Contents

Chapter 1

The Jackknife

The Jackknife is a tool for estimating the bias and variance of a statistic. It is based on the principle of repeatedly computing the statistic by removing at each time one observation. In this chapter, we introduce the theory for the Jackknife and prove consistency results for variance and bias estimation. The statistics considered are smooth functions of the sample mean. Such a class covers a wide range of estimators encountered in statistic. The organization of this chapter is as follows: Section 3.1 introduces the statistical framework and provides the intuition behind the bias and variance jackknife estimators; Section 1.2 contains some examples of applications; Section 1.3 proves the consistency of the jackknife variance and bias estimators; finally, Section 1.4 discusses some extensions.

1.1 The framework and some intuition for the Jackknife estimators

We observe an iid sample $\{X_i\}_{i=1}^n$ and want to make inference on an unknown population parameter θ. To this end, we use a statistic (an estimator)

$$T_n := T_n(X_1, .., X_n) \ .$$

The Jackknife for bias reduction. It often turns out that T_n is a biased estimator of θ and the Jackknife can be used to reduce its bias. To see why, let us start by introducing the bias of T_n

$$Bias(T_n) := \mathbb{E}\{T_n\} - \theta \ .$$

The Jackknife estimator of the bias of T_n is formally defined as

$$b_{jack} := (n-1)\left(\overline{T}_n - T_n\right) \text{, where } \overline{T}_n := \frac{1}{n}\sum_{i=1}^{n} T_{n-1,i} \qquad (1.1)$$

leave - one - out

and $T_{n-1,i}$ is the statistic computed by dropping the ith observation from the sample, i.e.

$$T_{n-1,i} := T_{n-1}(X_1, .., X_{i-1}, X_{i+1}, .., X_n)\,.$$

An estimator with the above structure is often called "leave-one-out" estimator. The *bias-corrected jackknife estimator* is therefore

$$T_{jack} := T_n - b_{jack}\,. \tag{1.2}$$

Tjack could still be biased.

T_{jack} will admit a lower bias compared to T_n. To heuristically justify this, assume that the bias of the statistic admits the following expansion:

$$\text{Bias}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right)\,. \quad = O\left(\frac{1}{n}\right) \tag{1.3}$$

The above expansion can be often derived after imposing smoothness conditions on the functional form of T_n. Since our present purpose is only to give a heuristic justification for using b_{jack}, we will not enter such details. By (1.2) we obtain

$$\begin{aligned}\text{Bias}(T_{jack}) =& \text{Bias}(T_n) - \mathbb{E}\{b_{jack}\}\\ =& \text{Bias}(T_n) - (n-1)\left[\mathbb{E}\{\overline{T}_n\} - \mathbb{E}\{T_n\}\right]\,.\end{aligned}$$

The iid assumption ensures that $\mathbb{E}\{\overline{T_n}\} = \mathbb{E}\{T_{n-1,1}\}$, so $\mathbb{E}\{\overline{T}_n\} - \mathbb{E}\{T_n\} = \text{Bias}(T_{n-1,1}) - \text{Bias}(T_n)$. Since Equation (1.3) also holds for the statistic $T_{n-1,1}$, the quantity in the square brackets on the RHS equals

$$\begin{aligned}\text{Bias}(T_{n-1,1}) - \text{Bias}(T_n) &= \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right) - \frac{a}{n} - \frac{b}{n^2} - O\left(\frac{1}{n^3}\right)\\ &= \frac{a}{n(n-1)} + b\frac{2n-1}{n^2(n-1)^2} + O\left(\frac{1}{n^3}\right)\,.\end{aligned}$$

Gathering together the previous two displays gives

$$\begin{aligned}\text{Bias}(T_{jack}) &= \text{Bias}(T_n) - \frac{a}{n} - b\frac{2n-1}{n^2(n-1)} + \cancel{O\left(\frac{1}{n^3}\right)} \quad O\left(\frac{1}{n^2}\right)\\ &= \frac{b}{n^2} - b\frac{2n-1}{n^2(n-1)} + O\left(\frac{1}{n^3}\right)\\ &= O\left(\frac{1}{n^2}\right)\,.\end{aligned}$$

By comparing the above display with (1.3) we obtain that the bias of T_{jack} is of lower order than the bias of T_n. This justifies T_{jack} as a bias-reduced version of T_n and b_{jack} as an estimator of $\text{Bias}(T_n)$.

Bias(Tjack) converges faster to 0.

Remark 1.1.1. The argument provided here is heuristic. In practice, the superior finite sample bias properties of T_{jack} compared to T_n are shown in simulations.

Remark 1.1.2. Notice that the b_{jack} does not require to know the analytic expression of $\text{Bias}(T_n)$ and can be computed in a straightforward way.

