Stats305a Étude 4

Due: Wednesday, December 14 at 5:00pm on Gradescope.

Question 4.1: The *sandwich estimator of variance* of an estimator generalizes typical variance estimates for linear regression models to work in situations with heteroskedastic noise, mis-specified models, and losses beyond the squared error. You will compare this sandwich estimate of variance, the bootstrap estimators, and the standard covariance estimates in linear models.

We we wish to estimate a parameter β in the (population) regression problem

$$\beta^\star = \operatorname*{argmin}_\beta \left\{ L(\beta) := \mathbb{E}[(x^T\beta - y)^2] \right\}$$

where (x, y) have some joint distribution (recall Question 2.1 in the exercises). Consider the empirical estimator

$$\widehat{\beta}_n = \operatorname*{argmin}_\beta \frac{1}{n} \sum_{i=1}^n (x_i^T\beta - y)^2$$

We wish to establish confidence intervals for $\beta^\star$ using methods that find covariance estimates $\widehat{\Sigma}$ so that we have the approximate distributional identity

$$\widehat{\beta}_n - \beta^\star \stackrel{\text{apx}}{\sim} \mathsf{N}(0, \widehat{\Sigma}). \tag{4.1}$$

We take $\stackrel{\text{apx}}{\sim}$ to mean "is approximately distributed as;" see the appendices for rigor. Given such an approximate distribution, we have $\widehat{\Sigma}^{-1/2}(\widehat{\beta}_n - \beta^\star) \stackrel{\text{apx}}{\sim} \mathsf{N}(0, I_d)$, so the confidence set

$$\mathcal{C}_n := \left\{ \beta \in \mathbb{R}^d \mid (\widehat{\beta}_n - \beta)^T \widehat{\Sigma}^{-1} (\widehat{\beta}_n - \beta) \le \chi^2_{d,1-\alpha} \right\} \tag{4.2}$$

where $\chi^2_{d,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2 random variable with d degrees of freedom, should have approximately the correct coverage

$$\mathbb{P}(\beta^\star \in \mathcal{C}_n) = 1 - \alpha + o(1).$$

Consider the following four methods to do this:

(i) Nonparametric bootstrap resampling: generate B different bootstrap resamples of the data, fit $\widehat{\beta}^{*b}$ on each resample, and set $\widehat{\Sigma}$ to be the empirical covariance of $\widehat{\beta}^{*b}$.

(ii) Parametric bootstrap: on the training data (X, Y), fit $\widehat{\beta}_n = (X^TX)^{-1}X^TY$. Resample the residuals $\widehat{\varepsilon} = Y - X\widehat{\beta}_n$ (with replacement) and fit models on new responses $Y^{*b} = X\widehat{\beta}_n + \widehat{\varepsilon}^{*b}$, where $\widehat{\varepsilon}^{*b}$ are resampled. Choose $\widehat{\Sigma}$ to be the empirical covariance of the refit models.

(iii) The "assumption-heavy" linear model: if the linear model $Y = X\beta^\star + \varepsilon$, $\varepsilon \sim \mathsf{N}(0, \sigma^2 I)$, holds then $\widehat{\beta}_n - \beta^\star \sim \mathsf{N}(0, \sigma^2(X^TX)^{-1})$, and for any consistent estimator $\widehat{\sigma}^2$ of the variance σ^2,

$$\widehat{\beta}_n - \beta^\star \stackrel{\text{apx}}{\sim} \mathsf{N}\left(0, \widehat{\sigma}^2(X^TX)^{-1}\right).$$

Use $\widehat{\Sigma} = \widehat{\sigma}^2(X^TX)^{-1}$.

(iv) The "assumption-light" model: whenever $y_i = \mathbb{E}[Y \mid X = x_i] + \varepsilon_i$, we have $Y = X\beta^\star + \eta + \varepsilon$, where η are "nonlinear" and uncorrelated with the x_i (recall Question 2.1), and we have

$$\widehat{\beta} - \beta^\star = (X^TX)^{-1}X^T(\eta + \varepsilon) \sim \left(0, (X^TX)^{-1} \sum_{i=1}^n (\eta_i + \varepsilon_i)^2 x_i x_i^T (X^TX)^{-1}\right)$$

(convince yourself of this). We have no access to the summation $\sum_{i=1}^n (\eta_i + \varepsilon_i)^2 x_i x_i^T$. Nonetheless, the "sandwich estimator of covariance," which replaces $\eta_i + \varepsilon_i$ with the empirical residual $\widehat{\varepsilon}_i := y_i - x_i^T \widehat{\beta}$, giving the estimate

$$\widehat{\Sigma} = (X^T X)^{-1} X^T \operatorname{diag}\left(\widehat{\varepsilon}_i^2\right) X (X^T X)^{-1},$$

is consistent. One can make fully rigorous (see Corollary C.8) the distributional identities

$$\sqrt{n}(\widehat{\beta} - \beta^\star) \stackrel{\text{apx}}{\sim} \mathsf{N}\left(0, (n^{-1} X^T X)^{-1} \frac{1}{n} \sum_{i=1}^n \widehat{\varepsilon}_i^2 x_i x_i^T (n^{-1} X^T X)^{-1}\right) = \mathsf{N}\left(0, n\widehat{\Sigma}\right).$$

With the procedures (i)–(iv) outlined above, we turn to the actual problem. Using the Abalone data from the UCI machine learning repository, you should compare the coverage of the confidence set (4.2) for each of the four methods for estimating a covariance. To do so, do the following experiment multiple times.

