How to Obtain Confidence Intervals without Simulation: The Delta Method

Justin Grimmer and Holger Kern

March 22, 2007

Suppose that you are stranded on a desert island, without a computer and you need to compute the function $g(\cdot)$ of some maximum-likelihood estimated parameter θ. Remembering the lessons of gov 2001, you know that you need to provide some measure of uncertainty about the estimate of $g(\theta)$ (even being marooned on a desert island is no excuse to poorly present results). What are you to do?

One way to proceed is to use the *delta-method*. The idea behind the delta method is that we can use a linear approximation of $g(\theta)$ to compute a pretty good approximation of the variance of our estimate. This brief handout describes how to compute that approximation, why it works, and then provides some examples.

1 How the Delta Method Works (This is pretty technical and not necessary)

The key concept for understanding the delta method is convergence in distribution. A sequence of random variables x_n converges to a random variable x if there exists N s.t. for all $n^{'} > N$, x and $x_{n'}$ have the same cdf. For example, we know that the t-statistic for testing the hypothesis that a coefficient is not equal to zero converges in distribution to a standardized normal distribution (Greene pg 907).

Let's suppose that we obtain a parameter estimate of θ, $\hat{\theta}$ using maximum likelihood estimation. Furthermore, let's suppose that,

$$\hat{\theta} \quad \rightarrow^{d} \quad \mathcal{N}(\theta, \sigma^2). \tag{1}$$

where $\rightarrow^{d}$ shows that the convergence is in distribution. This implies that, for sufficiently large n, $\hat{\theta}$ will be centered at θ with variance σ^2.

Suppose that we are interested in some function of the estimated parameter $g : \Theta \rightarrow \Re$, where Θ is the parameter space. Further, let's suppose that $g(\cdot)$ is continuous and differentiable for all $\theta \in \Theta$. Then, we have (Greene pg 913)[1],

$$g(\hat{\theta}) \quad \rightarrow^{d} \mathcal{N}(g(\theta), \sigma^2(g^{'}(\theta))^2)$$

[1]Note, these assumptions are stronger than necessary see Greene pg 913

What is the intuition behind this result? First, since $g(\cdot)$ is continuous we know that as $\hat{\theta}$ approaches θ, then $g(\hat{\theta})$ should approach $g(\theta)$. To compute the variance, we take the linear Taylor series approximation of $g(\hat{\theta})$,[2]

$$g(\hat{\theta}) \approx g(\theta) + g^{'}(\theta)(\hat{\theta} - \theta)$$

Now, we compute the variance using this approximation

$$\text{var}(g(\hat{\theta})) \approx \text{var}\left(g(\theta)\right) + \text{var}\left(g^{'}(\theta)(\hat{\theta} - \theta)\right) \tag{2}$$

Remember, that we are assuming that there exists a true parameter θ, which therefore has no variance, so Equation 2 equals,

$$\begin{aligned} &= \text{var}\left(g^{'}(\theta)\hat{\theta}\right) \\ &= g^{'}(\theta)^2 \text{var}\left(\hat{\theta}\right) \\ &= g^{'}(\theta)^2\sigma^2 \end{aligned} \tag{3}$$

Where the last line follows by Equation 1.
Now, we are equipped with the center and an estimate of the variance so we can use the maximum likelihood estimation.

2 Univariate Example

Suppose that we observe $Y_i \sim \mathcal{N}(\mu, 1)$ for $i = 1, \ldots, n$. We will suppose that all our observations are from the same normal distribution and each observation is independent. Therefore, our log-likelihood is,

$$ll(\mu|y) \doteq \sum_{i=1}^{n} -\frac{(y_i - \mu)^2}{2}. \tag{4}$$

Differentiating with respect to μ, we have

$$\begin{aligned} \frac{\partial ll(\mu|y)}{\partial \mu} &= \sum_{i=1}^{n}(y_i - \mu) \\ &= n\bar{y} - n\mu. \end{aligned}$$

Now, setting equal to zero and solving we have,

$$\bar{y} = \mu. \tag{5}$$

[2]This is linear because we have discarded the $g^{''}$ portion of the series. If we included this component, this would be the best quadratic approximation

Checking the second order condition and getting the hessian at the same time, we have,

$$\frac{\partial^2 ll(\mu|y)}{\partial \mu^2} = -n. \tag{6}$$

Therefore, our maximum likelihood estimate of μ is $\bar{y}$ and our variance of this estimate is $1/n$.