The consistency of b_{jack} as an estimator of $\text{Bias}(T_n)$ will be proved in Section 1.3

$$\overset{(T_n)}{Var\left(\frac{1}{n}\sum X_i\right)} = \frac{1}{n} \cdot Var(X)$$

$$Var(\boxed{\sqrt{n} \cdot T_n}) = Var(X) = V$$

V is Variance of X_i

The Jackknife for variance estimation. The jackknife can also be used to estimate the (asymptotic) variance of T_n. It is often the case that $\sqrt{n}(T_n - \theta) \rightsquigarrow \mathcal{N}(0, V)$, where V denotes the asymptotic variance of $\sqrt{n}T_n$. In some cases, the estimation of V is difficult to implement as ① V has a difficult expression, or ② the available estimators of V do note perform well in finite sample, see Section 1.2. The Jackknife provides a valuable alternative by giving an *automatic* method for the estimation of V.
To construct the Jackknife variance estimation, we start by expressing the bias reduced estimator T_{jack} as a sample mean. From the definitions of T_{jack} in (1.2) and b_{jack} in (1.1) we get

$$\begin{aligned}
T_{jack} &= T_n - (n-1)\left[\overline{T_n} - T_n\right] \\
&= nT_n - (n-1)\overline{T_n} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[nT_n - (n-1)T_{n-1,i}\right] \\
&=: \frac{1}{n}\sum_{i=1}^{n}\widetilde{T}_{n,i} \ .
\end{aligned}$$

The construction of the Jackknife variance estimator starts from the following idea: the variance of T_{jack} can approximate the variance of T_n. So, if $\widetilde{T}_{n,i}$ with $i = 1, \ldots, n$, can roughly be thought as being iid, then $Var(T_{jack}) \approx (1/n)Var(\widetilde{T}_{n,1})$. Thus, the jackknife variance estimator is defined as

$$Var(\widetilde{T}_{n,1}) \simeq \frac{1}{n-1}\underbrace{\sum_{j=1}^{n}\left[\widetilde{T}_{n,i} - \overline{\widetilde{T}_n}\right]^2}_{\text{Covariance of } \widetilde{T}}$$

$$\begin{aligned}
v_{jack} &:= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\widetilde{T}_{n,i} - \frac{1}{n}\sum_{j=1}^{n}\widetilde{T}_{n,j}\right]^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[nT_n - (n-1)T_{n-1,i} - \frac{1}{n}\sum_{j=1}^{n}(nT_n - (n-1)T_{n-1,j})\right]^2 \\
&= \frac{n-1}{n}\boxed{\sum_{i=1}^{n}\left[T_{n-1,i} - \frac{1}{n}\sum_{j=1}^{n}T_{n-1,j}\right]^2}. \qquad (1.4)
\end{aligned}$$

Similarly to b_{jack}, v_{jack} does not require knowing the analytic expression of the (asymptotic) variance of $\sqrt{n}T_n$ and can be straightforwardly computed.

1.2 Examples of Application

OLS inference with heteroskedasticity. We observe $\{Y_i, Z_i\}_{i=1}^n$, with Z_i valued in $\mathbb{R}^k$ and Y_i real valued. We are interested in making inference on the coefficients of the linear projection of Y onto X. So, consider the regression

$$Y_i = Z^T \beta_0 + \varepsilon \text{ with } \mathbb{E}\varepsilon Z = 0$$

Assume that $\mathbb{E}ZZ^T$ is invertible (i.e. there is no collinearity between the explanatory variables Z). Then, $\beta_0 = (\mathbb{E}ZZ^T)^{-1}\mathbb{E}ZY$ and the OLS estimator is

$$\widehat{\beta} := \left(\overline{ZZ^T}\right)^{-1} \overline{ZY} .$$

When the errors are heteroskedastic, $\mathbb{E}\{\varepsilon^2|Z\}$ is not constant, and from basic statistical textbooks we know that

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}(0, \Sigma) \text{ with } \Sigma = (\mathbb{E}ZZ^T)^{-1}(\mathbb{E}\varepsilon^2 ZZ^T)(\mathbb{E}ZZ^T)^{-1} ,$$

where $\rightsquigarrow$ denotes convergence in distribution (i.e., weak convergence). "Dividing" by the asymptotic standard deviation yields an asymptotically pivotal statistic[1]

$$\widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}(0, I)$$

where $\widehat{\Sigma}^{-1/2}$ is the unique matrix satisfying $\widehat{\Sigma}^{-1/2}\widehat{\Sigma}^{-1/2} = \widehat{\Sigma}^{-1}$ and

classic asymptotique estimator

$$\boxed{\widehat{\Sigma}} = \left(\overline{ZZ^T}\right)^{-1} \left(\overline{ZZ^T\widehat{\varepsilon}^2}\right) \left(\overline{ZZ^T}\right)^{-1} \text{ with } \widehat{\varepsilon}_i = Y_i - Z_i^T\widehat{\beta} .$$

$\widehat{\Sigma}$ is called "Heteroskedasticity Consistent Covariance matrix Estimator", in the sense that it estimates consistently the asymptotic variance of $\widehat{\beta}$ with and without heteroskedasticity, and it hence allows a correct inference on β_0. Such an estimator is also known as "Huber-White" estimator, from the names of the authors who proposed it. Despite this robusteness feature, a problem with such an estimator is that is might behave poorly in finite samples. The jackknife provides a valuable alternative. From Equation (1.4), the Jackknife variance estimator in this context writes as

$$\frac{n-1}{n} \sum_{i=1}^{n} \left[\widehat{\beta}_i - \frac{1}{n} \sum_{j=1}^{n} \widehat{\beta}_j \right]^2$$

[1] Recall that a statistic is asymptotically pivotal when its asymptotic distribution does not depend on unknown parameters.

