Advanced Microeconometrics, Autumn 2021: The Delta Method

Anders Munk-Nielsen   Jesper Riis–Vestergaard Sørensen

November 15, 2021

1 Motivation

Suppose that an estimator $\widehat{\boldsymbol{\theta}}_n$ for a parameter $\boldsymbol{\theta}_0$ is available, but the quantity of interest is $\mathbf{h}(\boldsymbol{\theta}_0)$ for some known function $\mathbf{h}$. A natural estimator of $\mathbf{h}(\boldsymbol{\theta}_0)$ is $\mathbf{h}(\widehat{\boldsymbol{\theta}}_n)$, sometimes referred to as the *plug-in estimator*. How do the asymptotic properties of $\mathbf{h}(\widehat{\boldsymbol{\theta}}_n)$ follow from those of $\widehat{\boldsymbol{\theta}}_n$?

A first result is an immediate consequence of a continuous-mapping theorem (C&T Theorem A.3). If the sequence $\widehat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}_0$ as $n \to \infty$ and $\mathbf{h}$ is continuous at $\boldsymbol{\theta}_0$, then $\mathbf{h}(\widehat{\boldsymbol{\theta}}_n)$ converges in probability to $\mathbf{h}(\boldsymbol{\theta}_0)$.

Of greater interest is a similar question concerning limiting distributions as these facilitate inference. In particular, if $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution as $n \to \infty$, is the same true for $\sqrt{n}\{\mathbf{h}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{h}(\boldsymbol{\theta}_0)\}$? If so, what is the relationship between their distributional limits?

2 Delta Method

Suppose for simplicity that $\widehat{\boldsymbol{\theta}}_n$ is a sequence $\widehat{\theta}_n$ of random scalars and that $\mathbf{h}$ is a scalar function $h : \mathbf{R} \to \mathbf{R}$. Consider the random scalar $\sqrt{n}\{h(\widehat{\theta}_n) - h(\theta_0)\}$, which we may write as the product

$$\sqrt{n}\{h(\widehat{\theta}_n) - h(\theta_0)\} = \frac{h(\widehat{\theta}_n) - h(\theta_0)}{\widehat{\theta}_n - \theta_0} \cdot \sqrt{n}(\widehat{\theta}_n - \theta_0).$$

Provided that $\widehat{\theta}_n$ is consistent and h differentiable at θ_0, one may show that the difference quotient converges in probability to the derivative $h'(\theta_0)$, i.e.

$$\frac{h(\widehat{\theta}_n) - h(\theta_0)}{\widehat{\theta}_n - \theta_0} \xrightarrow{p} h'(\theta_0).$$

Suppose that, in addition, $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converges in distribution to some limit Z (a scalar random variable). Since $h'(\theta_0)$ is constant, it follows from the product rule (CT Theorem A.12) that

$$\frac{h(\widehat{\theta}_n) - h(\theta_0)}{\widehat{\theta}_n - \theta_0} \cdot \sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} h'(\theta_0)\, Z.$$

Summing up,

$$\sqrt{n}\{h(\widehat{\theta}_n) - h(\theta_0)\} \xrightarrow{d} h'(\theta_0)\, Z.$$

This result is the simplest form of what is usually referred to as the *delta method*.

While the above result has nothing to do with normality, often Z *is* normally distributed. Consider therefore the case of $Z \sim \mathrm{N}(0, \sigma_0^2)$ with σ_0^2 denoting the limit variance. The fact that the normal family is closed under affine transformations (i.e. an affine transformation of a normally distributed random variable is normally distributed) then implies $h'(\theta_0)\, Z \sim \mathrm{N}(0, [h'(\theta_0)]^2\, \sigma_0^2)$. Hence, when the delta method is justified, in the case of a normal limit,

$$\sqrt{n}\{h(\widehat{\theta}_n) - h(\theta_0)\} \xrightarrow{d} \mathrm{N}\big(0, [h'(\theta_0)]^2\, \sigma_0^2\big).$$

Provided we can also consistently estimate the limit variance $[h'(\theta_0)]^2\, \sigma_0^2 =: v_0^2$, the delta method therefore provides a vehicle for constructing confidence intervals for $h(\theta_0)$.

Specifically, one may show that if both

1. $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathrm{N}(0, \sigma_0^2)$ for some $\sigma^2 \in \mathbf{R}_{++}$,

2. h is continuously differentiable at θ_0 with nonzero derivative $h'(\theta_0)$, and

3. $\widehat{\sigma}_n^2$ is a sequence of variance estimators consistent for σ_0^2,

then the random interval

$$\left[h(\widehat{\theta}_n) - 1.96 \cdot \frac{\widehat{v}_n}{\sqrt{n}}, h(\widehat{\theta}_n) + 1.96 \cdot \frac{\widehat{v}_n}{\sqrt{n}}\right]$$

with variance estimator given by

$$\widehat{v}_n^2 := [h'(\widehat{\theta}_n)]^2 \widehat{\sigma}_n^2$$

is an asymptotically valid 95 pct. confidence interval for $h(\theta_0)$. This claim follows from repeated applications of CT Theorems A.3 and A.12, a straightforward but tedious exercise, which we encourage you to carry out. Of course, asymptotic $100 \cdot (1 - \alpha)$ pct. confidence for some other $\alpha \in (0, 1)$ may be obtained by replacing 1.96 with the $(1 - \alpha/2)$ quantile of the standard normal distribution.

