Difference in Differences

New policies come into effect in various jurisdictions at various times.

Can we disentangle effects of a policy change from other changes across time or jurisdictions?

One commonly used method is a type of linear regression model called **difference in differences**, or **diff-in-diff**, or **DiD**.

Index jurisdictions by g and time periods by t. Then y_{gti} denotes the dependent variable for the i^{th} unit within jurisdiction g at time t.

Assume that, in the absence of the policy,

$$y_{gti} = \eta_g + \lambda_t + u_{gti}, \tag{1}$$

where η_g is a jurisdiction fixed effect and λ_t is a time fixed effect.

Since (1) includes two sets of fixed effects, it is often called the **two-way fixed effects**, or **TWFE**, model.

(1) imposes a common η_g on all time periods and a common λ_t on all jurisdictions. The latter is called the **parallel trends assumption**.

We cannot replace $\eta_g + \lambda_t$ by γ_{gt}, which would be much less restrictive, because then we could not identify the effects of the policy.

If a policy shifts $\mathrm{E}(y_{gti})$ by a constant δ, then (1) would include an additional term δ for any observation where the policy is active.

Let there be only two jurisdictions, denoted a and b, and two time periods, denoted 1 and 2.

If the policy is imposed in jurisdiction b in period 2 only, then we have four equations, one for each jurisdiction in each time period:

$$\begin{aligned} y_{a1i} &= \eta_a + \lambda_1 + u_{a1i}, \quad & y_{a2i} &= \eta_a + \lambda_2 + u_{a2i}, \\ y_{b1i} &= \eta_b + \lambda_1 + u_{b1i}, \quad & y_{b2i} &= \eta_b + \lambda_2 + \delta + u_{b2i}. \end{aligned} \tag{2}$$

Let $\bar{y}_{gt}$ and $\bar{u}_{gt}$ denote the average values of the y_{gti} and the u_{gti}, for $g = a, b$ and $t = 1, 2$. In this case, we can estimate δ using just the $\bar{y}_{gt}$.

(2) and our assumption about the policy effect imply that

$$\bar{y}_{a2} - \bar{y}_{a1} = \lambda_2 - \lambda_1 + (\bar{u}_{a2} - \bar{u}_{a1}), \tag{3}$$

and

$$\bar{y}_{b2} - \bar{y}_{b1} = \delta + \lambda_2 - \lambda_1 + (\bar{u}_{b2} - \bar{u}_{b1}). \tag{4}$$

Therefore,

$$(\bar{y}_{b2} - \bar{y}_{b1}) - (\bar{y}_{a2} - \bar{y}_{a1}) = \delta + (\bar{u}_{b2} - \bar{u}_{b1}) - (\bar{u}_{a2} - \bar{u}_{a1}). \tag{5}$$

The l.h.s. is the difference between two first differences, $\bar{y}_{b2} - \bar{y}_{b1}$ and $\bar{y}_{a2} - \bar{y}_{a1}$. The r.h.s. is δ plus a linear combination of the disturbances, which has mean zero.

The parameters λ_1 and λ_2 have vanished. The difference in differences on the left of equation (5) can be calculated. It gives us an estimate of δ.

When there are more than two time periods and/or jurisdictions, it is easier just to estimate a version of (1) with other regressors.

Define D^b_{gti} as a dummy variable that equals 1 if $g = b$ and 0 otherwise, and D^2_{gti} as a dummy variable that equals 1 if $t = 2$ and 0 otherwise.

Then equations (2) can be combined into just one equation:

$$y_{gti} = \beta_1 + \beta_2 D^b_{gti} + \beta_3 D^2_{gti} + \delta D^b_{gti} D^2_{gti} + u_{gti}. \tag{6}$$

The coefficient of interest is δ, which measures the effect of the treatment on jurisdiction b in period 2.

The first three coefficients in (6) are related to the ones in (2) as follows:

$$\beta_1 = \eta_a + \lambda_1, \quad \beta_2 = \eta_b - \eta_a, \quad \beta_3 = \lambda_2 - \lambda_1. \tag{7}$$

There are a few studies that use the DiD methodology with just two jurisdictions and two time periods, e.g. Card and Krueger (1994).

But it is impossible to allow for disturbances clustered by jurisdiction (and/or time period) without a reasonable number of jurisdictions (and/or a reasonable number of time periods).

