The Delta Method

1 Introduction

Econometricians often want to perform inference on nonlinear functions of model parameters. This requires them to estimate the standard error of a nonlinear function of parameter estimates or, more generally, the covariance matrix of a vector of such functions. One popular way to do so is called the delta method. It is based on asymptotic approximation.

For simplicity, let us start with the case of a single parameter (but this can be extended to a vector case). Suppose that we have estimated parameter θ, which might be one of the coefficients of a linear regression model, and that we are interested in the parameter $\gamma = g(\theta)$, where $g(\cdot)$ is a monotonic function that is continuously differentiable. In this situation, the obvious way to estimate γ is to use $\hat{\gamma} = g(\hat{\theta})$. Since $\hat{\theta}$ is a random variable, so is $\hat{\gamma}$. The problem is to estimate the variance of $\hat{\gamma}$.

Since $\hat{\gamma}$ is a function of $\hat{\theta}$, it seems logical that $\mathrm{Var}(\hat{\gamma})$ should be a function of $\mathrm{Var}(\hat{\theta})$. If $g(\theta)$ is a linear or affine function, then we already know how to calculate $\mathrm{Var}(\hat{\gamma})$[1]. The idea of the delta method is to find a linear approximation to $g(\theta)$, and then apply our rule for linear/affine functions.

2 Taylor's Theorem

We use Taylor's Theorem to obtain linear approximations to nonlinear functions. In its simplest form, Taylor's Theorem applies to functions of a scalar argument that are differentiable at least once on some real interval $[a, b]$, with the derivative

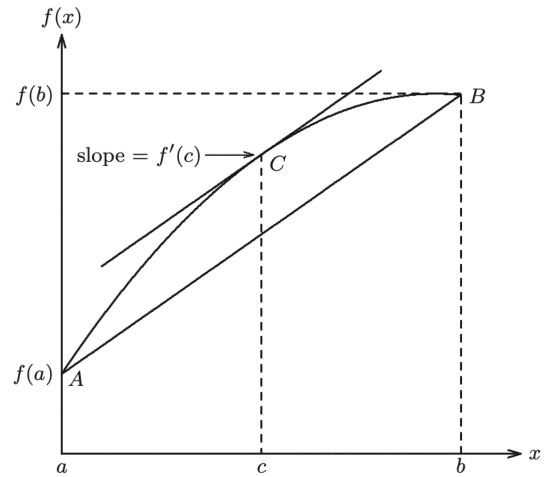
[1] The covariance matrix of $\hat{\boldsymbol{\beta}}$, say, can be used to calculate the variance of any linear (strictly speaking, affine) function of $\hat{\boldsymbol{\beta}}$. Suppose that we are interest the variance of $\hat{\phi}$, where $\phi = \mathbf{w}^\top \boldsymbol{\beta}$, $\hat{\phi} = \mathbf{w}^\top \hat{\boldsymbol{\beta}}$, and $\mathbf{w}$ is a k-vector of known coefficients. Since

$$\mathrm{Var}(\mathbf{b}) = \mathbb{E}\left[(\mathbf{b} - \mathbb{E}[\mathbf{b}])(\mathbf{b} - \mathbb{E}[\mathbf{b}])^\top\right],$$

we have

$$\begin{aligned}
\mathrm{Var}(\mathbf{w}^\top \hat{\boldsymbol{\beta}}) &= \mathbb{E}\left[\mathbf{w}^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbf{w}\right] \\
&= \mathbf{w}^\top \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\right] \mathbf{w} \\
&= \mathbf{w}^\top (\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}) \mathbf{w} \\
&= \mathbf{w}^\top \mathrm{Var}(\hat{\boldsymbol{\beta}}) \mathbf{w} \\
&= \mathrm{Var}(\mathbf{w}^\top \hat{\boldsymbol{\beta}}) = \mathrm{Var}(\hat{\phi}).
\end{aligned}$$

Figure 1: **Taylor's Theorem**



a continuous function on $[a, b]$. The following figure shows the graph of such a function, $f(x)$, for $x \in [a, b]$.

