_ix_i'\right)^{-1}}^{\equiv Q_{XX}^{-1}}\right)} \overbrace{- (X'X)^{-1}R\left(R'(X'X)^{-1}R\right)^{-1}}^{\equiv \hat{M} \xrightarrow{p} M} \overbrace{\sqrt{n}\left(R'\hat{\beta} - C\right)}^{R'\left(\sqrt{n}\left(\hat{\beta}-\beta\right)\right) \text{ since } C = R'\beta} \\ &\xrightarrow{d} \mathcal{N}(0, Cov), \end{aligned}$$

where Cov is derived by $\sqrt{n}\left(\tilde{\beta} - \beta\right)\sqrt{n}\left(\tilde{\beta} - \beta\right)'$. Clearly,

$$\begin{aligned} Cov &= \sqrt{n}\left(\tilde{\beta} - \beta\right)\sqrt{n}\left(\tilde{\beta} - \beta\right)' \\ &= n\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)' - n\hat{M}\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)' - n\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\hat{M}' + n\hat{M}\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\hat{M}' \\ &\xrightarrow{p} V_\beta - MV_\beta - V_\beta M' + MV_\beta M' \\ &= V_\beta - Q_{XX}^{-1}R\left(R'Q_{XX}^{-1}R\right)^{-1}R'V_\beta - V_\beta R\left(R'Q_{XX}^{-1}R\right)^{-1}R'Q_{XX}^{-1} + Q_{XX}^{-1}R\left(R'Q_{XX}^{-1}R\right)^{-1}R'V_\beta. \end{aligned}$$

Can we do better?

The answer is yes. We may set up a minimun distance estimation as

$$\begin{aligned} \min_{\beta} \quad & \mathcal{J}(\beta) = n\left(\hat{\beta}-\beta\right)' \hat{W}\left(\hat{\beta}-\beta\right) \\ \text{s.t.} \quad & \text{Constraints,} \end{aligned}$$

where $\hat{\beta}$ is an OLS estimator (treated as given). Note that β here is a choive variable, not the true parameter.

Remark. The Constrained Least Squares (CLS) is a special case where $\hat{W} = Q_{XX}$. □

Now, consider the SSE (what is β below. choice variable or true para?)

$$\begin{aligned} SSE &= \sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2 \\ &= \sum_{i=1}^{n}\left(x_i'\hat{\beta} + \hat{e}_i - x_i'\beta\right)^2 \\ &= \sum_{i=1}^{n}\left(\hat{e}_i + x_i'\left(\hat{\beta}-\beta\right)\right)^2 \\ &= \sum_{i=1}^{n}\hat{e}_i^2 + \left(\hat{\beta}-\beta\right)'\left(\sum_{i=1}^{n}x_ix_i'\right)\left(\hat{\beta}-\beta\right) + 2\cancelto{0}{\sum_{i=1}^{n}\hat{e}_i x_i\left(\hat{\beta}-\beta\right)} \\ &= \sum_{i=1}^{n}\hat{e}_i^2 + \left(\hat{\beta}-\beta\right)'\left(\sum_{i=1}^{n}x_ix_i'\right)\left(\hat{\beta}-\beta\right), \end{aligned}$$

where we define $\mathcal{J}(\beta)$ as the last term $\left(\hat{\beta}-\beta\right)'\left(\sum_{i=1}^{n}x_ix_i'\right)\left(\hat{\beta}-\beta\right)$ with $\hat{W} = \sum_{i=1}^{n}x_ix_i'$.

After obtaining $\mathcal{J}(\beta)$, we want to conduct the minimun distance estimation. That is, we solve the system

$$\begin{aligned} \min_{\beta} \quad & \mathcal{J}(\beta) \\ \text{s.t.} \quad & R'\beta = C \quad (\text{Note that } R'\beta_0 = C). \end{aligned}$$

The corresponding Lagrange function is

$$\begin{aligned} \mathcal{L} &= \frac{1}{2}\mathcal{J}\left(\beta, \hat{W}\right) + \lambda'(R'\beta - C) \\ &= \frac{n}{2}\left(\hat{\beta}-\beta\right)'\hat{W}\left(\hat{\beta}-\beta\right) + \lambda'(R'\beta - C), \end{aligned}$$

and FOC w.r.t. β and λ yeilds

$$\frac{\partial \mathcal{L}}{\partial \beta} = -n\hat{W}\left(\hat{\beta}-\tilde{\beta}\right) + R\tilde{\lambda} = 0 \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = R'\tilde{\beta} - C = 0. \tag{4}$$

Extending (3) solves $\tilde{\beta}$:

$$\tilde{\beta} = \hat{\beta} - \frac{1}{n}\hat{W}^{-1}R\tilde{\lambda}.$$

Substituting () for (4) gives

$$R'\left(\hat{\beta} - \frac{1}{n}\hat{W}^{-1}R\tilde{\lambda}\right) - C = 0 \quad \Longleftrightarrow \quad \tilde{\lambda} = n\left(R'\hat{W}^{-1}R\right)^{-1}\left(R'\hat{\beta} - C\right).$$

Lastly, we use $\tilde{\lambda}$ to solve $\tilde{\beta}$ in (3):

$$\begin{aligned} & -n\hat{W}\left(\hat{\beta}-\tilde{\beta}\right)+nR\left(R'\hat{W}^{-1}R\right)^{-1}\left(R'\hat{\beta}-C\right)=0 \\ \iff & \tilde{\beta}=\hat{\beta}-\hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}\left(R'\hat{\beta}-C\right). \end{aligned}$$

Consistency

Given $\hat{\beta}\xrightarrow{p}\beta_0$ (true parameter) and $R'\hat{\beta}-C\xrightarrow{p}0$, we obtain $\tilde{\beta}\xrightarrow{p}\beta_0$.