$$P(|\hat{t}| > q_{1-\frac{\alpha}{2}}) = \mathbb{E}[\mathbb{1}_{\{|\hat{t}| > q_{1-\frac{\alpha}{2}}\}}]$$

$$\simeq \frac{1}{H} \cdot \sum_{h=1}^{H} \mathbb{1}_{\{|\hat{t}_h| > q_{1-\frac{\alpha}{2}}\}}$$

$$v_{jack} = \frac{n-1}{n} \cdot \sum_{i=1}^{n} [T_{n-1,i} - \bar{T}_n]^2 , \quad \bar{T}_n = \frac{1}{n} \sum_{i=1}^{n} T_{n-1,i}$$

$$\hat{t}_{jack} = \frac{\hat{\beta} - \beta_0}{\sqrt{v_{jack}}} \qquad \hat{t}_{OLS} = \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{\Sigma}}}$$

$$\hat{\theta} = \frac{1}{n} \Sigma$$

$$\sigma^2 \boxed{Var(\hat{\theta})} = \frac{1}{n} \cdot Var(x) \qquad n \times \sigma^2$$

$$\frac{n-1}{n} \cdot (n-1) \cdot Cov(\tilde{T})$$

$$\sigma_n^2$$

where $\widehat{\beta}_j$ represents the OLS estimator based on all observations except the jth. Then, inference on β_0 can be based on

$$\frac{\widehat{\beta} - \beta_0}{\sqrt{\upsilon_{jack}}} .$$

Inference when the variance has a difficult expression. The jackknife can also be employed when the asymptotic variance has a difficult expression. Several examples are reported in Shao and Tu (1995), Chapter 2.

Bias-improved 2SLS estimator. Two-stages least squares estimators (2SLS) are used in quasi-experimental studies to estimate treatment effects with imperfect compliance, i.e. Local Average Treatment Effects (LATE). Consider a linear model with endogenous regressors

$$Y = Z^T \beta_0 + \varepsilon \text{ with } \mathbb{E} W\varepsilon = 0 \text{ and } \mathbb{E} Z\varepsilon \neq 0 \,,$$

where Y is a scalar, Z, W are two vectors and ε is unobserved. W is an instrumental variable. The coefficient β_0 is identified as

$$\beta_0 = \left[(\mathbb{E}ZW^T)(\mathbb{E}WW^T)^{-1}(\mathbb{E}WZ^T)\right]^{-1}(\mathbb{E}ZW^T)(\mathbb{E}WW^T)^{-1}(\mathbb{E}WY)$$

and the 2SLS estimator is

$$\widehat{\beta} = \left[(\overline{ZW^T})(\overline{WW^T})^{-1}(\overline{WZ^T})\right]^{-1}(\overline{ZW^T})(\overline{WW^T})^{-1}(\overline{WY}) \,.$$

Differently from the OLS, the 2SLS estimator is biased. The jackknife can be used to reduce its bias. The bias formula in (1.1) adapted to $\widehat{\beta}$ is

$$b_{jack} := \frac{n-1}{n} \sum_{j=1}^{n} \left(\widehat{\beta}_j - \widehat{\beta}\right)$$

where $\widehat{\beta}_j$ denotes the 2SLS estimtor that uses all the observations except the jth. The bias-corrected 2SLS estimator based on the jackknife is then $\widehat{\beta} - b_{jack}$.

1.3 Consistency results

In this section we provide the consistency results for υ_{jack} and b_{jack}. We will assume that the population parameter θ takes the following form

$$\theta = g(\mu) \text{ where } \mu = \mathbb{E}\{X\}$$

and g is a fixed function. The estimator T_n is a function of the sample mean

$$T_n := g(\overline{X}_n) \ , \text{ where } \overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

Assumption A. *$\mathbb{E}XX^T$ is well defined and finite.*

Assumption B. *g is twice continuously differentiable in a neighborhood of μ.*

1.3.1 Consistency of the Jackknife variance estimator

Before turning to the consistency of υ_{jack} we show how such an estimator can be used to construct confidence intervals about θ. Under Assumption A and B, using a second-order Taylor expansion (see Proposition A.0.11) yields that with probability approaching one (wpa1)

$$\begin{aligned} T_n - \theta =& \nabla g(\mu)^T(\overline{X}_n - \mu) \\ &+ \frac{1}{2}(\overline{X}_n - \mu)^T \nabla^2 g(\mu)(\overline{X}_n - \mu) \\ &+ \frac{1}{2}(\overline{X}_n - \mu)^T \left[\nabla^2 g(\xi_n) - \nabla^2 g(\mu)\right](\overline{X}_n - \mu) \end{aligned} \tag{1.5}$$