The coordinates of A are $(a, f(a))$ and the coordinates of B are $(b, f(b))$. Thus, the slope of the line AB is $(f(b) - f(a))/(b - a)$. What drives the theorem is the observation that there must always be a value between a and b, like c in the figure, at which the derivative $f'(c)$ is equal to the slope of AB. This is a consequence of the continuity of the derivative. If it were not continuous, and the graph of $f(x)$ had a corner, the slope might always be greater than $f'(c)$ on one side of the corner, and always be smaller on the other. But if $f'(x)$ is continuous on $[a, b]$, then there must exist c such that:

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

This can be rewritten as

$$f(b) = f(a) + (b - a)f'(c).$$

If we let $h = b - a$, then, since c lies between a and b, it must be the case that $c = a + th$, for some $t \in (0, 1)$. Thus we obtain:

$$f(a + h) = f(a) + hf'(a + th), \tag{1}$$

which is the simplest expression for Taylor's Theorem.

It is more usual to just set $t = 0$, so as to obtain a linear approximation to the function $f(x)$ for x in the neighbourhood of a. This approximation, called a first-order Taylor expansion about a, is:

$$f(a+h) \approx f(a) + hf'(a), \tag{2}$$

where the RHS of this equation is an affine function of h.

Taylor's Theorem can be extended in order to provide approximations that are quadratic or cubic functions, or polynomials of any desired order. The exact statement of the theorem, with terms proportional to powers of h up to h^p is:

$$f(a+h) = f(a) + \sum_{i=1}^{p-1} \frac{h^i}{i!} f^{(i)}(a) + \frac{h^p}{p!} f^{(p)}(a+th). \tag{3}$$

Here, $f^{(i)}$ is the i-th derivative of f, and once more $t \in (0,1)$. The approximate version of the theorem sets $t = 0$, and gives rise to a p-th order Taylor expansion about a. A common example in macroeconomics of the latter is the second-order Taylor expansion:

$$f(a+h) \approx f(a) + hf'(a) + \frac{1}{2}h^2 f''(a).$$

Both versions of Taylor's Theorem require as a regularity condition that $f(x)$ should have a p-th derivative that is continuous on $[a, a+h]$.

There are also multivariate versions of Taylor's Theorem, and we will need them from time to time. If $f(\mathbf{x})$ is a now a scalar-valued function of the m-vector $\mathbf{x}$, then, for $p = 1$, Taylor's Theorem states that, if $\mathbf{h}$ is also an m-vector, then:

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \sum_{j=1}^{m} h_j f_j(\mathbf{x}+t\mathbf{h}), \tag{4}$$

where h_j is the j-th component of $\mathbf{h}$, f_j is the partial derivative of f with respect to its j-th argument, and, as before, $t \in (0,1)$.

3 Delta method for a scalar parameter

Let $\hat{\theta}$ be root-n consistent[2] and asymptotically normal, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{a}{\sim} N(0, V^\infty(\hat{\theta})), \tag{5}$$

where $V^\infty(\hat{\theta})$ denotes the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$. In order to find the asymptotic distribution of $\hat{\gamma} = g(\hat{\theta})$, we perform a first-order Taylor expansion of $g(\hat{\theta})$ about θ_0. We get:

$$\hat{\gamma} = g(\hat{\theta}) \approx \underbrace{g(\theta_0)}_{\gamma_0} + g'(\theta_0)(\hat{\theta} - \theta_0), \tag{6}$$

[2] i.e. $\sqrt{n}(\hat{\theta} - \theta_0) = \mathrm{O}_p(1)$.