The delta method is not limited to the scalar case. One may show that if $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to $\mathbf{Z}$, a random element of $\mathbf{R}^K$, and the function $\mathbf{h} : \mathbf{R}^K \to \mathbf{R}^L$ is differentiable at $\boldsymbol{\theta}_0$, then

$$\sqrt{n}\{\mathbf{h}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{h}(\boldsymbol{\theta}_0)\} \xrightarrow{d} \nabla \mathbf{h}(\boldsymbol{\theta}_0)\, \mathbf{Z},$$

where $\nabla \mathbf{h}(\boldsymbol{\theta}_0)$ denotes the gradient of $\mathbf{h}$ at $\boldsymbol{\theta}_0$ [i.e. the $L \times K$ matrix of partial derivatives $(\partial/\partial\theta_k)\, h_\ell(\boldsymbol{\theta}_0)$]. If, in addition, $\mathbf{Z}$ is $\mathrm{N}(\mathbf{0}_{K\times 1}, \boldsymbol{\Sigma}_0)$ distributed for some $K \times K$ variance matrix $\boldsymbol{\Sigma}_0$, then the affine transformation $\nabla \mathbf{h}(\boldsymbol{\theta}_0)\, \mathbf{Z}$ is distributed $\mathrm{N}\left(\mathbf{0}_{L\times 1}, \nabla \mathbf{h}(\boldsymbol{\theta}_0)\, \boldsymbol{\Sigma}_0 \nabla \mathbf{h}(\boldsymbol{\theta}_0)'\right)$. Hence, when the (multivariate) delta method is justified, in the case of a normal limit, we get the simplification

$$\sqrt{n}\{\mathbf{h}(\widehat{\boldsymbol{\theta}}_n) - \mathbf{h}(\boldsymbol{\theta}_0)\} \xrightarrow{d} \mathrm{N}\left(\mathbf{0}_{L\times 1}, \nabla \mathbf{h}(\boldsymbol{\theta}_0)\, \boldsymbol{\Sigma}_0 \nabla \mathbf{h}(\boldsymbol{\theta}_0)'\right).$$

Provided we can consistently estimate the limit variance $\nabla \mathbf{h}(\boldsymbol{\theta}_0)\, \boldsymbol{\Sigma}_0 \nabla \mathbf{h}(\boldsymbol{\theta}_0)' =: \mathbf{V}_0$—now a matrix—the delta method therefore provides a vehicle for constructing confidence intervals for the *elements of* $\mathbf{h}(\boldsymbol{\theta}_0)$ or some combination thereof.

Specifically, one may show that if both

1. $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_d \mathrm{N}(\mathbf{0}_{k\times 1}, \boldsymbol{\Sigma}_0)$ for some $\boldsymbol{\Sigma}_0 \in \mathbf{R}^{K\times K}$ symmetric positive definite,
2. $\mathbf{h}$ is continuously differentiable at $\boldsymbol{\theta}_0$ with derivative $\nabla \mathbf{h}(\boldsymbol{\theta}_0)$ of full rank, and
3. $\widehat{\boldsymbol{\Sigma}}_n$ is a sequence of variance estimators consistent for $\boldsymbol{\Sigma}_0$,

then each random interval

$$\left[h_\ell(\widehat{\boldsymbol{\theta}}_n) - 1.96 \cdot \sqrt{\frac{\widehat{V}_{n,\ell\ell}}{n}}, h_\ell(\widehat{\boldsymbol{\theta}}_n) + 1.96 \cdot \sqrt{\frac{\widehat{V}_{n,\ell\ell}}{n}} \right],$$

with $\widehat{V}_{n,\ell\ell}$ denoting the ℓth diagonal element of the variance estimator $\widehat{\mathbf{V}}_n$ given by the sandwich formula

$$\widehat{\mathbf{V}}_n := \nabla \mathbf{h}(\widehat{\boldsymbol{\theta}}_n) \widehat{\boldsymbol{\Sigma}}_n \nabla \mathbf{h}(\widehat{\boldsymbol{\theta}}_n)',$$

is an asymptotically valid 95 pct. confidence interval for $h_\ell(\theta_0)$.

3 Probit Marginal Effects

Often we formulate a model for the conditional mean $\mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}]$ of an outcome Y as a function of regressor realizations $\mathbf{x}$. One object which may be of interest is the *marginal* *effect* of some regressor X_k at a particular point $\mathbf{x}_0$. In the case of a continuous conditioning variable, when speaking of the "marginal effect at $\mathbf{x}_0$" we mean the partial derivative

$$\mathrm{ME}_k(\mathbf{x}_0) := \frac{\partial}{\partial x_k} \mathrm{E}[Y|\, \mathbf{X} = \mathbf{x}_0].$$

(For discrete regressors, simply replace differentials with differences in conditional expectations.) Via differentiation, our model for the conditional mean implies a model for the marginal effect(s). When $\mathrm{E}[Y|\, \mathbf{X} = \mathbf{x}]$ is linear in $\mathbf{x}$, marginal effects are given by the model parameters and, thus, constant. In nonlinear models, this relationship is more complicated. Moreover, $\mathrm{ME}_k(\mathbf{x}_0)$ will generally depend on the point $\mathbf{x}_0$ in a nontrivial manner. However, for a fixed point of interest $\mathbf{x}_0$, $\mathrm{ME}_k(\mathbf{x}_0)$ is given by a potentially complicated—but known—function of model parameters.