(a) Treat the full dataset (of size N) as the population, and take $\beta^\star$ to be the least-squares minimizer on the entire data.

(b) Take a subsample of size $n = N/2$ of the full dataset, and then fit $\widehat{\beta}_n$ on this subsample. Then perform each of the methods (i)–(iv) on the subsample, using $B = 400$ bootstrap replicates. For $\alpha = .05$ and $\alpha = .1$, check whether $\beta^\star$ belongs to the constructed confidence region $\mathcal{C}_n$.

(c) Repeat step (b) 100 times to get estimates of the coverage probabilities and report your results. Explain your observations. Include your code, but feel free to use any packages you like.

A few educational notes: We demonstrate how to make the approximations in (iv) above rigorous in appendices B–C to this exercise. Roughly, what is happening is the following more general phenomenon. Imagine we construct an M-estimator by loss minimization, so that for a loss $\ell(\theta; z)$ for parameter θ on datum z, we choose

$$\widehat{\theta}_n = \operatorname*{argmin}_\theta \frac{1}{n} \sum_{i=1}^n \ell(\theta; Z_i)$$

for an i.i.d. sequence of Z_i. Then (see Appendix B) we typically have that for the population loss $L(\theta) = \mathbb{E}[\ell(\theta; Z)]$ and associated Hessian $\nabla^2 L(\theta)$, the convergence (in the sense of distributions)

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \stackrel{d}{\to} \mathsf{N}\left(0, \nabla^2 L(\theta^\star)^{-1} \operatorname{Cov}(\nabla \ell(\theta^\star; Z)) \nabla^2 L(\theta^\star)^{-1}\right)$$

holds. Of course, the quantities on the right are population and hence unavailable; instead, we replace them by their empirical counterparts, that is,

$$\widehat{\Sigma} = \nabla^2 L_n(\widehat{\theta}_n)^{-1} \frac{1}{n} \sum_{i=1}^n \nabla \ell(\widehat{\theta}_n; Z_i) \nabla \ell(\widehat{\theta}_n; Z_i)^T \nabla^2 L_n(\widehat{\theta}_n)^{-1},$$

where $L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta; Z_i)$ is the empirical loss. Then one gets the distributional convergence

$$\sqrt{n} \widehat{\Sigma}^{-1/2} (\widehat{\theta}_n - \theta^\star) \stackrel{d}{\to} \mathsf{N}(0, I).$$

See Appendix C. By inspection, the normal limits for the least-squares error are simply an application of these general arguments to the particulars of the squared error $\ell(\beta; x, y) = \frac{1}{2}(x^T \beta - y)^2$, which satisfies $\nabla \ell(\beta; x, y) = x(x^T \beta - y)$ and hence $\nabla \ell(\widehat{\beta}; x, y) = x\widehat{\varepsilon}$, as well as the Hessian identity $\nabla^2 \ell(\beta; x, y) = xx^T$.

For more on "sandwich" estimators, see, for example, White [4], discussion by Freedman [2], as well as survey papers by Buja et al. [1].

A Convergence in distribution and probability

In this section, we provide a few of the definitions that will be necessary to obtain the distributional convergence results that we leverage in Question 4.1. We will be necessarily terse, as this section simply provides definitions, but we will also try to give a bit of intuition and motivation.

We begin with the two modes of convergence we require. We state convergence in distribution originally in terms of d-dimensional vectors, where we treat the inequality $z \le t$ for $z, t \in \mathbb{R}^d$ to mean that each entry j satisfies $z_j \le t_j$.

Definition A.1. *Let Z_n be a sequence of random vectors and Z a random vector. Then Z_n* converges in distribution to *Z, written $Z_n \xrightarrow{d} Z$, if*

$$\mathbb{P}(Z_n \le t) \to \mathbb{P}(Z \le t)$$

for all $t \in \mathbb{R}^d$ at which the mapping $t \mapsto \mathbb{P}(Z \le t)$ is continuous. If Z has a continuous distribution, then the convergence simply occurs for all t.

There are several equivalent definitions of convergence in distribution, which actually apply for random elements that take values in a metric space, and which sometimes are useful.

Proposition A.1 (Portmanteau Lemma)**.** *For random vectors, the following are equivalent:*

(a) $Z_n \xrightarrow{d} Z$.

(b) For all bounded and continuous functions f, $\mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]$.

(c) For all bounded and Lipschitz continuous functions f, $\mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]$.