In general, there are $G \geq 2$ clusters, of which G_1 are treated in at least some of T time periods and G_0 are never treated.

$$y_{gti} = \beta_1 + \sum_{j=2}^{G} \beta_j \mathrm{DJ}_{gti}^{j} + \sum_{k=1}^{T-1} \beta_{G+k} \mathrm{DT}_{gti}^{k} + \delta \mathrm{TR}_{gti} + u_{gti}. \tag{8}$$

The DJ_{gti}^{j} are jurisdiction dummies equal to 1 when $g = j$, and the DT_{gti}^{k} are time dummies equal to 1 when $t = k$. TR_{gti} is a treatment dummy.

$\mathrm{TR}_{gti} = 1$ for treated observations in the G_1 treated clusters. It equals 0 for the remaining observations in those clusters and for all observations in the G_0 untreated clusters.

- If standard errors are clustered by jurisdiction, there must be at least a moderate number of treated jurisdictions.
- If standard errors are clustered by time period, there must be at least a moderate number of treated time periods.

Equation (8) may also include explanatory (control) variables, but not ones that vary only by time or only by jurisdiction.

We can only identify the parameter δ if some jurisdictions are treated in some periods and not treated in others.

- When computing test statistics or confidence intervals based on equations like (8), it is obligatory to use a CRVE.
- In principle, one could cluster by time period, by jurisdiction, by both of them, or by time-jurisdiction pairs.
- Clustering by jurisdiction-period pairs is a bad idea. Clustering by jurisdiction or two-way clustering works better.

Recent highly influential work on DiD relaxes the assumption about common treatment effects and allows effects of treatment to vary over time. In such cases, the usual TWFE model can be misleading.

- de Chaisemartin and D'Haultefoeuille (AER, 2020)
- Callaway and Sant'Anna (Journal of Econometrics, 2021)
- Sun and Abraham (Journal of Econometrics, 2021)
- de Chaisemartin and D'Haultefoeuille (Ects. Journal, 2023)

The Delta Method

One popular way to estimate the standard error of a nonlinear function of parameter estimates is to use the **delta method**.

We estimate a scalar parameter θ, and we are interested in $\gamma \equiv g(\theta)$, where $g(\cdot)$ is a monotonic function that is continuously differentiable.

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 $\hat{\theta}$, $\text{Var}(\hat{\gamma})$ should be a function of $\text{Var}(\hat{\theta})$. If $g(\theta) = w\theta$, we already know how to calculate $\text{Var}(\hat{\gamma})$:

$$\text{Var}(\hat{\gamma}) = w^2 \, \text{Var}(\hat{\theta}). \tag{9}$$

This is a special case of $\text{Var}(\hat{\gamma}) = \boldsymbol{w}^\top \text{Var}(\hat{\boldsymbol{\theta}}) \boldsymbol{w}$.

The delta method simply finds a linear approximation to $g(\theta)$ and then applies (9) to this approximation.

Taylor's Theorem

Taylor's Theorem applies to functions that are continuously differentiable at least once on some real interval $[a, b]$.

The figure shows the graph of such a function, $f(x)$, for $x \in [a, b]$. The coordinates of A are $(a, f(a))$, and those of B are $(b, f(b))$. Thus the slope of the line AB is $\big(f(b) - f(a)\big) / (b - a)$.