$\sqrt{n}(\bar{x}_n - \mu) \to N(0, \Sigma)$

$\sqrt{n}(\bar{x}_n - \mu) = O_P(1)$

with $\xi_n \in [\overline{X}_n, \mu]$. By the imposed assumptions, the CLT yields that $\overline{X}_n - \mu = \underline{O_P(n^{-1/2})}$. bounded in probability By the LLN and the Continuous Mapping Theorem $\nabla^2 g(\xi_n) - \nabla^2 g(\mu) = o_P(1)$. Thus, the above display ensures that

$$\sqrt{n}(T_n - \theta) = \boxed{\nabla g(\mu)^T \sqrt{n}(\overline{X}_n - \mu) + o_P(1)} \rightsquigarrow \mathcal{N}(0, V), \text{ with } V := \nabla g(\mu)^T Var(X) \nabla g(\mu)$$

and

$$\frac{T_n - \theta}{\sigma_n} \rightsquigarrow \mathcal{N}(0,1), \text{ where } \sigma_n^2 := n^{-1}\nabla g(\mu)^T Var(X)\nabla g(\mu) \ . \tag{1.6}$$

If υ_{jack} is consistent for σ_n^2, i.e. $\upsilon_{jack}/\sigma_n^2 \xrightarrow{P} 1$, then by Slutsky's theorem

$$\frac{T_n - \theta}{\sqrt{\upsilon_{jack}}} = \frac{\sigma_n}{\sqrt{\upsilon_{jack}}} \cdot \frac{T_n - \theta}{\sigma_n} \rightsquigarrow \mathcal{N}(0,1) \ .$$

Hence confidence intervals for θ can be easily constructed. The following Theorem shows the consistency of υ_{jack}.

Proposition 1.3.1. *Under Assumptions A and B*

$$\frac{\upsilon_{jack}}{\sigma_n^2} \xrightarrow{P} 1$$

Proof. We first prove some preliminary results that will be used multiple times in the present proof. Let $\overline{X}_{n-1,i} = \sum_{j\neq i} X_j/(n-1)$. Notice that

$$\begin{aligned}\overline{X}_{n-1,i} - \overline{X}_n &= \left(\frac{1}{n-1} - \frac{1}{n}\right)\sum_j X_j - \frac{X_i}{n-1}\\ &= \frac{\overline{X}_n}{n-1} - \frac{X_i}{n-1}\ . \end{aligned} \tag{1.7}$$

So, by Assumption A (see the comments below)

$$\begin{aligned}(n-1)\sum_i ||\overline{X}_{n-1,i} - \overline{X}_n||^2 &= \frac{1}{n-1}\sum_i ||\overline{X}_n - X_i||^2\\ &= \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^T(X_i - \overline{X}_n)\\ &= \text{Trace}\left(\frac{1}{n-1}\sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right)\\ &= \text{Trace}\,(\text{Var}(X)) + o_P(1)\ , \end{aligned} \tag{1.8}$$

where the third equality has used the fact that $a^Ta = \text{Trace}(a^Ta) = \text{Trace}(a\ a^T)$ for any vector a.

We now enter the core of the theorem. The LLN ensures that $||\overline{X}_n - \mu|| = o_P(1)$. By the above display, $\max_{i=1,..,n} ||\overline{X}_{n-1,i} - \overline{X}_n||^2 \leq \sum_{i=1}^n ||\overline{X}_{n-1,i} - \overline{X}_n||^2 = O_P(n^{-1})$. This implies that $\max_{i=1,..,n} ||\overline{X}_{n-1,i} - \mu|| = o_P(1)$. Hence, Assumption B and a 1st order Taylor expansion yield that wpa1

$$\begin{aligned}T_{n-1,i} - T_n =& g(\overline{X}_{n-1,i}) - g(\overline{X}_n)\\ =& \nabla g(\overline{X}_n)^T(\overline{X}_{n,i} - \overline{X}_n) + R_{n,i}\,,\\ \text{where } R_{n,i} :=& \left[\nabla g(\xi_{n,i}) - \nabla g(\overline{X}_n)\right]^T(\overline{X}_{n-1,i} - \overline{X}_n) \end{aligned} \tag{1.9}$$

and $\xi_{n,i} \in [\overline{X}_{n-1,i}, \overline{X}_n]$. By Equation (1.7) $\sum_i(\overline{X}_{n-1,i} - \overline{X}_n) = 0$, so using (1.9)

$$\frac{1}{n}\sum_i (T_{n,i} - T_n) = \frac{1}{n}\sum_i R_{n,i} =: \overline{R}_n \tag{1.10}$$

We thus obtain

$$\begin{aligned}
v_{jack} =& \frac{n-1}{n}\sum_i \left[T_{n-1,i} - \frac{1}{n}\sum_j T_{n-1,j}\right]^2 \text{ (by (1.4))} \\
=& \frac{n-1}{n}\sum_i \left[T_{n-1,i} - T_n - \frac{1}{n}\sum_j (T_{n,j} - T_n)\right]^2 \\
=& \frac{n-1}{n}\sum_i \left[\nabla g(\overline{X}_n)^T(\overline{X}_{n-1,i} - \overline{X}_n) + R_{n,i} - \overline{R}_n\right]^2 \text{ (by (1.10) and (1.9))} \\
=& \frac{n-1}{n}\nabla g(\overline{X}_n)^T \cdot \sum_i (\overline{X}_{n-1,i} - \overline{X}_n)(\overline{X}_{n-1,i} - \overline{X}_n)^T \cdot \nabla g(\overline{X}_n) \\
&+ \frac{n-1}{n}\sum_i (R_{n,i} - \overline{R}_n)^2 \\
&+ 2\frac{n-1}{n}\sum_i (R_{n,i} - \overline{R}_n)(\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla g(\overline{X}_n) =: A_n + B_n + 2C_n \ .
\end{aligned}$$

By the previous display, the result is proved if $A_n/\sigma_n^2 \xrightarrow{P} 1$, $B_n/\sigma_n^2 \xrightarrow{P} 0$, and $C_n/\sigma_n^2 \xrightarrow{P} 0$.