(d) If C is a closed set, then $\limsup_n \mathbb{P}(Z_n \in C) \le P(Z \in C)$.

(e) If O is open, then $\liminf \mathbb{P}(Z_n \in O) \ge \mathbb{P}(Z \in O)$.

(f) Let the boundary of a set A be $\operatorname{bd} A = \operatorname{cl} A \setminus \operatorname{int} A$. *Then* $\mathbb{P}(Z_n \in A) \to \mathbb{P}(Z \in A)$ *whenever A is a* continuity set, *meaning* $\mathbb{P}(Z \in \operatorname{bd} A) = 0$.

As this proposition is beyond the scope of the course, we simply refer the reader to any standard text on Asymptotic statistics. Some treatments include van der Vaart [3, Ch. 2.1, Lemma 2.2].

Sometimes we need to deal with convergence in probability, which is stronger than convergence in distribution.

Definition A.2. *A sequence of random vectors Z_n* converges in probability *to Z, written $Z_n \xrightarrow{p} Z$, if $\mathbb{P}(\|Z_n - Z\| \ge \epsilon) \to 0$ for all $\epsilon > 0$.*

Convergence in probability is stronger than convergence in distribution:

Lemma A.1. *Let $Z_n \xrightarrow{p} Z$. Then $Z_n \xrightarrow{d} Z$.*

Proof Let f be 1-Lipschitz and bounded by 1. Then for any $0 < \epsilon \le 2$,

$$|\mathbb{E}[f(Z_n) - f(Z)]| \le \mathbb{E}[\|Z_n - Z\| \wedge 2] \le \epsilon \mathbb{P}(\|Z_n - Z\| \le \epsilon) + 2\mathbb{P}(\|Z_n - Z\| \ge \epsilon).$$

The latter term tends to zero and the first term on the right hand side is bounded by ϵ. □

It will also be convenient for us to use what is known as *stochastic big-Oh notation*, which generalizes typical (asymptotic) big-Oh notation to random sequences.

Definition A.3 (Big-Oh notation)**.** *Let Z_n be random vectors and R_n be a random or non-random sequence of nonnegative numbers. Then $Z_n = O_P(R_n)$, read* big-Oh-P *of R_n, if for all $\epsilon > 0$ there exists $C < \infty$ such that*

$$\lim_{n\to\infty} \mathbb{P}(\|Z_n\| \geq CR_n) \leq \epsilon.$$

Additionally, $Z_n = o_P(R_n)$, read little-Oh-P *of R_n, if for all $c > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\|Z_n\| \geq cR_n) = 0.$$

As a simple example of Definition A.3, we have $Z_n - Z = o_P(1)$ if and only if $Z_n \xrightarrow{p} Z$, which in turn occurs if and only if $Z_n - Z \xrightarrow{p} 0$.

A frequently useful result is the *continuous mapping theorem*, which states that

Theorem A.2 (Continuous mapping)**.** *Let f be continuous on a set C such that $\mathbb{P}(Z \in C) = 1$. Then if $Z_n \xrightarrow{d} Z$, $f(Z_n) \xrightarrow{d} f(Z)$, and if $Z_n \xrightarrow{p} Z$, then $f(Z_n) \to f(Z)$.*

Proof We prove the convergence $f(Z_n) \xrightarrow{d} f(Z)$ assuming that f is continuous; the extension uses a more sophisticated application of the Portmanteau lemma than we use here (see, e.g., van der Vaart [3, Theorem 2.3]). In this case, for any bounded continuous g, the function $h = g \circ f$ is bounded and continuous, so $\mathbb{E}[g(f(Z_n))] \to \mathbb{E}[g(f(Z))]$.

Assume now that $Z_n \xrightarrow{p} Z$ and let $\epsilon > 0$ be arbitrary. Let B_δ be the set of points x such that there exists y with $\|y - x\| < \delta$ but $\|f(x) - f(y)\| \geq \epsilon$. Then $\|f(Z_n) - f(Z)\| \geq \epsilon$ implies that either $Z \in B_\delta$ or $\|Z_n - Z\| \geq \delta$. Then

$$\mathbb{P}(\|f(Z_n) - f(Z)\| \geq \epsilon) \leq \mathbb{P}(Z \in B_\delta) + \underbrace{\mathbb{P}(\|Z_n - Z\| \geq \delta)}_{\to 0},$$

and $\lim_{\delta \downarrow 0} B_\delta \cap C = \emptyset$. Taking $\delta \to 0$ gives that $\mathbb{P}(\|f(Z_n) - f(Z)\| \geq \epsilon) \to 0$. □

Combining the big-Oh notation in Definition A.3 with continuous mapping, we have the following corollary, whose proof we leave as an exercise.

Corollary A.3 (Some O_P notation)**.** *The following hold.*

(i) If $Z_n \xrightarrow{d} Z$ and $E_n = o_P(1)$, then $Z_n + E_n \xrightarrow{d} Z$.