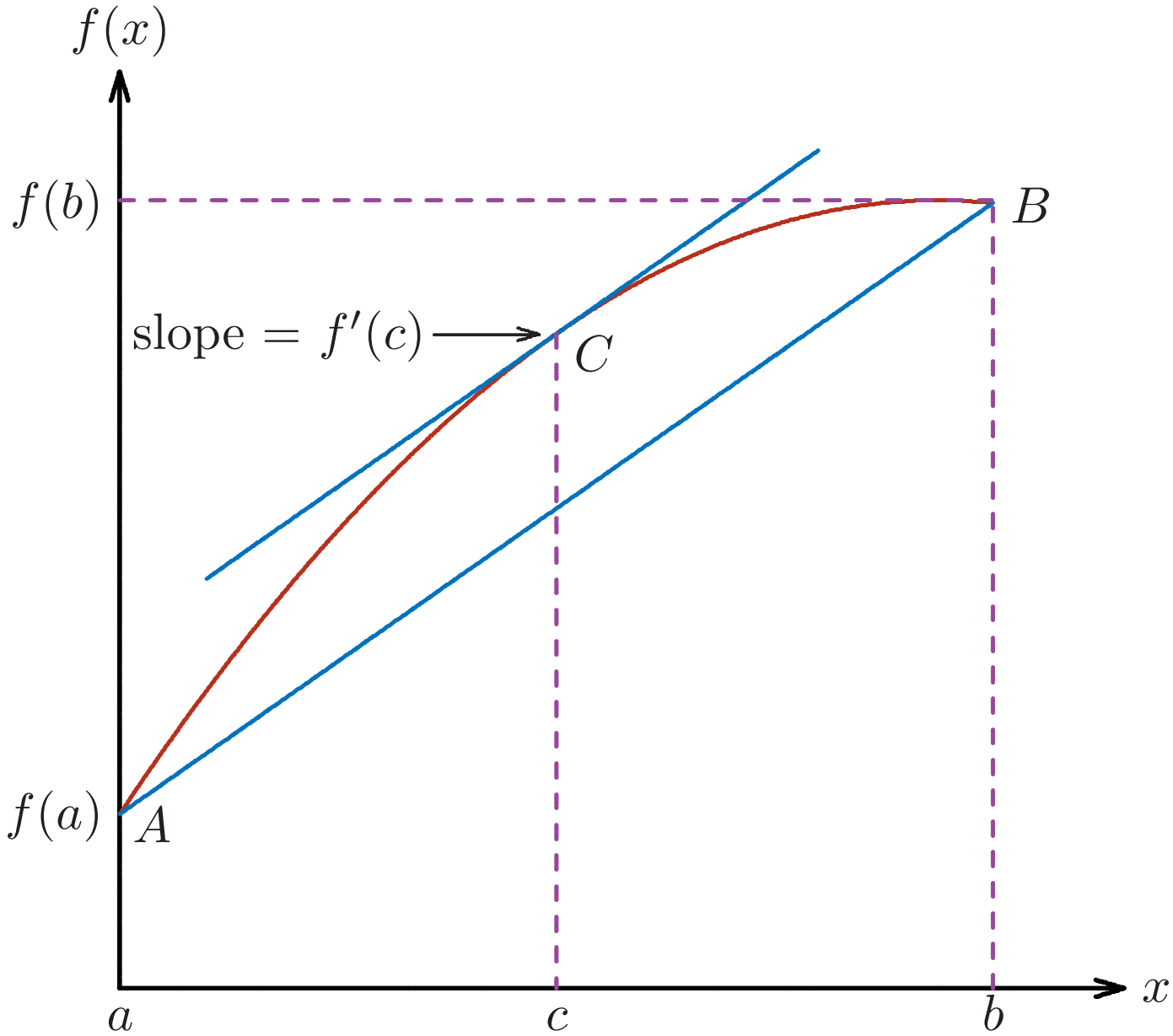
There must always be a value between a and b, like c in the figure, at which the derivative $f'(c)$ is equal to the slope of AB.

This is a consequence of the continuity of the derivative. If $f'(x)$ is continuous on $[a, b]$, then there must exist c such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{10}$$

This can be rewritten as

$$f(b) = f(a) + (b - a) f'(c). \tag{11}$$


$f(x)$
$f(b)$
B
slope $= f'(c)$
C
$f(a)$
A
x
a
c
b

If we let $h = b - a$, then, since c lies between a and b, it must be the case that $c = a + \lambda h$ for some λ between 0 and 1. Thus we obtain

$$f(a+h) = f(a) + hf'(a+\lambda h). \tag{12}$$

This result is also known as the **Mean Value Theorem**.

More commonly, we set $\lambda = 0$ to obtain a linear approximation to $f(x)$ for $x \approx a$. The **first-order Taylor expansion** around a is

$$f(a+h) \cong f(a) + hf'(a), \tag{13}$$

where the symbol "$\cong$" means "is approximately equal to."

The **second-order Taylor expansion** is

$$f(a+h) \cong f(a) + hf'(a) + \frac{1}{2}h^2 f''(a), \tag{14}$$

where $f(x)$ must have a p^{th} derivative that is continuous on $[a, a+h]$.