Now, let's suppose that we want to know the distribution of the square of μ. Then our function $f : \Re \to \Re_+$, $f(\mu) = \mu^2$, and $f^{'}|_{\mu=\mu^*} = 2\mu^*$. Then, using the results from above we have that our statistic is distributed

$$f(\hat{\mu}) = \hat{\mu}^2 \sim \mathcal{N}(\mu^2, \frac{4\mu^2}{n}) \tag{7}$$

See the `R` code from section to see that this is equivalent to the simulation based method.

3 Multivariate Case (Once again, pretty technical)

What do we do if we have a vector of random variables, like the vector $\hat{\boldsymbol{\beta}}$ from a regression? Luckily, we have a very similar result. Suppose that there is a function $F : \boldsymbol{B} \to \Re$, where $\boldsymbol{\beta} \in \boldsymbol{B}$, and $\boldsymbol{B} \subset \Re^k$. Further, let's suppose that F is continuous and differentiable with respect to each β_k in the vector $\boldsymbol{\beta}$. Then, the theorem that drives the delta method tells us that, for sufficiently large n that,

$$F(\hat{\boldsymbol{\beta}}) \to^d \mathcal{N}\left(F(\boldsymbol{\beta}), \frac{\partial F(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} \text{var}(\hat{\boldsymbol{\beta}}) \frac{\partial F(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right) \tag{8}$$

The intuition for Equation 8 is similar to the univariate case. The mean arises because $F(\cdot)$ is continuous, and the variance can be thought of as resulting from a linear approximation. In this case, the linear approximation of $F(\hat{\boldsymbol{\beta}})$ is ,

$$F(\hat{\boldsymbol{\beta}}) \approx F(\boldsymbol{\beta}) + \frac{\partial F(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{9}$$

and the variance of this expression is

$$\text{var}(F(\hat{\boldsymbol{\beta}})) \approx (\frac{\partial F(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}})^T \text{var}(\hat{\boldsymbol{\beta}}) \frac{\partial F(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} \tag{10}$$

Which follows from a similar argument to the univariate case (replacing the squared portion with the analogous operation for matrix algebra).

4 Multivariate Example

Suppose that we are given a set of observations $Y_i \sim \text{Bernoulli}(y_i|\pi_i)$, where $\pi_i = F(X_i\hat{\boldsymbol{\beta}})$. Therefore, in this case $F : \boldsymbol{B} \rightarrow [0, 1]$, where $F(X_i\hat{\boldsymbol{\beta}}) = \Pr(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{\beta})$ and we will suppose that F is some cumulative distribution function with corresponding pdf f. How do we use the above results?

Well, once we obtain our MLE results for $\hat{\boldsymbol{\beta}}$, we can realize that,

$$\begin{aligned} \frac{\partial F(\boldsymbol{X_i}\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} &= \frac{\partial F(\boldsymbol{X_i}\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{X_i}\hat{\boldsymbol{\beta}}} \frac{\partial \boldsymbol{X_i}\hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}} \\ &= f(\boldsymbol{X_i}\hat{\boldsymbol{\beta}})\boldsymbol{X_i} \end{aligned}$$

Therefore, we have

$$\text{var}\left(F(\boldsymbol{X_i}\hat{\boldsymbol{\beta}})\right) \approx f(\boldsymbol{X_i}\hat{\boldsymbol{\beta}})^2 \boldsymbol{X_i}^T \text{var}(\hat{\boldsymbol{\beta}}) \boldsymbol{X_i} \tag{11}$$

where T denotes the transpose.

In the section code, we show how to use this result to compute the standard error for predicted values and first differences when we assume that F is the cumulative t-distribution, or when we are doing "Robit" regression (Gelman and Hill 2006).

5 Further Reading

This is a pretty good source on the delta method and probit/logit regressions

`http://www.indiana.edu/~jslsoc/stata/spostci/spost_deltaci.pdf`

This is the must-have guide to Econometrics
William Greene (2003). *Econometric Analysis*. Pearson.