(ii) If r_n is a (deterministic) increasing sequence and $r_n Z_n \xrightarrow{d} Z$ and $E_n = o_P(r_n^{-1}$, then $r_n(Z_n + E_n) \xrightarrow{d} Z$

(iii) If r_n is any deterministic seqeunce, then $Z_n = O_P(r_n)$ implies that $Z_n = o_P(R_n)$ for any sequence R_n satisfying $R_n/r_n \to \infty$.

The final set of asymptotic results we require are a few results that upgrade convergence in distribution of individual vectors to a type of joint convergence in distribution, though they require that one of the random variables converge to a constant. (As otherwise the results may fail.) We begin with a relatively simple observation that convergence in distribution to a constant implies convergence in probability:

Lemma A.2. *Let $Z_n \xrightarrow{d} c$ where c is a constant. Then $Z_n \xrightarrow{p} c$.*

Proof Let $K < \infty$ and define $f_K(x) = (1 - K\|x - c\|)_+$. Then f_K is bounded and K-Lipschitz, and it evidently satisfies $\mathbb{E}[f_K(X_n)] \to 1$. But $f_K(x) \le 1 - \mathbf{1}\{\|x - c\| \ge 1/K\} = \mathbf{1}\{\|x - c\| < 1/K\}$, and so

$$\liminf_{n\to\infty} \mathbb{P}(\|X_n - c\| < 1/K) \ge \liminf_n \mathbb{E}[f_K(X_n)] = 1.$$

As K was arbitrary this gives the result. □

With Lemma A.2 in hand, we can now prove Slutsky's theorems, which upgrade convergence of pairs of random vectors to joint convergence.

Theorem A.4 (Slutsky)**.** *Assume that $X_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{p} c$, where c is a constant. Then*

$$(X_n, Y_n) \xrightarrow{d} (Z, c).$$

Proof Let f be a 1-Lipschitz function bounded by 1. Then

$$\mathbb{E}[f(X_n, Y_n) - f(Z, c)] = \mathbb{E}[f(X_n, Y_n) - f(X_n, c)] + \underbrace{\mathbb{E}[f(X_n, c) - f(Z, c)]}_{\to 0},$$

as $x \mapsto f(x, c)$ is Lipchitz continuous and bounded. Applying Jensen's inequality to the first term, we have

$$|\mathbb{E}[f(X_n, Y_n) - f(X_n, c)]| \le \mathbb{E}[|f(X_n, Y_n) - f(X_n, c)|] \le \mathbb{E}[2 \wedge \|Y_n - c\|] \le 2\mathbb{P}(\|Y_n - c\| \ge \epsilon) + \epsilon$$

for all $\epsilon > 0$. Take $n \to \infty$ and apply the Portmanteau result (Proposition A.1). □

An important corollary of Theorem A.4 is that we can combine it with the continuous mapping theorem to get that various inverse matrices converge.

Corollary A.5. *Let $C_n \xrightarrow{p} C$ where C is an invertible matrix, and let $X_n \xrightarrow{d} Z$. Then $C_n^{-1} X_n \xrightarrow{d} C^{-1} X$.*

The corollary is an immediate consequence of the fact that the inverse mapping $A \mapsto A^{-1}$ is continuous at any A for which the inverse exists, and of course multiplication $(A, x) \mapsto Ax$ is continuous. Note that implicitly in the corollary—and this highlights the power of the Continuous Mapping Theorem A.2—is that the inverses C_n^{-1} make sense (eventually), even though individual C_n may not be invertible.

B Convergence of M-estimators

We consider the following standard setting, which is a generalization of Question 4.1. We assume we have losses $\ell(\theta; z)$ where $\ell(\theta; z)$ measures the fidelity of a parameter θ for example z. A few typical examples include linear regression, where $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$ for covariates x and response y and

$$\ell(\theta; x, y) = \frac{1}{2}(x^T\theta - y)^2;$$

absolute (ℓ_1) regression, where $\ell(\theta; x, y) = |x^T\theta - y|$; or smooth robust regression losses, including $\ell(\theta; x, y) = \log(1 + \exp(y - x^T\theta)) + \log(1 + \exp(x^T\theta - y))$. The standard M-estimator is then to minimize the empirical loss

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i)$$

for Z_i drawn i.i.d. from the underlying population. Defining the population loss

$$L(\theta) := \mathbb{E}[\ell(\theta; Z)],$$

typical results show that (under some appropriate regularity conditions) the empirical minimizer of L_n has an asymptotically normal distribution with a given variance.

Some such conditions include the following:

Assumption 1 (Regularity conditions for an M-estimator)**.** *Let $\theta^\star = \operatorname{argmin}_\theta L(\theta)$. The following hold:*

(i) The losses $\ell(\cdot; z)$ are convex and twice differentiable for each z.

(ii) The Hessian $\nabla^2 L(\theta)$ is positive definite at $\theta^\star$.