Taylor's Theorem can be extended to polynomials of any desired order. If that order is p, the analog of (12) is

$$f(a+h) = f(a) + \sum_{i=1}^{p-1} \frac{h^i}{i!} f^{(i)}(a) + \frac{h^p}{p!} f^{(p)}(a+\lambda h). \tag{15}$$

Here $f^{(i)}$ is the i^{th} derivative of f, and once more $0 < \lambda < 1$.

If $f(\boldsymbol{x})$ is a scalar-valued function of the m-vector $\boldsymbol{x}$, then, when $\boldsymbol{h}$ is also an m-vector, the first-order Taylor expansion is

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \sum_{j=1}^{m} h_j f_j(\boldsymbol{x}+\lambda \boldsymbol{h}), \tag{16}$$

where h_j is the j^{th} component of $\boldsymbol{h}$, f_j is the partial derivative of f with respect to its j^{th} argument, and, as before, $0 < \lambda < 1$.

The Delta Method for a Scalar Parameter

If $\hat{\theta}$ is root-N consistent and asymptotically normal, then

$$N^{1/2}(\hat{\theta} - \theta_0) \overset{a}{\sim} \mathrm{N}\big(0, V^{\infty}(\hat{\theta})\big), \tag{17}$$

where $V^{\infty}(\hat{\theta})$ is the asymptotic variance of $N^{1/2}(\hat{\theta} - \theta_0)$.

To find the asymptotic distribution of $\hat{\gamma} = g(\hat{\theta})$, we perform a first-order Taylor expansion of $g(\hat{\theta})$ around θ_0. Using (12), we obtain

$$\hat{\gamma} \cong g(\theta_0) + g'(\theta_0)(\hat{\theta} - \theta_0), \tag{18}$$

where $g'(\theta_0)$ is the first derivative of $g(\theta)$, evaluated at θ_0.

Multiplying both sides of (18) by $N^{1/2}$ and letting $\gamma_0 \equiv g(\theta_0)$ and $g_0' \equiv g'(\theta_0)$, we obtain

$$\mathrm{N}^{1/2}(\hat{\gamma} - \gamma_0) \overset{a}{=} g_0' N^{1/2}(\hat{\theta} - \theta_0). \tag{19}$$

Since the r.h.s. of (19) is g_0' times something asymptotically normal with mean 0, so must be $N^{1/2}(\hat{\gamma} - \gamma_0)$.

The variance of $N^{1/2}(\hat{\gamma} - \gamma_0)$ is clearly $(g_0')^2 V^\infty(\hat{\theta})$, and so we conclude that

$$N^{1/2}(\hat{\gamma} - \gamma_0) \stackrel{a}{\sim} \mathrm{N}\big(0, (g_0')^2 V^\infty(\hat{\theta})\big). \tag{20}$$

This shows that $\hat{\gamma}$ is root-N consistent and asymptotically normal whenever $\hat{\theta}$ is both of those things.

From (13), if the standard error of $\hat{\theta}$ is s_θ, then the s.e. of $\hat{\gamma}$ is

$$s_\gamma \equiv \left|g'(\hat{\theta})\right| s_\theta. \tag{21}$$

Any asymptotically valid standard error for $\hat{\theta}$ can be used. It may be robust to heteroskedasticity, and/or to autocorrelation, and/or to within-cluster correlation.

Consider the case in which $\gamma = \theta^2$. Then $g'(\theta) = 2\theta$, and (21) tells us that $s_\gamma = 2|\hat{\theta}|s_\theta$.

Confidence Intervals and the Delta Method

When the finite-sample distributions of estimates are far from the limiting normal distribution, we cannot expect any asymptotic procedure to perform well.

Whenever $g(\theta)$ is nonlinear, it is impossible that $\hat{\theta}$ and $\hat{\gamma}$ should *both* be normally distributed in finite samples, as the delta method pretends.

- Suppose that $\hat{\theta}$ really does happen to be normally distributed. Then, unless $g(\cdot)$ is linear, $\hat{\gamma}$ cannot possibly be normally, or even symmetrically, distributed.
- Similarly, if $\hat{\gamma}$ is normally distributed, $\hat{\theta}$ cannot be either normally or symmetrically distributed.

Moreover, since s_γ generally depends on $\hat{\theta}$, the numerator of a t statistic for γ is not independent of the denominator. But independence is essential for a t statistic to follow the Student's t distribution.

