Identifiability & Asymptotics (Lab 7)

BST 235: Advanced Regression and Statistical Learning

Alex Levis, Fall 2019

1 Identifiability in statistical models

In statistical inference, the concept of *identifiability* is fundamental. If we wish to be able to estimate a parameter from data, it has to be the case that different parameters induce different observed data distributions.

Consider for instance, the parametric models

(a) $P_\mu \in \mathcal{F} = \{P_\mu \,|\, \mu \in \mathbb{R}\}$, with $P_\mu = \mathcal{N}(\mu, 1)$.

(b) $P_{\boldsymbol{\theta}} \in \mathcal{F} = \{P_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in \mathbb{R}^2\}$, with $P_{\boldsymbol{\theta}} = \mathcal{N}(\theta_1 + \theta_2, 1)$, where $\boldsymbol{\theta} = [\theta_1 \;\; \theta_2]^T$.

Since $P_\mu = P_{\mu'} \implies \mu = \mu'$, we would say μ is identifiable. On the other hand, $P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta'}}$, where $\boldsymbol{\theta} = (1, 0)$ and $\boldsymbol{\theta'} = (0, 1)$ — we cannot distinguish between these two parameters, as they induce the same distribution. In this case, we would say $\boldsymbol{\theta}$ is non-identifiable.

In the abstract case, suppose we observe data $\mathcal{S} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n) \sim P_{\boldsymbol{\theta}}$, with parameter $\boldsymbol{\theta} \in \Theta$ indexing the statistical model $\mathcal{F} = \{P_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in \Theta\}$. The model, or equivalently the parameter $\boldsymbol{\theta}$, is called *identifiable* if

$$P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta'}} \implies \boldsymbol{\theta} = \boldsymbol{\theta'}.$$

More generally, a function g with domain Θ is called *identifiable* if

$$P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta'}} \implies g(\boldsymbol{\theta}) = g(\boldsymbol{\theta'}).$$

Equivalently, $g(\boldsymbol{\theta}) \neq g(\boldsymbol{\theta'}) \implies P_{\boldsymbol{\theta}} \neq P_{\boldsymbol{\theta'}}$. In some cases, the full model is not identifiable, but the model can be thought of as "partially identifiable" if there are some functions g of $\boldsymbol{\theta}$ that are identifiable, i.e., there are some functions of the parameters which we can hope to estimate.

2 Identifiability in the linear model

The homoscedastic Gaussian linear model (i.e., assumptions (A), (B), (C), (D) from the notes) posits that

$$\mathbf{Y} \,|\, \mathbb{X} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2 I_n),$$

for some $\boldsymbol{\beta} \in \mathbb{R}^d$, $\sigma^2 \geq 0$. Let $\boldsymbol{\theta} = [\boldsymbol{\beta}^T \;\; \sigma^2]^T \in \mathbb{R}^d \times [0, \infty) =: \Theta$ denote the full set of parameters. Then $\boldsymbol{\theta} \in \Theta$ indexes the parametric model $\mathcal{F} = \{P^{(n)}_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in \Theta\}$ for the conditional distribution of $\mathbf{Y}$ given $\mathbb{X}$, where $P^{(n)}_{\boldsymbol{\theta}} = \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2 I_n)$. The next exercise shows that the function $g(\boldsymbol{\theta}) = \boldsymbol{\beta}$ is not identifiable when $\mathbb{X}$ is not full column rank.

Exercise 1. In the homoscedastic Gaussian linear model, show that $\boldsymbol{\beta}$ is not identifiable from the distribution of $\mathbf{Y}$ given $\mathbb{X}$ when $\text{rank}(\mathbb{X}) < d$.

Since $\text{rank}(\mathbb{X}) < d$, we know by rank-nullity that $\dim(\mathcal{N}(\mathbb{X})) \geq 1$, so take $\mathbf{v} \neq 0$ such that $\mathbb{X}\mathbf{v} = \mathbf{0}_n$. Then fixing $\sigma^2 \in [0, \infty)$, let $\boldsymbol{\theta}_1 = [\boldsymbol{\beta}^T \ \ \sigma^2]^T$, $\boldsymbol{\theta}_2 = [(\boldsymbol{\beta} + \mathbf{v})^T \ \ \sigma^2]^T$, and see that

$$P_{\boldsymbol{\theta}_1}^{(n)} = \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2 I_n) = \mathcal{N}_n(\mathbb{X}(\boldsymbol{\beta} + \mathbf{v}), \sigma^2 I_n) = P_{\boldsymbol{\theta}_2}^{(n)},$$

but $\boldsymbol{\beta} \neq \boldsymbol{\beta} + \mathbf{v}$. Thus $\boldsymbol{\beta}$ is not identifiable.