(I) We show that $A_n/\sigma_n^2 \xrightarrow{P} 1$.

By using the definition of A_n, the definition of σ_n^2, and (1.7) we obtain

$$\begin{aligned}
\frac{A_n}{\sigma_n^2} =& \frac{(n-1)\cdot \nabla g(\overline{X}_n)^T \cdot \sum_i (\overline{X}_{n-1,i} - \overline{X}_n)(\overline{X}_{n-1,i} - \overline{X}_n))^T \cdot \nabla g(\overline{X}_n)}{\nabla g(\mu)^T \mathrm{Var}(X) \nabla g(\mu)} \\
=& \frac{\nabla g(\overline{X}_n)^T \cdot (n-1)^{-1}\sum_i (\overline{X}_n - X_i)(\overline{X}_n - X_i)^T \cdot \nabla g(\overline{X}_n)}{\nabla g(\mu)^T \mathrm{Var}(X) \nabla g(\mu)} \text{ (by (1.7))} \\
=& 1 + o_P(1)
\end{aligned}$$

where the last line follows from $\nabla g(\overline{X}_n) = \nabla g(\mu) + o_P(1)$ and $(n-1)^{-1}\sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T = \mathrm{Var}(X) + o_P(1)$.

(II) We show that $B_n/\sigma_n^2 \xrightarrow{P} 0$.

By definition of B_n and σ_n^2 we get

$$\begin{aligned}
\frac{B_n}{\sigma_n^2} =& \frac{n-1}{\sigma_n^2}\cdot\frac{1}{n}\sum_i (R_{n,i} - \overline{R}_n)^2 \\
\leq& \frac{n-1}{\sigma_n^2}\cdot\frac{1}{n}\sum_i R_{n,i}^2 \\
\leq& \frac{n-1}{\nabla g(\mu)^T \mathrm{Var}(X) \nabla g(\mu)}\sum_i R_{n,i}^2 \\
\leq& \frac{n-1}{\nabla g(\mu)^T \mathrm{Var}(X) \nabla g(\mu)}\sum_i ||\nabla g(\xi_{n,i}) - \nabla g(\overline{X}_n)||^2 \ ||\overline{X}_{n-1,i} - \overline{X}_n||^2
\end{aligned}$$

where the last line follows from the definition of $R_{n,i}$ in (1.9) and the Cauchy-Schwartz inequality. From Equation (1.8), to prove the desired result it suffices to show that

$\max_{i=1,..,n} |||\nabla g(\xi_{n,i}) - \nabla g(\overline{X}_n)|| = o_P(1)$. To this end, notice that

$$\begin{aligned}\max_{i=1,..,n} ||\nabla g(\xi_{n,i}) - \nabla g(\overline{X}_n)||^2 \leq & 4 \cdot \max_{i=1,..,n} ||\nabla g(\xi_{n,i}) - \nabla g(\mu)||^2 \\ & + 4 \cdot ||\nabla g(\overline{X}_n) - \nabla g(\mu)||^2 .\end{aligned}$$

By the continuity of ∇g at μ, $\overline{X}_n = \mu + o_P(1)$, and the continuous mapping theorem, the last term on the RHS of the above display is $o_P(1)$. For the first term, using again the continuity of ∇g at μ, for an arbitrary $\epsilon > 0$ there exists a $\delta > 0$ such that if $||x - \mu|| < \delta$ then $||\nabla g(x) - \nabla g(\mu)|| < \epsilon$. Hence,

$$P\left(\max_{i=1,..,n} ||\nabla g(\xi_{n,i}) - \nabla g(\mu))|| \geq \epsilon\right) \leq P\left(\max_{i=1,..,n} ||\xi_{n,i} - \mu|| \geq \delta\right) . \tag{1.11}$$

Since $\xi_{n,i} \in [\overline{X}_{n-1,i}, \overline{X}_n]$ we have that $\max_{i=1,..,n} ||\xi_{n,i} - \mu|| \leq \max_{i=1,..,n} ||\overline{X}_{n-1,i} - \overline{X}_n|| + ||\overline{X}_n - \mu||$, with $||\overline{X}_n - \mu|| = O_P(n^{-1/2})$ (by the CLT) and $\max_{i=1,..,n} ||\overline{X}_{n-1,i} - \overline{X}_n||^2 \leq \sum_i ||\overline{X}_{n-1,i} - \overline{X}_n||^2 = O_P(n^{-1})$ (see Equation (1.8)). So,

$$\max_{i=1,...,n} \|\xi_{i,n} - \mu\| = O_P(n^{-1/2}) , \tag{1.12}$$

the RHS of (1.11) is $o(1)$, and we conclude.