(iii) There exists $M_2 : \mathcal{Z} \to \mathbb{R}_+$ with $\overline{M_2} := \mathbb{E}[M_2(Z)] < \infty$ such that

$$\left\|\nabla^2\ell(\theta_0; z) - \nabla^2\ell(\theta_1; z)\right\|_{\text{op}} \le M_2(z)$$

for all θ_0, θ_1 near $\theta^\star$.

(iv) The covariance $C := \mathbb{E}[\nabla\ell(\theta^\star; Z)\nabla\ell(\theta^\star; Z)^T]$ exists.

Theorem B.6. *Let the regularity conditions in Assumption 1 hold. Then $\widehat{\theta}_n \xrightarrow{p} \theta^\star$, and*

$$\widehat{\theta}_n - \theta^\star = -\nabla^2 L(\theta^\star)^{-1}\nabla L_n(\theta^\star) + o_P(n^{-1/2}). \tag{B.1}$$

Consequently,

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \mathsf{N}\left(0, \nabla^2 L(\theta^\star)^{-1} C \nabla^2 L(\theta^\star)^{-1}\right).$$

Proof The proof proceeds in two parts. The first is to prove consistency, that is, that $\widehat{\theta}_n \xrightarrow{p} \theta^\star$. The second, using this consistency, proves the asymptotic expansion B.1.

To prove consistency, we will show that any point sufficiently far from $\theta^\star$ must necessarily suffer large empirical loss, so that it cannot be $\widehat{\theta}_n$. For this, we require an auxiliary lemma that relies on convexity of the loss.

Lemma B.1. *Let $b, c > 0$, and let f be a convex function with $f(\theta_1) \ge f(\theta_0) + c$ for all θ_1 satisfying $\|\theta_1 - \theta_0\| = b$. Then*

$$f(\theta) \ge f(\theta_0) + c \cdot \frac{\|\theta - \theta_0\|}{b} \quad \text{for } \|\theta - \theta_0\| > b.$$

Proof Let $\|\theta - \theta_0\| > b$, and define $t = b/\|\theta - \theta_0\|$. The point $\theta_1 = \theta_0 + t(\theta - \theta_0) = (1-t)\theta_0 + t\theta$ then satisfies $\|\theta_1 - \theta_0\| = b$, and

$$f(\theta_1) = f((1-t)\theta_0 + t\theta) \le (1-t)f(\theta_0) + tf(\theta),$$

so that

$$tf(\theta) \geq f(\theta_1) - f(\theta_0) + tf(\theta_0) \geq c + tf(\theta_0).$$

Dividing through by $t = b/\left\|\theta - \theta_0\right\|$ gives the result. □

Now, fix $b > 0$ to be chosen presently, and define the minimal eigenvalue

$$\Lambda_n(b) := \inf_{\|\theta - \theta^\star\| \leq b} \lambda_{\min}(\nabla^2 L_n(\theta)).$$

Then by a Taylor expansion, for all $\|\theta - \theta^\star\| \leq b$ we have

$$L_n(\theta) \geq L_n(\theta^\star) + \nabla L_n(\theta^\star)^T(\theta - \theta^\star) + \frac{1}{2}\Lambda_n(b)\left\|\theta - \theta^\star\right\|^2. \tag{B.2}$$

Let $\lambda_\star = \lambda_{\min}(\nabla^2 L(\theta^\star))$ and b be chosen small enough that the (mean) Hessian Lipschitz constant $\overline{M_2} = \mathbb{E}[M_2(Z)]$ satisfies $b\overline{M_2} \leq \frac{1}{4}\lambda_\star$. Then

$$\Lambda_n(b) \geq \lambda_{\min}(\nabla^2 L_n(\theta^\star)) - \sup_{\|\theta - \theta^\star\| \leq b} \frac{1}{n}\sum_{i=1}^n M_2(Z_i)\left\|\theta - \theta^\star\right\| \xrightarrow{p} \lambda_\star - \overline{M_2}b \geq \frac{3}{4}\lambda_\star,$$

we define the event $\mathcal{E}_1 := \{\Lambda_n(b) \geq \frac{1}{2}\lambda_\star\}$, which we see happens eventually.[1] Let $\epsilon > 0$ also be arbitrary, and define the additional event that

$$\mathcal{E}_2 := \left\{\|\nabla L_n(\theta^\star)\| \leq \epsilon\right\},$$

which also happens eventually by the strong law of large numbers. So on $\mathcal{E}_1 \cap \mathcal{E}_2$, we have by inequality (B.2) that

$$L_n(\theta) \geq L_n(\theta^\star) - \epsilon\left\|\theta - \theta^\star\right\| + \frac{\lambda_\star}{4}\left\|\theta - \theta^\star\right\|^2,$$

and if $\|\theta - \theta^\star\| = b$, then $L_n(\theta) \geq L_n(\theta^\star) + b(\frac{\lambda^\star}{4}b - \epsilon)$. As $\epsilon > 0$ was arbitrary, we can take it smaller than $\lambda_\star b/4$, yielding that for all $\|\theta - \theta^\star\| = b$ we have $L_n(\theta) > L_n(\theta^\star)$. Applying Lemma B.1 gives that we must therefore have

$$\|\widehat{\theta}_n - \theta^\star\| \leq b \quad \text{on} \quad \mathcal{E}_1 \cap \mathcal{E}_2.$$

As $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \to 1$ and $b > 0$ is arbitrary, we can take $b \downarrow 0$ to obtain $\widehat{\theta}_n - \theta^\star = o_P(1)$.