Lemma 1. In the homoscedastic Gaussian linear model, a function h of $\boldsymbol{\beta}$ (think $g(\boldsymbol{\theta}) = h(\boldsymbol{\beta})$) is identifiable if and only if for some ϕ with domain $\mathcal{C}(\mathbb{X})$, $h(\boldsymbol{\beta}) = \phi(\mathbb{X}\boldsymbol{\beta})$, for all $\boldsymbol{\beta} \in \mathbb{R}^d$.

Proof. We first show that $g(\boldsymbol{\theta}) = \mathbb{X}\boldsymbol{\beta}$ is identifiable, which will imply $h(\boldsymbol{\beta}) = \phi(\mathbb{X}\boldsymbol{\beta})$ is identifiable, for any given ϕ. But the former fact is immediate, since for $\boldsymbol{\theta}_1 = [\boldsymbol{\beta}_1^T \ \ \sigma_1^2]^T, \boldsymbol{\theta}_2 = [\boldsymbol{\beta}_2^T \ \ \sigma_2^2]^T \in \Theta$, then $P_{\boldsymbol{\theta}_1}^{(n)} = P_{\boldsymbol{\theta}_2}^{(n)}$ implies that the means of these conditional distributions are equal, i.e., $\mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2$.

Conversely, suppose there does not exist a function ϕ with domain $\mathcal{C}(\mathbb{X})$ satisfying $h(\boldsymbol{\beta}) = \phi(\mathbb{X}\boldsymbol{\beta})$, for all $\boldsymbol{\beta} \in \mathbb{R}^d$. Then there must exist $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d$ with $\mathbb{X}\boldsymbol{\beta}_1 = \mathbb{X}\boldsymbol{\beta}_2$, but $h(\boldsymbol{\beta}_1) \neq h(\boldsymbol{\beta}_2)$. Fixing $\sigma^2 \in [0, \infty)$, we can take $\boldsymbol{\theta}_1 = [\boldsymbol{\beta}_1^T \ \ \sigma^2]^T$, $\boldsymbol{\theta}_2 = [\boldsymbol{\beta}_2^T \ \ \sigma^2]^T$, and see that

$$P_{\boldsymbol{\theta}_1}^{(n)} = \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}_1, \sigma^2 I_n) = \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}_2, \sigma^2 I_n) = P_{\boldsymbol{\theta}_2}^{(n)},$$

yet $h(\boldsymbol{\beta}_1) \neq h(\boldsymbol{\beta}_2)$. Therefore, $h(\boldsymbol{\beta})$ is not identifiable. □

The following result and exercise will help us understand identifiability of linear functions of the regression parameters.

Lemma 2. Let $B \in \mathbb{R}^{k \times d}$ and $C \in \mathbb{R}^{\ell \times d}$. Then

$$\mathcal{R}(B) \subseteq \mathcal{R}(C) \iff \exists M \in \mathbb{R}^{k \times \ell} \text{ such that } B = MC.$$

Proof. If $\exists M \in \mathbb{R}^{k \times \ell}$ such that $B = MC$, then

$$\mathcal{R}(B) = \mathcal{R}(MC) = \mathcal{C}(C^T M^T) \subseteq \mathcal{C}(C^T) = \mathcal{R}(C),$$

where the inclusion follows trivially from the definition of the column space. Conversely, suppose $\mathcal{R}(B) \subseteq \mathcal{R}(C)$, or equivalently that for all $\mathbf{v} \in \mathbb{R}^k$, there exists $\mathbf{u} \in \mathbb{R}^\ell$ such that $B^T\mathbf{v} = C^T\mathbf{u}$. We can find $\mathbf{u}_i \in \mathbb{R}^\ell$ such that $C^T\mathbf{u}_i = B^T e_i^{(k)} = B_{i\cdot}$, where $B_{i\cdot}$ denotes the i-th row of B, for $i = 1, \ldots, k$, so let $M = [\mathbf{u}_1 \cdots \mathbf{u}_k]^T$. Then, observe that

$$B = \begin{bmatrix} B_{1\cdot}^T \\ \vdots \\ B_{k\cdot}^T \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1^T C \\ \vdots \\ \mathbf{u}_k^T C \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_k^T \end{bmatrix} C = MC,$$

concluding the argument. □

A linear function of the regression parameters, $A\boldsymbol{\beta}$, is called *estimable* if $g(\boldsymbol{\theta}) = A\boldsymbol{\beta}$ is identifiable.