(III) We prove that $C_n/\sigma_n^2 \xrightarrow{P} 0$.
From the definition of C_n and the Cauchy-Schwartz inequality

$$\begin{aligned}\left|\frac{C_n}{\sigma_n^2}\right| &= \frac{n-1}{n}\left|\sum_i \frac{(R_{n,i} - \overline{R}_n)}{\sigma_n} \frac{(\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla g(\overline{X}_n)}{\sigma_n}\right| \\ &\leq \left\{\sum_i \frac{(R_{n,i} - \overline{R}_n)^2}{\sigma_n^2}\right\}^{1/2} \cdot \left\{\sum_i \frac{\left[(\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla g(\overline{X}_n)\right]^2}{\sigma_n^2}\right\}^{1/2} \\ &\leq \left\{\frac{n}{n-1}\frac{B_n}{\sigma_n^2}\right\}^{1/2} \cdot \left\{\frac{n}{n-1}\frac{A_n}{\sigma_n^2}\right\}^{1/2} = o_P(1) ,\end{aligned}$$

where the last two equalities follow from the definitions of A_n and B_n, and the results of (I) and (II). □

1.3.2 Consistency of the Jackknife bias estimator

We start by defining the asymptotic bias of T_n. Consider the expansion in (1.5). Since $\xi_n \in [\overline{X}_n, \mu]$, by the LLN $\xi_n \xrightarrow{P} \mu$. So, under Assumption B we can apply the continuous mapping theorem and obtain that $\nabla^2 g(\xi_n) = \nabla^2 g(\mu) + o_P(1)$. Recall that by the CLT

$\overline{X}_n - \mu = O_P(n^{-1/2})$. Hence, the third term on the RHS of (1.5) is $o_P(n^{-1})$ and we have

$$\begin{aligned} T_n - \theta =& \nabla g(\mu)^T (\overline{X}_n - \mu) \\ &+ \frac{1}{2}(\overline{X}_n - \mu)^T \nabla^2 g(\mu)(\overline{X}_n - \mu) \\ &+ o_P\left(\frac{1}{n}\right) \ . \end{aligned} \tag{1.13}$$

Now, the first term on the RHS is $O_P(n^{-1/2})$ and has null expectation. The second term is of order $O_P(n^{-1})$ and its expectation is (see the comments below)

$$\begin{aligned} \mathbb{E}\left\{\frac{1}{2}(\overline{X}_n - \mu)^T \nabla^2 g(\mu)(\overline{X}_n - \mu)\right\} =& \frac{1}{2}\mathbb{E} \text{ Trace}\left\{(\overline{X}_n - \mu)^T \nabla^2 g(\mu)(\overline{X}_n - \mu)\right\} \\ =& \frac{1}{2}\mathbb{E} \text{ Trace}\left\{\nabla^2 g(\mu)(\overline{X}_n - \mu)(\overline{X}_n - \mu)^T\right\} \\ =& \frac{1}{2} \text{ Trace}\left(\nabla^2 g(\mu)\ \mathbb{E}\left\{(\overline{X}_n - \mu)(\overline{X}_n - \mu)^T\right\}\right) \\ =& \frac{1}{2n}\text{Trace}\left(\nabla^2 g(\mu)\ \text{Var}(X)\right) \ . \end{aligned}$$

Indeed, for any conformable matrix B and vector c we have $c^T Bc = Trace(c^T Bc) = Trace(Bcc^T)$, which gives the first and second equality. The third equality follows by commuting expectation and $Trace$. Finally, the last equality is ensures by the iid assumption.
Now, the last term on the RHS of (1.13) can be ignored as it is $o_P(n^{-1})$ and hence of smaller order than the other two terms on the RHS. So, *heuristically*

$$\mathbb{E}\{T_n\} - \theta \approx \frac{a}{n} \text{ with } a := \frac{1}{2} \text{ Trace}\left(\nabla^2 g(\mu)\ \text{Var}(X)\right) \ .$$

By the above display, we can define a/n as the *asymptotic bias* of T_n. At the end of this section we will show that $b_{jack} = a/n + o_P(n^{-1})$, so that the Jackknife bias estimator is consistent for the asymptotic bias of T_n. This rate implies that the bias-reduced Jackknife estimator T_{jack} has the same asymptotic distribution as T_n, i.e.

$$\begin{aligned} \sqrt{n}(T_{jack} - \theta) =& \sqrt{n}(T_n - \theta) + \sqrt{n} b_{jack} \\ =& \sqrt{n}(T_n - \theta) + \sqrt{n}\left(\frac{a}{n} + o_P(n^{-1})\right) \\ =& \sqrt{n}(T_n - \theta) + o_P(1) \ . \end{aligned}$$

In other words, the bias correction b_{jack} does not have an impact on the asymptotic behavior of T_n. Its advantage stands in the small-sample behavior of T_{jack}, in the sense that in finite sample T_{jack} will generally display a lower bias than T_n.[2]

[2] The *heuristic* argument for this is contained in Section 1.1.