To illustrate this point, consider the probit model, which models a binary outcome Y as

$$Y = \mathbf{1}(\mathbf{X}'\boldsymbol{\beta}_0 + \varepsilon > 0),$$

where ε is distributed $\mathrm{N}(0,1)$ independently of regressors $\mathbf{X}$. With $\Phi(t) := \int_{-\infty}^{t} (2\pi)^{-1/2} \mathrm{e}^{-u^2/2} \mathrm{d}u$ denoting the standard normal cumulative distribution function (CDF), we obtain

$$\mathrm{P}(Y = 1|\, \mathbf{X} = \mathbf{x}) = \Phi(\mathbf{x}'\boldsymbol{\beta}_0),$$

where we have invoked independence of ε and $\mathbf{X}$, normality of ε, and symmetry of the standard normal distribution about zero. (Check it!) By virtue of Y being binary, this is also the conditional mean, so

$$\mathrm{E}[Y|\, \mathbf{X} = \mathbf{x}] = \Phi(\mathbf{x}'\boldsymbol{\beta}_0).$$

Writing $\varphi(t) = (2\pi)^{-1/2} \mathrm{e}^{-t^2/2}$ for the standard normal probability density function (PDF), by means of the chain rule for differentiation we see that

$$\mathrm{ME}_k(\mathbf{x}_0) = \varphi(\mathbf{x}_0'\boldsymbol{\beta}_0)\, \beta_{0j}.$$

The right-hand side, hence the marginal effect of X_j, depends on both model parameters and the point of evaluation in a nonlinear manner.

Suppose now that we want a (say, 95 pct.) confidence interval for $\mathrm{ME}_k(\mathbf{x}_0) = \varphi(\mathbf{x}_0'\boldsymbol{\beta}_0)\, \beta_{0k}$. The right-hand side is the kth coordinate of the function $\mathbf{h} : \mathbf{R}^K \to \mathbf{R}^K$ defined by

$$\mathbf{h}(\boldsymbol{\beta}) := \varphi(\mathbf{x}_0'\boldsymbol{\beta})\, \boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbf{R}^K,$$

evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Inspection shows that φ is differentiable with derivative given by $\varphi'(t) = -t\varphi(t)$. (Do it!) It thus follows from the product and chain rules for differentiation that the function $\mathbf{h}$ is differentiable with gradient

$$\nabla\mathbf{h}(\boldsymbol{\beta}) = \varphi(\mathbf{x}_0'\boldsymbol{\beta})\left[\mathbf{I}_K - (\mathbf{x}_0'\boldsymbol{\beta})\boldsymbol{\beta}\mathbf{x}_0'\right] \in \mathbf{R}^{K\times K}.$$

(Check it!) The right-hand side is element-by-element continuous in $\boldsymbol{\beta}$, so $\mathbf{h}$ is everywhere continuously differentiable, hence continuously differentiable at $\boldsymbol{\beta}_0$.

Suppose now that we have access to an independent sample of n pairs $(Y_i, \mathbf{X}_i)$ and estimate $\boldsymbol{\beta}_0$ using maximum likelihood. Then, under certain conditions, the maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\beta}}_n$ satisfies $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}_{K\times 1}, \boldsymbol{\Sigma}_0)$ with a consistently estimable limit variance $\boldsymbol{\Sigma}_0$ (see lecture slides for details).

The delta method therefore applies to show that the $K \times 1$ vector $\widehat{\mathbf{ME}}(\mathbf{x}_0) := \mathbf{h}(\widehat{\boldsymbol{\beta}}_n)$ of estimators of marginal effects $\mathbf{ME}(\mathbf{x}_0) = \mathbf{h}(\boldsymbol{\beta}_0)$ satisfies

$$\sqrt{n}\{\widehat{\mathbf{ME}}(\mathbf{x}_0) - \mathbf{ME}(\mathbf{x}_0)\} \xrightarrow{d} \mathrm{N}(\mathbf{0}_{K\times 1}, \mathbf{V}_0)$$

where the limit variance takes the form (check!):

$$\mathbf{V}_0 = \varphi^2(\mathbf{x}_0'\boldsymbol{\beta}_0)\left[\mathbf{I}_K - (\mathbf{x}_0'\boldsymbol{\beta}_0)\boldsymbol{\beta}_0\mathbf{x}_0'\right]\boldsymbol{\Sigma}_0\left[\mathbf{I}_K - (\mathbf{x}_0'\boldsymbol{\beta}_0)\boldsymbol{\beta}_0\mathbf{x}_0'\right]'.$$