Exercise 2. Show that $A\boldsymbol{\beta}$ is estimable iff $\mathcal{R}(A) \subseteq \mathcal{R}(\mathbb{X})$.

Suppose $A \in \mathbb{R}^{q\times d}$. The result is obtained from the following sequence of equivalences:

$$\begin{aligned} A\boldsymbol{\beta} \text{ is estimable} &\iff \exists \phi : \mathcal{C}(\mathbb{X}) \to \mathbb{R}^q \text{ such that for all } \boldsymbol{\beta} \in \mathbb{R}^d, A\boldsymbol{\beta} = \phi(\mathbb{X}\boldsymbol{\beta}), \text{ by Lemma 1,} \\ &\iff \exists M \in \mathbb{R}^{q\times n} \text{ such that } A = M\mathbb{X}, \text{ by linearity of matrix multiplication,} \\ &\iff \mathcal{R}(A) \subseteq \mathcal{R}(\mathbb{X}), \text{ by Lemma 2.} \end{aligned}$$

Although seemingly simple, the equivalence invoking linearity really requires additional care in the 'only if' direction. To elaborate, suppose $\phi : \mathcal{C}(\mathbb{X}) \to \mathbb{R}^q$ satisfies $A\boldsymbol{\beta} = \phi(\mathbb{X}\boldsymbol{\beta})$, for all $\boldsymbol{\beta} \in \mathbb{R}^d$, so that $T_A = \phi \circ T_{\mathbb{X}}$, where T_B is the linear transformation associated with the matrix B. Since $T_{\mathbb{X}}$ is surjective on its image $\mathcal{C}(\mathbb{X})$, Fact 1 below implies that $\phi \in \mathcal{L}(\mathcal{C}(\mathbb{X}), \mathbb{R}^q)$. By Fact 2, we can extend ϕ to $\phi^* \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^q)$, satisfying $\phi^*(\mathbf{w}) = \phi(\mathbf{w})$, for all $\mathbf{w} \in \mathcal{C}(\mathbb{X})$. By material from Lab 2, there must then be a (unique) matrix $M \in \mathbb{R}^{q\times n}$ such that $M\mathbf{w} = \phi^*(\mathbf{w})$, for all $\mathbf{w} \in \mathbb{R}^n$. Therefore,

$$M\mathbb{X}\boldsymbol{\beta} = \phi^*(\mathbb{X}\boldsymbol{\beta}) = \phi(\mathbb{X}\boldsymbol{\beta}) = A\boldsymbol{\beta},$$

for all $\boldsymbol{\beta} \in \mathbb{R}^d$, so $M\mathbb{X} = A$, as claimed.

Fact 1: Suppose for vector spaces U, V, W, $T \in \mathcal{L}(U, W)$ is a linear map, $S \in \mathcal{L}(U, V)$ is surjective, and $T = f \circ S$ for some $f : V \to W$. Then $f \in \mathcal{L}(V, W)$.

Proof. If f is not linear, then $\exists\, \alpha, \beta \in \mathbb{F}$ and $v_1, v_2 \in V$ such that $f(\alpha v_1 + \beta v_2) \neq \alpha f(v_1) + \beta f(v_2)$. By surjectivity, $\exists\, u_1, u_2 \in U$ such that $v_1 = S(u_1)$, $v_2 = S(u_2)$. By linearity of S,

$$T(\alpha u_1 + \beta u_2) = f(S(\alpha u_1 + \beta u_2)) = f(\alpha v_1 + \beta v_2) \neq \alpha f(v_1) + \beta f(v_2) = \alpha T(u_1) + \beta T(u_2),$$

contradicting linearity of T. Hence f is linear. □

Fact 2: Suppose V is a finite-dimensional vector space, and $U \subseteq V$ is a linear subspace. If W is another vector space and $T \in \mathcal{L}(U, W)$, then there exists $T^* \in \mathcal{L}(V, W)$ satisfying $T^*(v) = T(v)$, for all $v \in U$, i.e., T^* extends T from U to all of V.