Proposition 1.3.2. *Under Assumptions A and B*

$$n\ b_{jack} = a + o_P(1)$$

If moreover $\nabla^3 g$ is continuous in a neighborhood of μ and $\mathbb{E}||X||^4 < \infty$ then

$$n\ b_{jack} = a + O_P\left(n^{-1/2}\right)\ .$$

Proof. From the definition of b_{jack} in (1.1)

$$\begin{aligned} b_{jack} =&(n-1)\left(\frac{\sum_i T_{n-1,i}}{n} - T_n\right) \\ =&\frac{n-1}{n}\sum_i (T_{n-1,i} - T_n)\ . \end{aligned}$$

As obtained in the proof of Proposition 1.3.1 $\max_{i=1,..,n} ||\overline{X}_{n-1,i} - \mu|| \le \max_{i=1,..,n} ||\overline{X}_{n-1,i} - \overline{X}_n|| + \|\overline{X}_n - \mu\| = o_P(1)$. So, by Assumption B and a second order Taylor expansion (see Proposition A.0.11), wpa1

$$\begin{aligned} T_{n-1,i} - T_n = g(\overline{X}_{n-1,i}) - g(\overline{X}_n) =&\nabla g(\overline{X}_n)^T(\overline{X}_{n-1,i} - \overline{X}_n) \\ &+ \frac{1}{2}(\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla^2 g(\xi_{n,i})(\overline{X}_{n-1,i} - \overline{X}_n)\ , \end{aligned}$$

where $\xi_{n,i} \in [\overline{X}_{n-1,i}, \overline{X}_n]$. From (1.7) $\sum_i(\overline{X}_{n-1,i} - \overline{X}_n) = 0$. So, from the previous two displays

$$\begin{aligned} n\ b_{jack} =&(n-1)\sum_i (T_{n-1,i} - T_n) \\ =&\frac{n-1}{2}\sum_i (\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla^2 g(\xi_{n,i})(\overline{X}_{n-1,i} - \overline{X}_n) \\ =&\frac{1}{2(n-1)}\sum_i (X_i - \overline{X}_n)^T \nabla^2 g(\xi_{n,i})(X_i - \overline{X}_n)\ \text{(from (1.7))} \\ =&\frac{1}{2(n-1)}\sum_i \text{Trace}\left\{\nabla^2 g(\xi_{n,i})(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right\} \\ =&\frac{1}{2(n-1)}\sum_i \text{Trace}\left\{\left[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)\right](X_i - \overline{X}_n)((X_i - \overline{X}_n)^T\right\} \\ &+ \frac{1}{2(n-1)}\sum_i \text{Trace}\left\{\nabla^2 g(\mu)(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right\} \\ =:&A_n + B_n \end{aligned} \tag{1.14}$$

where for the fourth equality we have used the property that $\text{Trace}(c^T Bc) = \text{Trace}(Bcc^T)$

for any conformable matrix B and vector c. Now,

$$\begin{aligned} B_n =& \frac{1}{2}\text{ Trace}\left\{\nabla^2 g(\mu)\frac{1}{n-1}\sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right\} \\ =& \frac{1}{2}\text{ Trace}\left\{\nabla^2 g(\mu)\text{Var}(X)\right\} + o_P(1) \\ =& a + o_P(1)\ , \end{aligned}$$

where we have used $(n-1)^{-1}\sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T = \text{Var}(X) + o_P(1)$ and the definition of a. So, from the previous two displays to show the first part of the proposition it is sufficient to prove that $A_n = o_P(1)$. To this end, denote with $X_i^{(s)}$ and $\overline{X}_n^{(s)}$ the sth component of X_i and $\overline{X}_n$. Also, denote with $[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)]_{s,m}$ the element on the sth row and mth column of $[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)]$. Then, from the definition of A_n

$$\begin{aligned} |A_n| \le& \frac{2}{n-1}\sum_{i=1}^n \left|\text{Trace}\left\{\left[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)\right](X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right\}\right| \\ =& \frac{2}{n-1}\sum_{i=1}^n \left|\sum_{s,m}\left[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)\right]_{s,m}(X_i^{(m)} - \overline{X}_n^{(m)})(X_i^{(s)} - \overline{X}_n^{(s)})\right| \\ \le& \frac{1}{2(n-1)}\sum_{i=1}^n ||\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)||\sum_{s,m}\left|(X_i^{(s)} - \overline{X}_n^{(s)})(X_i^{(m)} - \overline{X}_n^{(m)})\right| \\ \le& \max_{i=1,..,n} ||\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)||\sum_{s,m}\frac{1}{2(n-1)}\sum_{i=1}^n\left|(X_i^{(s)} - \overline{X}_n^{(s)})(X_i^{(m)} - \overline{X}_n^{(m)})\right| \end{aligned} \tag{1.15}$$

where $||\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)||$ denotes the Forbenious norm of the matrix $\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)$. By proceeding similarly to Equation (1.11) and (1.12), we get that $\max_{i=1,..,n} ||\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)|| = o_P(1)$. Also, by Assumption A for each s, m the average on the RHS of the above display is $O_P(1)$. Thus, $|A_n| = o_P(1)$ and we have proved the first part of the proposition.