For the asymptotic normality and expansion (B.1), we rely on a different Taylor expansion. We first recall that if $f : \mathbb{R}^d \to \mathbb{R}$ has M-Lipschitz Hessian, then

$$\begin{aligned}
\nabla f(y) &= \nabla f(x) + \int_0^1 \nabla^2 f(x + t(y - x))(y - x)dt \\
&= \nabla f(x) + \nabla^2 f(x)(y - x) + \underbrace{\int_0^1 (\nabla^2 f(x + t(y - x)) - \nabla^2 f(x))dt}_{=:E} \cdot (y - x) \\
&= \nabla f(x) + (\nabla^2 f(x) + E)(y - x),
\end{aligned}$$

[1] Technically we've only proved that $\mathbb{P}(\mathcal{E}_1) \to 1$, but it also happens almost surely.

where the error matrix $E = \int_0^1 (\nabla^2 f(x + t(y-x)) - \nabla^2 f(x))dt$ satisfies $\|E\|_{\rm op} \le \int_0^1 Mt \, \|y - x\| \, dt = \frac{M}{2} \|y - x\|$. With this in mind, we see that because $\widehat{\theta}_n$ minimizes $L_n(\theta)$, we have

$$0 = \nabla L_n(\widehat{\theta}_n) = \nabla L_n(\theta^\star) + (\nabla^2 L_n(\theta^\star) + E_n)(\widehat{\theta}_n - \theta^\star),$$

where the error matrix E_n satisfies

$$\|E_n\|_{\rm op} \le \frac{1}{n} \sum_{i=1}^n M_2(Z_i) \|\widehat{\theta}_n - \theta^\star\|.$$

But we know that $\widehat{\theta}_n \xrightarrow{p} \theta^\star$ by the consistency argument above, and so $\|E_n\|_{\rm op} = o_P(1)$. The strong law of large numbers gives $\nabla^2 L_n(\theta^\star) = \nabla^2 L(\theta^\star) + o_P(1)$, so finally, we apply Slutsky's Theorem A.4 and continuous mapping to obtain

$$\widehat{\theta}_n - \theta^\star = (\nabla^2 L_n(\theta^\star) + E_n)^{-1} \nabla L_n(\theta^\star) = \nabla^2 L(\theta^\star)^{-1} \nabla L_n(\theta^\star) + o_p(1) \cdot \nabla L_n(\theta^\star).$$

But $\mathbb{E}[\|\nabla L_n(\theta^\star)\|^2] = \frac{1}{n} \mathbb{E}[\|\nabla \ell(\theta^\star; Z)\|^2] = \frac{1}{n} \operatorname{tr}(C)$, where $C = \operatorname{Cov}(\nabla \ell(\theta^\star; Z))$, and so the final term is $o_P(1/\sqrt{n})$. This gives equality (B.1).

The final claim—that $\sqrt{n}(\widehat{\theta}_n - \theta^\star)$ is asymptotically normal—is an application of the central limit theorem and Corollary A.3 on big-Oh notation. Because $\mathbb{E}[\nabla \ell(\theta^\star; Z)] = 0$ and so $\operatorname{Cov}(\sqrt{n} \nabla L_n(\theta^\star)) = C$, we have $\sqrt{n} \nabla L_n(\theta^\star) \xrightarrow{d} \mathsf{N}(0, C)$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) = \nabla^2 L(\theta^\star)^{-1} \sqrt{n} \nabla L_n(\theta^\star) + \underbrace{\sqrt{n} \cdot o_P(1/\sqrt{n})}_{\xrightarrow{p} 0},$$

which has the desired convergence by applying continuous mapping. □

C Pivotal asymptotic estimates via M-estimators

While the asymptotics in Theorem B.6 are powerful—and indeed, are the "standard" asymptotic normality result for M-estimators—it interesting to develop actionable estimators that allow us to provide so-called *pivotal asymptotics*, by which we mean estimates $\widehat{\Sigma}$ of the covariance

$$\Sigma := \nabla^2 L(\theta^\star)^{-1} \operatorname{Cov}(\nabla \ell(\theta^\star; Z)) \nabla^2 L(\theta^\star)^{-1}$$

that satisfy $\widehat{\Sigma} \xrightarrow{p} \Sigma$. Notably, with such a consistent estimator of the covariance, we have