The obvious asymptotic confidence interval for γ is

$$\left[\hat{\gamma} - s_\gamma z_{1-\alpha/2},\ \ \hat{\gamma} + s_\gamma z_{1-\alpha/2}\right], \tag{22}$$

where s_γ is the delta method estimate, and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

This confidence interval can be expected to work well when:

1. the finite-sample distribution of $\hat{\gamma}$ is well approximated by the normal distribution;
2. s_γ is a reliable estimator of the standard deviation of $\hat{\gamma}$;
3. s_γ is approximately independent of $\hat{\gamma}$.

But it may work badly otherwise. It may **over-cover** or **under-cover** (the latter seems to be more common). It may even over-cover at one end and under-cover at the other.

With complicated nonlinear models, is it reasonable to assume that points 1, 2, and 3 are true for the parameters we care about?

Instead of using an interval based on the delta-method standard error s_γ, we can transform the interval for the underlying parameter θ,

$$\left[\hat{\theta} - s_\theta z_{1-\alpha/2},\ \ \hat{\theta} + s_\theta z_{1-\alpha/2}\right]. \tag{23}$$

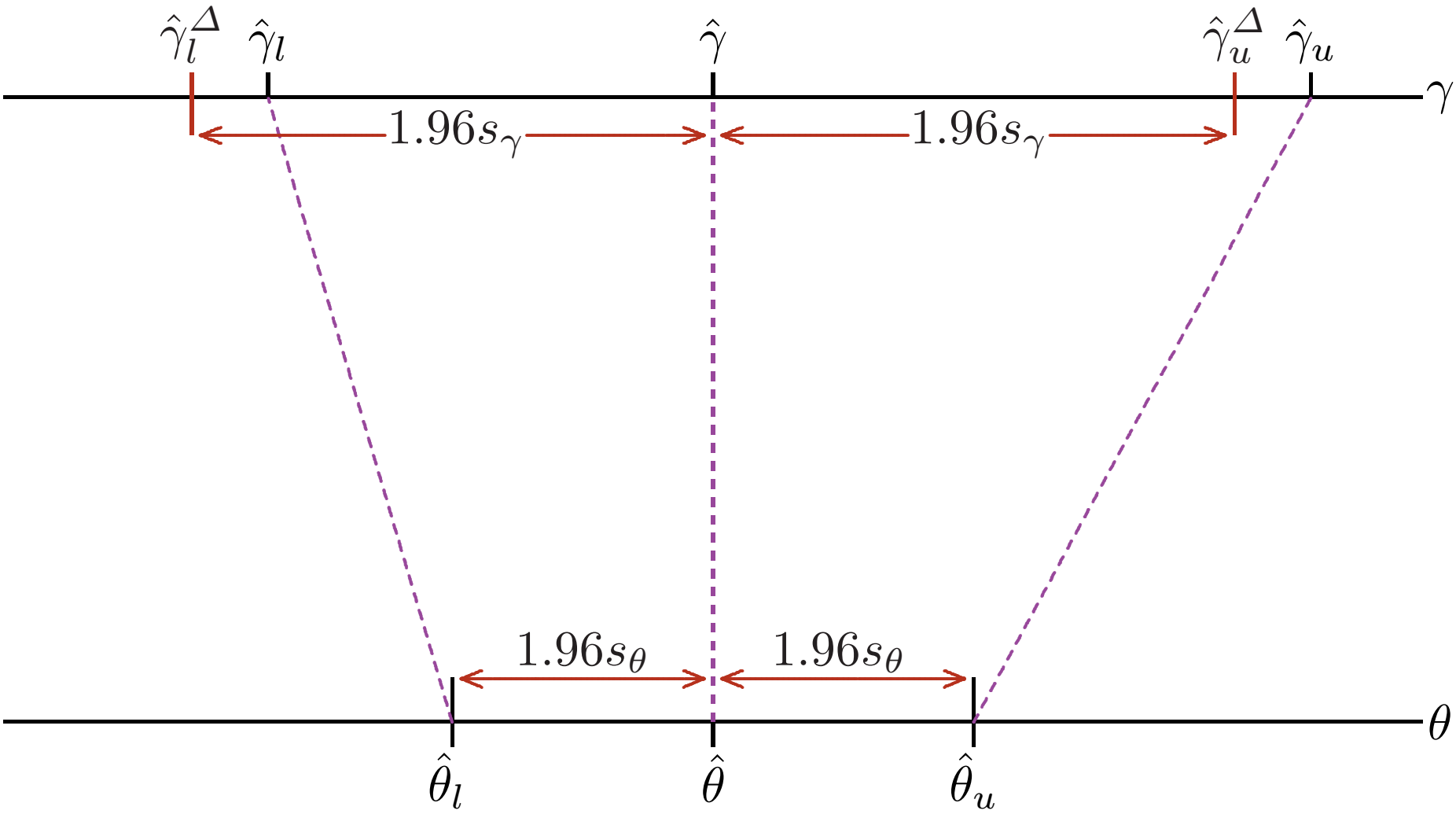
Transforming the endpoints of this interval by the function $g(\cdot)$ gives the following interval for γ:

$$\left[g(\hat{\theta} - s_\theta z_{1-\alpha/2}),\ \ g(\hat{\theta} + s_\theta z_{1-\alpha/2})\right]. \tag{24}$$

This formula assumes that $g'(\theta) > 0$. If $g'(\theta) < 0$, the two ends of the interval would have to be interchanged.

Whenever $g(\theta)$ is a nonlinear function, the confidence interval (24) must be asymmetric.

See the figure, which compares the delta-method interval (22) with the transformed interval (24) for the case $\gamma = \theta^2$.


$\hat{\gamma}_l^{\Delta}$
$\hat{\gamma}_l$
$\hat{\gamma}$
$\hat{\gamma}_u^{\Delta}$
$\hat{\gamma}_u$
γ
$1.96s_\gamma$
$1.96s_\gamma$
$1.96s_\theta$
$1.96s_\theta$
θ
$\hat{\theta}_l$
$\hat{\theta}$
$\hat{\theta}_u$

The Delta Method for Several Parameters

The result (21) can be extended to the case in which $\boldsymbol{\theta}$ is a k-vector and $\boldsymbol{\gamma}$ is an l-vector, with $l \leq k$.

The relation between $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is $\boldsymbol{\gamma} \equiv \boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{g}(\boldsymbol{\theta})$ is an l-vector of monotonic functions that are continuously differentiable.

We start from the asymptotic result that

$$N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} \mathrm{N}\big(\mathbf{0}, \boldsymbol{V}^{\infty}(\hat{\boldsymbol{\theta}})\big), \tag{25}$$

where $\boldsymbol{V}^{\infty}(\hat{\boldsymbol{\theta}})$ is the asymptotic covariance matrix of $N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

Using (25) and a first-order Taylor expansion of $\boldsymbol{g}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$,

$$N^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \overset{a}{\sim} \mathrm{N}\big(\mathbf{0}, \boldsymbol{G}_0 \boldsymbol{V}^{\infty}(\hat{\boldsymbol{\theta}}) \boldsymbol{G}_0^{\top}\big), \tag{26}$$

where $\boldsymbol{G}_0$ is an $l \times k$ matrix with typical element $\partial g_i(\theta)/\partial \theta_j$, evaluated at $\boldsymbol{\theta}_0$. It is the **Jacobian** of the transformation.

The asymptotic covariance matrix $\boldsymbol{G}_0 \boldsymbol{V}^{\infty}(\hat{\boldsymbol{\theta}}) \boldsymbol{G}_0^{\top}$ is an $l \times l$ matrix.

It has full rank l if $\boldsymbol{V}^{\infty}(\hat{\boldsymbol{\theta}})$ is nonsingular and $\boldsymbol{G}_0$ has rank l.

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

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\gamma}}) \equiv \hat{\boldsymbol{G}} \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{G}}^{\top}, \tag{27}$$

where $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{G}} \equiv \boldsymbol{G}(\hat{\boldsymbol{\theta}})$.

- Like the scalar delta-method standard error estimator (21), the covariance matrix estimator (27) may or may not prove to have good properties in finite samples.
- Wald tests based on (27) often over-reject severely in finite samples, especially when testing more than a few restrictions.

We can often obtain much more reliable results by using **bootstrap methods**, including **bootstrap standard errors**, **bootstrap covariance matrices**, **bootstrap tests**, and **bootstrap confidence intervals**.