To prove the second part, notice that when $\mathbb{E}||X||^4 < \infty$ we can apply the CLT and have $(n-1)^{-1}\sum_i X_i X_i^T = \mathbb{E}XX^T + O_P(n^{-1/2})$, so that

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)(X_i-\overline{X}_n)^T =&\frac{1}{n-1}\sum_{i=1}^{n}X_iX_i^T-\frac{1}{n-1}\sum_{i=1}^{n}X_i\overline{X}_n^T\\ &-\overline{X}_n\frac{1}{n-1}\sum_{i=1}^{n}X_i^T+\frac{n}{n-1}\overline{X}_n\overline{X}_n^T\\ =&\mathbb{E}XX^T-\mu\mu^T+O_P\left(\frac{1}{\sqrt{n}}\right)\\ =&\text{Var}(X)+O_P\left(\frac{1}{\sqrt{n}}\right)\ .\end{aligned}$$

Using the above display and the definition of B_n in (1.14) we find that

$$B_n=\frac{1}{2}\text{Trace}\left\{\nabla^2 g(\mu)\text{Var}(X)\right\}+O_P(n^{-1/2})\ .$$

Hence, from (1.14) it suffices to show that $|A_n|=O_P(n^{-1/2})$. To this end, from the Equation (1.15) and the arguments below it, we only have to prove that $\max_{i=1,..,n}||\nabla^2 g(\xi_{n,i})-\nabla^2 g(\mu)||=O_P(n^{-1/2})$. Consider a single component of $\nabla^2 g(\xi_{n,i})-\nabla^2 g(\mu)$ and drop its matrix index for notational simplicity. Since by assumption $\nabla^3 g$ is continuous in a neighborhood of μ, by a first order Taylor expansion there exists a C and a $\delta>0$ such that

$$\left|\nabla^2 g(x)-\nabla^2 g(\mu)\right|\leq C||x-\mu||\text{ for all }x\in\text{Ball}(\mu,\delta)\ .$$

Recall that from (1.12) $\max_{i=1,..,n}||\xi_{n,i}-\mu||=O_P(n^{-1/2})$, so the above display implies that wpa1

$$\max_{i=1,..,n}||\nabla^2 g(\xi_{n,i})-\nabla^2 g(\mu)||\leq C\max_{i=1,..,n}||\xi_{n,i}-\mu||=O_P(n^{-1/2})$$

which delivers the desired result.

□

1.4 Extensions

In some cases the jackknife does not provide a consistent estimator of the variance, i.e. $\upsilon_{jack}/\sigma_n^2$ does not converge in probability to 1. One of such cases is the estimation of the variances of the sample quantiles, e.g. the variance of the sample median. This lack of consistency is due to the lack of smoothness of the function g, see Shao and Tu (1995). This should not be surprising: the smoothness of g has been heavily used in the proof of Proposition 1.3.1. A valid alternative to the Jackknife in these situations is the *deleted d jackknife*. This can be seen as a generalization of the Jackknife studied in the previous pages, where instead of removing one observation we remove d observations and recompute the statistic on the remaining sample. In particular, let $\mathbf{s}$ be a subset of size $n-d$ of the indices $\{1,\dots,n\}$ and let $T_{n-d,\mathbf{s}}$ be the statistic computed with such observations. The *deleted-d jackknife variance estimator* is given by

$$\upsilon_{jack-d} := \frac{n-d}{d}\frac{1}{N}\sum_{\mathbf{s}\in\mathcal{S}}\left[T_{n-d,\mathbf{s}} - \frac{1}{N}\sum_{\mathbf{s}\in\mathcal{S}}T_{n-d,\mathbf{s}}\right]^2 ,$$

where $\mathcal{S}$ represents the collection of all subgroup of $\{1,\dots,n\}$ of size $n-d$, where each group differs from the others by at least one element, and $N = \binom{n}{n-d}$ is the number of elements in $\mathcal{S}$. In combinatorial calculus $\mathcal{S}$ is also called collection of combinations of $\{1,\dots,n\}$ of class $n-d$. Notice that when $d=1$, υ_{jack-d} corresponds to the jackknife studied in the previous pages. The asymptotic properties of υ_{jack-d} are obtained in Shao and Tu (1995) to whom we refer the reader. We will see in the following chapters that such a principle can be used also to estimate the distribution of a statistic, and it has a close connection with the *subsampling method*.

The jackknife principle can also be used to estimate the variance of a functional, where $T_n = g(F_n)$ and F_n is the empirical cdf. Notice that the empirical cdf is a function $x \mapsto (1/n)\sum_i 1(X_i \leq x)$, so g maps such functions into $\mathbb{R}$. More details can be found in Shao and Tu (1995).

Notes

The first author to propose the jackknife method for bias estimation has been Quenouille (1949). Some time later, Tukey (1958) realized that the same principle could be used to estimate the variance of a statistic and popularized the Jackknife.

The presentation given in this chapter follows closely Chapter 2 of Shao and Tu (1995) who provide a very comprehensive and detailed theory of the Jackknife.

MacKinnon and White (1985) have been the first to propose the Jackknife for estimating the variance of the OLS estimator in the presence of heteroskedasticity. They provide a large simulation study showing that it outperforms the popular White-Huber estimator.

An idea similar to the Jackknife has been proposed by Angrist, Imbens, and Krueger (1999) for bias correcting the 2SLS estimator. They label their method as JIVE (Jackknife Instrumental Variable Estimator), although it is quite different from a Jackknife bias corrected estimator. Later, Hahn, Hausman, and Kuersteiner (2004) proposed a bias corrected jackknife estimator for the 2SLS, as that seen in Section 1.2. They provide evidence about the finite sample performances of such an estimator and show that in practice it outperforms several estimators for the linear model with endogenous regressors.

Davidson and MacKinnon (2006) also show empirical evidence about several estimators for the IV model.