$$\begin{aligned} \sqrt{n} \widehat{\Sigma}^{-1/2} (\widehat{\theta}_n - \theta^\star) &\xrightarrow{d} \mathsf{N}(0, I) \quad \text{and} \\ n (\widehat{\theta}_n - \theta^\star)^T \widehat{\Sigma}^{-1} (\widehat{\theta}_n - \theta^\star) &\xrightarrow{d} \|Z\|_2^2 \quad \text{where} \quad Z \sim \mathsf{N}(0, I_d) \end{aligned} \tag{C.1}$$

by the continuous mapping and Slutsky's theorems. From this, one has the following immediate approximate confidence set for $\theta^\star$: for a desired level α, set

$$\mathcal{C}_{n,\alpha} := \left\{ \theta \in \mathbb{R}^d \mid n (\widehat{\theta} - \theta)^T \widehat{\Sigma}^{-1} (\widehat{\theta} - \theta) \le \chi^2_{d, 1-\alpha} \right\},$$

where $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)$ quantile of a χ^2 random variable with d degrees of freedom (i.e., of $\|Z\|_2^2$ for $Z \sim \mathsf{N}(0, I_d)$). Then by the definition of convergence in distribution, the limits (C.1) guarantee

$$\mathbb{P}(\theta^\star \in \mathcal{C}_{n,\alpha}) \to 1 - \alpha \quad \text{as} \quad n \to \infty.$$

Thus, to achieve asymptotically valid confidence intervals, it suffices to give a consistent estimate of Σ. The next theorem does this.

Theorem C.7. *Let the conditions of Theorem B.6 hold and additionally assume that the Hessian satisfies* $\mathbb{E}[\|\nabla^2 \ell(\theta^\star; Z)\|_{\text{op}}^2] < \infty$ *and* $\mathbb{E}[M_2^2(Z)] < \infty$. *Define*

$$\widehat{\Sigma} := \nabla^2 L_n(\widehat{\theta}_n)^{-1} \frac{1}{n} \sum_{i=1}^n \nabla \ell(\widehat{\theta}_n; Z_i) \nabla \ell(\widehat{\theta}_n; Z_i)^T \nabla^2 L_n(\widehat{\theta}_n)^{-1}.$$

Then $\widehat{\Sigma} \xrightarrow{p} \Sigma$.

Proof By the continuous mapping theorem, it suffices to show the Hessian and covariance convergences

$$\nabla^2 L_n(\widehat{\theta}_n) \xrightarrow{p} \nabla^2 L(\theta^\star) \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n \nabla \ell(\widehat{\theta}_n; Z_i) \nabla \ell(\widehat{\theta}_n; Z_i)^T \xrightarrow{p} \text{Cov}(\nabla \ell(\theta^\star; Z)). \tag{C.2}$$

We begin with the Hessian convergence in expression (C.2). By Assumption 1, we immediately see that

$$\nabla^2 L_n(\widehat{\theta}_n) - \nabla^2 L_n(\theta^\star) \le \underbrace{\frac{1}{n} \sum_{i=1}^n M_2(Z_i)}_{\xrightarrow{a.s.} \overline{M_2} = \mathbb{E}[M_2(Z)]} \underbrace{\|\widehat{\theta}_n - \theta^\star\|_2}_{\xrightarrow{p} 0} = o_P(1),$$

and similarly $\nabla^2 L_n(\theta^\star) \xrightarrow{a.s.} \nabla^2 L(\theta^\star)$ by the strong law of large numbers.

The covariance convergence in the limits (C.2) is harder. For this, we let $G_i = \nabla \ell(\theta^\star; Z_i)$, so that G_i are i.i.d. with $\mathbb{E}[G_i] = 0$ and $C := \text{Cov}(G_i) = \text{Cov}(\nabla \ell(\theta^\star; Z))$. Also let $H_i = \nabla^2 \ell(\theta^\star; Z_i)$ be the random Hessian for example i. Then

$$\nabla \ell(\widehat{\theta}_n; Z_i) = G_i + \underbrace{\nabla \ell(\widehat{\theta}_n; Z_i) - \nabla \ell(\theta^\star; Z_i)}_{r_{n,i}}$$

where the remainder $r_{n,i}$ term is

$$r_{n,i} = \int_0^1 \nabla^2 \ell(\theta^\star + t(\widehat{\theta}_n - \theta^\star); Z_i) dt \cdot (\widehat{\theta}_n - \theta^\star) = H_i(\widehat{\theta}_n - \theta^\star) + \underbrace{\int_0^1 (H_i - \nabla^2 \ell(\theta^\star + t(\widehat{\theta}_n - \theta^\star); Z_i)) dt \cdot (\widehat{\theta}_n - \theta^\star)}_{=: \Delta_{n,i}},$$

by Taylor's theorem, and the error $\Delta_{n,i}$ satisfies $\|\Delta_{n,i}\| \le \frac{1}{2} M_2(Z_i) \|\widehat{\theta}_n - \theta^\star\|_2^2$. Thus,

$$\begin{aligned} \nabla \ell(\widehat{\theta}_n; Z_i) \nabla \ell(\widehat{\theta}_n; Z_i)^T &= (G_i + H_i(\widehat{\theta}_n - \theta^\star) + \Delta_{n,i})(G_i + H_i(\widehat{\theta}_n - \theta^\star) + \Delta_{n,i})^T \\ &= G_i G_i^T + H_i(\widehat{\theta}_n - \theta^\star)(\widehat{\theta}_n - \theta^\star)^T H_i^T \\ &\quad + \Delta_{n,i} \Delta_{n,i}^T + H_i(\widehat{\theta}_n - \theta^\star)(G_i + \Delta_{n,i})^T + (G_i + \Delta_{n,i}) H_i(\widehat{\theta}_n - \theta^\star)^T. \end{aligned} \tag{C.3}$$