where $g'(\theta_0)$ is the first derivative of $g(\theta)$ evaluated at θ_0. Given the root-n consistency of $\hat{\theta}$, we can rearrange the above equation as an asymptotic equality. Two deterministic quantities are said to be asymptotically equal if they tend to the same limits as $n \to \infty$. Similarly, two random quantities are said to be asymptotically equal if they tend to the same limits in probability. As usual, we need a power of n to make things work correctly. Here, we multiply both sides by $\sqrt{n}$. If we denote $g(\theta_0)$, which is the true value of γ, by γ_0, then we can write:

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \overset{a}{=} g'(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0). \tag{7}$$

So this equation shows that $\sqrt{n}(\hat{\gamma} - \gamma_0)$ is asymptotically normal with mean 0, since the RHS is just $g'(\theta_0)$ times a quantity that is asymptotically normal with mean 0 by (5). The variance of $\sqrt{n}(\hat{\gamma} - \gamma_0)$ is clearly $g'(\theta_0)^2 V^\infty(\hat{\theta})$, and so we can say:

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \overset{a}{\sim} N(0, g'(\theta_0)^2 V^\infty(\hat{\theta})). \tag{8}$$

This says that $\hat{\gamma}$ is root-n consistent and asymptotically normal when $\hat{\theta}$ is.

4 The vector case

Suppose that $\boldsymbol{\theta}$ is a k-vector, and $\boldsymbol{\gamma}$ is an l-vector, where $l \leq k$. The relation between $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is $\boldsymbol{\gamma} = \mathbf{g}(\boldsymbol{\theta})$, where $\mathbf{g}(\boldsymbol{\theta})$ is an l-vector of monotonic functions that are continuously differentiable. The vector version of (5) is:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} N(\mathbf{0}, \mathbf{V}^\infty(\hat{\boldsymbol{\theta}})). \tag{9}$$

A first-order Taylor expansion of $\mathbf{g}(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ yields:

$$\hat{\boldsymbol{\gamma}} = \mathbf{g}(\hat{\boldsymbol{\theta}}) \approx \underbrace{\mathbf{g}(\boldsymbol{\theta}_0)}_{\boldsymbol{\gamma}_0} + \nabla\mathbf{g}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

and with a bit rearranging, and multiplying by $\sqrt{n}$, we get the following asymptotic equality:

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \overset{a}{=} \nabla\mathbf{g}(\boldsymbol{\theta}_0)\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{10}$$

where $\nabla\mathbf{g}(\boldsymbol{\theta}_0)$ is an $l \times k$ matrix with typical element $\partial g_i(\theta)/\partial\theta_j$, evaluated at $\boldsymbol{\theta}_0$, which we denote simply as $\mathbf{G}(\boldsymbol{\theta}_0)$. Much like (8), we can say that $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ is asymptotically normal with mean 0 due to (9). The variance of $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ is clearly $\mathbf{G}(\boldsymbol{\theta}_0)\mathbf{V}^\infty(\hat{\boldsymbol{\theta}})\mathbf{G}(\boldsymbol{\theta}_0)^\top$, and so we can write:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} N(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}_0)\mathbf{V}^\infty(\hat{\boldsymbol{\theta}})\mathbf{G}(\boldsymbol{\theta}_0)^\top). \tag{11}$$

The asymptotic covariance matrix that appears in the above equation is an $l \times l$ matrix, and it has full rank l if $\mathbf{V}^\infty(\hat{\boldsymbol{\theta}})$ is nonsingular and the matrix of derivatives $\mathbf{G}(\boldsymbol{\theta}_0)$ has full rank l.

In practice, the covariance matrix of $\hat{\boldsymbol{\gamma}}$ may be estimated by the matrix:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\gamma}}) = \mathbf{G}(\hat{\boldsymbol{\theta}})\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\mathbf{G}(\hat{\boldsymbol{\theta}})^{\top}. \tag{12}$$

This result is useful, but should be used with caution (like most things based on asymptotic theory). As in the scalar case, $\hat{\boldsymbol{\gamma}}$ cannot possibly be normally distributed if $\hat{\boldsymbol{\theta}}$ is.