Proof. See here for a discussion when $W = \mathbb{R}$ — the same ideas apply here. Extend a basis of U to a basis for V, and define T^* on the basis to equal T for any basis vector for U, and 0_W for any basis vector for $V \setminus U$. With this, define T^* to be the linear combination of T^* applied to the basis representation of its input. □

We conclude this section with a statement of results, analogous to what we have seen in the full rank setting, that apply even when $\mathbb{X}$ is not full rank. The proofs involve working with the SVD of $\mathbb{X}$, the pseudoinverse $(\mathbb{X}^T\mathbb{X})^-$, and some of this may be left for the next homework.

Lemma 3. Suppose that $r = \text{rank}(\mathbb{X}) \leq d$, and let $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^-\mathbb{X}^T\mathbf{Y}$, the least squares estimator.

(i) Assume the homoscedastic linear model (i.e., assumptions (A) - (C)), and suppose $\mathbf{c} \in \mathcal{R}(\mathbb{X})$. Then $\mathbf{c}^T\widehat{\boldsymbol{\beta}}$ is the BLUE of $\mathbf{c}^T\boldsymbol{\beta}$. Under normality (i.e., if additionally (D) holds), we also have $\mathbf{c}^T\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{c}^T\boldsymbol{\beta}, \sigma^2\mathbf{c}^T(\mathbb{X}^T\mathbb{X})^-\mathbf{c})$.

(ii) Let $\widehat{\sigma}^2 = \frac{1}{n}\|(I_n - \widehat{P}_{\mathbb{X}})\mathbf{Y}\|^2$. Under assumptions (A), (B), (C), (D), $\frac{n\widehat{\sigma}^2}{n-r} \,|\, \mathbb{X} \sim \chi^2_{n-r}(0)$.

(iii) Assuming (A), (B), (C), (D), $\mathbf{c}^T\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\sigma}^2 \,|\, \mathbb{X}$, and $\frac{\mathbf{c}^T\widehat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{\sqrt{\frac{n\widehat{\sigma}^2}{n-r}\mathbf{c}^T(\mathbb{X}^T\mathbb{X})^-\mathbf{c}}} \,|\, \mathbb{X} \sim t_{n-r}(0)$.

3 Review of asymptotics in probability theory

We review here some definitions and major results in large sample theory, that will be helpful in studying the random design setting, among other areas.

Suppose $(\mathbf{X}_n)_{n=1}^{\infty}$, with $\mathbf{X}_n = [X_{n1} \cdots X_{nm}]^T$, is a sequence of random vectors, on the probability space $(\Omega, \mathcal{A}, P)$, and let $\mathbf{X} = [X_1 \cdots X_m]^T$ be another random m-vector on this space.

(1) If $\mathbf{X}_n$ converges pointwise to $\mathbf{X}$, in the sense that $\mathbf{X}_n(\omega) \to \mathbf{X}(\omega)$ as $n \to \infty$, for all $\omega \in \Omega$, then we say $\mathbf{X}_n$ *converges surely to* $\mathbf{X}$.

(2) If $\exists\, N \in \mathcal{A}$ with $P(N) = 0$, such that $\mathbf{X}_n(\omega) \to \mathbf{X}(\omega)$ as $n \to \infty$, for all $\omega \in \Omega \setminus N$, meaning

$$P\left[\lim_{n\to\infty} \mathbf{X}_n = \mathbf{X}\right] = 1,$$

then we say $\mathbf{X}_n$ *converges almost surely to* $\mathbf{X}$ — we write $\mathbf{X}_n \overset{\text{a.s.}}{\to} \mathbf{X}$.

(3) If $\forall \epsilon, \delta > 0$, $\exists\, n_{\epsilon,\delta} \in \mathbb{N}$ such that

$$P\left[\|\mathbf{X}_n - \mathbf{X}\| > \epsilon\right] \leq \delta, \text{ for all } n \geq n_{\epsilon,\delta},$$

then we say $\mathbf{X}_n$ *converges in probability to* $\mathbf{X}$. Equivalently, for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\left[\|\mathbf{X}_n - \mathbf{X}\| > \epsilon\right] = 0 \iff \lim_{n\to\infty} P\left[\|\mathbf{X}_n - \mathbf{X}\| \leq \epsilon\right] = 1.$$

In this case, we write $\mathbf{X}_n \overset{P}{\to} \mathbf{X}$.