Asymptotic normality

$$\begin{aligned} \sqrt{n}\left(\tilde{\beta}-\beta_0\right) &= \sqrt{n}\left(\hat{\beta}-\beta_0\right)-\hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}R'\overbrace{\sqrt{n}\left(\hat{\beta}-\beta_0\right)}^{\xrightarrow{d}\mathcal{N}(0,V_\beta)} \\ &\xrightarrow{d} \mathcal{N}\left(0,Cov\right), \end{aligned}$$

where Cov is

$$\begin{aligned} Cov &= V_\beta-W^{-1}R\left(R'W^{-1}R\right)^{-1}R'V_\beta-V_\beta R\left(R'W^{-1}R\right)^{-1}R'W^{-1} \\ &\quad +W^{-1}R\left(R'W^{-1}R\right)^{-1}R'V_\beta R\left(R'W^{-1}R\right)^{-1}R'W^{-1}. \end{aligned}$$

It shows that the most efficient choice of W is V_β^{-1}. Therefore, the covariance matrix alters to

$$Cov=V_\beta-V_\beta R\left(R'V_\beta^{-1}R\right)^{-1}R'V_\beta.$$

In general, $\tilde{\beta}_{MD}$ (minimun distance) is more efficient than $\tilde{\beta}_{CLS}$.

Short summary

CLS: We solve $\min_\beta\sum_{i=1}^n\left(y_i-x_i'\beta\right)^2$ s.t. $R'\beta=C \implies \tilde{\beta}_{CLS}$.

MD Estimation: We solve $\min_\beta\left(\hat{\beta}-\beta\right)'\hat{W}\left(\hat{\beta}-\beta\right)$ s.t. $R'\beta=C \implies \tilde{\beta}_{MD}$.

Note that CLS is a special case where

$$\hat{W}=\frac{1}{n}\sum_{i=1}^n x_ix_i'\xrightarrow{p}W=\mathbb{E}\left[x_ix_i'\right],$$

but the efficient weight matrix is

$$\hat{W}=\left(\frac{1}{n}\sum_{i=1}^n x_ix_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_ix_i'\hat{e}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n x_ix_i'\right)^{-1}\xrightarrow{p}W=\mathbb{E}\left[x_ix_i'\right]^{-1}\mathbb{E}\left[x_ix_i'e_i^2\right]\mathbb{E}\left[x_ix_i'\right]^{-1}=V_\beta^{-1}.$$

Consequently, $\tilde{\beta}_{MD}$ is more efficient than $\tilde{\beta}_{CLS}$.

Example. Given a regression $y_i=x_{i1}'\beta_1+x_{i2}'\beta_2+e_i$ with a constraint $\beta_2=0$, we can show that the estimator from regression without x_{2i} is identical with the CLS estimator with $\beta_2=0$.

Another example can be found at Page. 269 in Hansen's textbook. □

Misspecification

(This section refers to Hansen's textbook, CH8.13.)

In the case that $R'\beta = C^\star \neq C$, the MD estimator alters to

$$\tilde{\beta}_{MD} = \hat{\beta} - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}\left(R'\hat{\beta} - C\right) \xrightarrow{p} \beta - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}(C^\star - C) \equiv \beta_n^\star.$$

The asymptotic normality becomes

$$\begin{aligned}
\sqrt{n}\left(\tilde{\beta}_{MD} - \beta_n^\star\right) &= \sqrt{n}\left(\hat{\beta} - \beta\right) - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}\sqrt{n}\left(R'\hat{\beta} - C^\star\right) \\
&= \sqrt{n}\left(\hat{\beta} - \beta\right) - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}\sqrt{n}\left(R'\hat{\beta} - R'\beta\right) \\
&= \left(I - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}R\right)\sqrt{n}\left(\hat{\beta} - \beta\right) \\
&\xrightarrow{d} \mathcal{N}\left(0, V_\beta(W)\right),
\end{aligned}$$

where $V_\beta(W)$ is the same asymptotic covariance in the case without misspecification. **why??????**

Another case for the misspecification issue might be in the form of $R'\beta_n = C + \delta\sqrt{n}$. In this case, $R'\hat{\beta} - C = R'\left(\hat{\beta} - \beta_n\right) + \delta\sqrt{n}$, and the MD estimator is

$$\begin{aligned}
\tilde{\beta}_{MD} &= \hat{\beta} - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}\left(R'\hat{\beta} - C\right) \\
&= \hat{\beta} - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}R'\left(\hat{\beta} - \beta_n\right) - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}R'\delta\sqrt{n}.
\end{aligned}$$

The asymptotic normality in this case becomes

$$\begin{aligned}
\sqrt{n}\left(\tilde{\beta}_{MD} - \beta_n\right) &= \overbrace{\sqrt{n}\left(\hat{\beta} - \beta_n\right)}^{\xrightarrow{d}\mathcal{N}(0,V_\beta)} - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}R'\overbrace{\sqrt{n}\left(\hat{\beta} - \beta_n\right)}^{\xrightarrow{d}\mathcal{N}(0,V_\beta)} - \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}R'\delta \\
&\xrightarrow{d} \mathcal{N}\left(0, V_\beta(W)\right) \overbrace{- \hat{W}^{-1}R\left(R'\hat{W}^{-1}R\right)^{-1}R'\delta}^{\equiv \delta^\star} \\
&= \mathcal{N}\left(\delta^\star, V_\beta(W)\right).
\end{aligned}$$