We can bound each of these outer product matrices by a combination of Cauchy-Schwarz and triangle inequalities. We have

$$\left\| \frac{1}{n} \sum_{i=1}^n H_i(\widehat{\theta}_n - \theta^\star)(\widehat{\theta}_n - \theta^\star)^T H_i^T \right\|_{\text{op}} \le \|\widehat{\theta}_n - \theta^\star\|_2^2 \cdot \frac{1}{n} \sum_{i=1}^n \|H_i\|_{\text{op}}^2 = O_P(1/n) \cdot O_P(1) = O_P(1/n)$$

by Theorem B.6 and the assumption that $\mathbb{E}[\|\nabla^2 \ell(\theta^\star; Z)\|_{\rm op}^2] < \infty$, as $\widehat{\theta}_n - \theta^\star = O_P(1/\sqrt{n})$. Similarly,

$$\left\| \frac{1}{n} \sum_{i=1}^n \Delta_{n,i} \Delta_{n,i}^T \right\|_{\rm op} \le \frac{1}{4} \sum_{i=1}^n M_2(Z_i)^2 \left\| \widehat{\theta}_n - \theta^\star \right\|_2^4 = O_P(1) \cdot O_P(1/n^2) = O_P(1/n^2)$$

because $4 \left\| \Delta_{n,i} \right\|_2^2 \le M_2(Z_i)^2 \|\widehat{\theta}_n - \theta^\star\|_2^4$. Finally, we have

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n H_i (\widehat{\theta}_n - \theta^\star)(G_i + \Delta_{n,i})^T \right\|_{\rm op} &\le \left\| \widehat{\theta}_n - \theta^\star \right\|_2 \frac{1}{n} \sum_{i=1}^n \left\| H_i \right\|_{\rm op} \left\| G_i + \Delta_{n,i} \right\| \\
&\le \left\| \widehat{\theta}_n - \theta^\star \right\|_2 \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n \left\| H_i \right\|_{\rm op}^2} \sqrt{\frac{2}{n} \sum_{i=1}^n (\|G_i\|^2 + \|\Delta_{n,i}\|^2)} \\
&= O_P(1/\sqrt{n}) \cdot O_P(1) \cdot O_P(1),
\end{aligned}$$

where we have applied the triangle inequality and Cauchy-Schwarz. Combining these identities with the expansion (C.3), we see that

$$\frac{1}{n} \sum_{i=1}^n \nabla \ell(\widehat{\theta}_n; Z_i) \nabla \ell(\widehat{\theta}_n; Z_i)^T = \frac{1}{n} \sum_{i=1}^n G_i G_i^T + O_P(1/\sqrt{n}) \xrightarrow{p} \mathrm{Cov}(G_i) = \mathrm{Cov}(\nabla \ell(\theta^\star; Z))$$

as desired, completing the proof of the convergences (C.2) and hence the theorem. □

An immediate corollary of the theorem applies to the sandwich estimator of variance for regression. For regression problems with squared error $\ell(\beta; x, y) = \frac{1}{2}(x^T \beta - y)^2$, note that $\nabla^2 \ell(\beta; x, y) = xx^T$, so that if $\mathbb{E}[\|x\|_2^4] < \infty$ and $\mathbb{E}[xx^T] \succ 0$, all the conditions of Theorem C.7 automatically hold (even with $M_2 \equiv 0$). So then with $\widehat{\varepsilon}_i = y_i - x_i^T \widehat{\beta}$, we have

Corollary C.8. *Let the preceding conditions hold and let $\widehat{\beta}$ be the ordinary least squares estimator. Then*

$$\widehat{\Sigma} := \frac{1}{n} (n^{-1} X^T X)^{-1} X^T \operatorname{diag}(\widehat{\varepsilon}) X (n^{-1} X^T X)^{-1}$$

satisfies

$$\sqrt{n} \widehat{\Sigma}^{-1/2} (\widehat{\beta} - \beta^\star) \xrightarrow{d} \mathsf{N}(0, I_d).$$

References

[1] A. Buja, L. Brown, R. Berk, E. George, E. Pitkin, M. Traskin, K. Zhang, and L. Zhao. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4): 523–544, 2019.

[2] D. A. Freedman. On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60(4):299–302, 2006.

[3] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[4] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.