(4) Let $F_n(\mathbf{x}) = P[X_{n1} \leq x_1, \ldots, X_{nm} \leq x_m]$ and $F(\mathbf{x}) = P[X_1 \leq x_1, \ldots, X_m \leq x_m]$. If

$$F_n(\mathbf{x}) \to F(\mathbf{x}), \text{ as } n \to \infty,$$

for all $\mathbf{x} \in \mathbb{R}^m$ where F is continuous, then we say $\mathbf{X}_n$ *converges in distribution to* $\mathbf{X}$ — we write $\mathbf{X}_n \overset{\mathcal{D}}{\to} \mathbf{X}$.

Convergence of types (1) or (2) is often far stronger than will be possible in practice, so we typically work with the weaker types of convergence, (3) and (4). Recall the hierarchy

$$(1) \implies (2) \implies (3) \implies (4),$$

and (3) $\iff$ (4) when $\mathbf{X}$ is constant. Note also that for convergence almost surely and in probability, $\mathbf{X}_n$ converges to $\mathbf{X}$ if and only if X_{nj} converges to X_j for $j = 1, \ldots, m$. The same does **not** hold for convergence in distribution (cf. Cramer-Wold device). In addition to the famous results below, recall the continuous mapping theorems, Slutsky's theorem, and the many uses of characteristic functions.

Weak law of large numbers: Suppose $(\mathbf{X}_n)_{n=1}^{\infty}$ is an iid sequence of random vectors, with $\mathbb{E}(\mathbf{X}_1) = \boldsymbol{\mu}$. Letting $\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$,

$$\overline{\mathbf{X}}_n \overset{P}{\to} \boldsymbol{\mu}.$$

Central limit theorem: Suppose $(\mathbf{X}_n)_{n=1}^{\infty}$ is an iid sequence of random vectors, with $\mathbb{E}(\mathbf{X}_1) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{X}_1) = \Sigma$. Then

$$\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \overset{\mathcal{D}}{\to} \mathcal{N}_m(\mathbf{0}_m, \Sigma).$$

4 Order in probability notation (if we have time)

Succinct notation can often help simplify derivations and clarify complex ideas. Stochastic order notation, in particular the use of $o_p(\cdot)$ and $O_p(\cdot)$ is one useful shorthand in many calculations. We briefly review definitions and basic properties here.

- Convergence in probability: if $\mathbf{X}_n \xrightarrow{P} \mathbf{0}_m$, then we write

  $$\mathbf{X}_n = o_p(1).$$

  In greater generality, for a given sequence of random variables $U_n \neq 0$, we write

  $$\mathbf{X}_n = o_p(U_n) \iff \frac{\mathbf{X}_n}{U_n} = o_p(1).$$

- Stochastic boundedness: we say the sequence $\mathbf{X}_n$ is stochastically bounded if for any $\epsilon > 0$, there exists $M > 0$ and $N \in \mathbb{N}$ such that

  $$P[\|\mathbf{X}_n\| > M] < \epsilon, \forall n \geq N.$$

  In this case, we write $\mathbf{X}_n = O_p(1)$, and for a sequence $U_n \neq 0$,

  $$\mathbf{X}_n = O_p(U_n) \iff \frac{\mathbf{X}_n}{U_n} = O_p(1).$$

As with usual $o(\cdot)$ and $O(\cdot)$ notation, the equality should be understood as the function on the left *belonging to* the class $o_p(a_n)$ or $O_p(a_n)$ — these can be seen as the set of all random sequences on this probability space that satisfy the definition. Note the following basic properties:

- $o_p(1) + o_p(1) = o_p(1)$
- $o_p(1) + O_p(1) = O_p(1)$
- $o_p(1)O_p(1) = o_p(1)$
- $U_n o_p(1) = o_p(U_n)$
- $U_n O_p(1) = O_p(U_n)$
- $o_p(O_p(1)) = o_p(1)$
- If $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$, then $\mathbf{X}_n = O_p(1)$.

If you are interested, see Chapter 2.2 of Asymptotic Statistics (van der Vaart, 1998) for more.