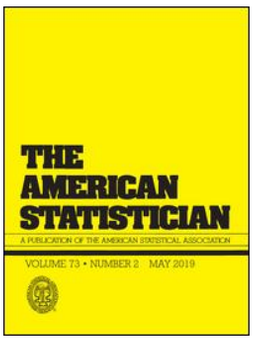

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

On the Mean Value Theorem for Estimating Functions

Alexandre Galvão Patriota

To cite this article: Alexandre Galvão Patriota (2019): On the Mean Value Theorem for Estimating Functions, The American Statistician, DOI: 10.1080/00031305.2018.1558110

To link to this article: https://doi.org/10.1080/00031305.2018.1558110

Accepted author version posted online: 15 Jan 2019.
Published online: 13 May 2019.

THE AMERICAN STATISTICIAN
2019, VOL. 00, NO. 0, 1–4: Short Technical Note
https://doi.org/10.1080/00031305.2019.1558110

On the Mean Value Theorem for Estimating Functions

Alexandre Galvão Patriota

Departamento de Estatística, IME, Universidade de São Paulo, São Paulo/SP, Brazil

ABSTRACT

Feng et al. revealed that the usual mean value theorem (MVT) should not be applied directly to a vector-valued function (e.g., the score function or a general estimating function under a multiparametric model). This note shows that the application of the Cramer–Wold's device to a corrected version of the MVT is sufficient to obtain standard asymptotics for the estimators attained from vector-valued estimating functions.

ARTICLE HISTORY
Received April 2018
Accepted November 2018

KEYWORDS
Asymptotics; Mean value theorem; Score function; Taylor's expansion.

1. Introduction

One of the goals in statistics is to estimate an unknown quantity $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^p$ is the parameter space, by the roots of an estimating function $U_n : \Theta \to \mathbb{R}^p$. For instance, if the log-likelihood function of the statistical model is concave and differentiable at each $\theta \in \Theta$, then the root of its derivative with respect to θ is the maximum-likelihood estimator. It is well known that, under some regularity conditions on U_n, the estimator $\hat{\theta}_n$ such that $U_n(\hat{\theta}_n) = 0$ is consistent and asymptotically normally distributed (van der Vaart 1998).

The mean value theorem (MVT) and Taylor's expansions are common techniques employed to study the limiting distribution of $z_n = \sqrt{n}(\hat{\theta}_n - \theta)$. Typically, this topic is studied in a graduate course of mathematical statistics. A common didactic artifice to justify the asymptotic distribution of z_n is by means of a straightforward "extension" to the vector-valued estimating function $U_n(\theta)$ of the MVT derived for real-valued functions, see Feng et al. (2013) and also Equation (1) of Section 2. Many authors have employed this "extension" in the statistical literature, namely, Barnett (1976), Wu (1981), Andersen and Gill (1982), McCullagh (1983), Serfling (1993, Lemma B on p. 153), Wang (1999), and Jacod and Sørensen (2018); for specific details of some of these cases, go to Section 3. Although it is simple and easy to understand, Feng et al. (2013) provided a vector-valued function where such an "extension" is false and offered two ways to circumvent the problem one of which is a consecrated MVT for vector-valued functions already used in some books of asymptotic theory (see Barndorff-Nielsen and Cox 1996; Ferguson 1996; van der Vaart 1998; Shao 2003, to cite a few). This consecrated version makes use of an integral operation on the derivative of U_n with respect to θ, see the rule (3) on page 20 in Ferguson (1996), and, hence, this technique is not readily applicable without imposing further conditions on the estimating function.

In this short communication, we show how to correct the invalid, but often employed, MVT for vector-valued estimating functions without changing the main mathematical tools. We also give a simple argument to justify the asymptotic theory, see Section 2. To illustrate this, some examples are presented in Section 3. In Section 4, a brief final remark is offered.

2. Mean Value Theorem for Estimating Functions

We assume the following regularity conditions throughout this article:

(A)Θ is an open convex set,
(B)U_n is continuously differentiable at each $\theta \in \Theta$.

As pointed out by Feng et al. (2013), one very common identity used by many books and papers is

$$U_n(\hat{\theta}_n) = U_n(\theta) + U_n'(\tilde{\theta}_n)(\hat{\theta}_n - \theta), \tag{1}$$

where $\tilde{\theta}_n = h\hat{\theta}_n + (1-h)\theta$ for some $h \in (0,1)$ and $U_n'(\theta) = \partial U_n(\theta)/\partial\theta^\top$ is the first derivative of U_n with respect to θ. From Equation (1) and some extra regularity conditions, it is possible to show that

$$\frac{1}{\sqrt{n}} I_\theta^{-1} U_n(\theta) = z_n + o_p(1), \tag{2}$$

where $z_n = \sqrt{n}(\hat{\theta}_n - \theta)$. However, Feng et al. (2013) showed a function, whose domain and codomain are subsets of $\mathbb{R}^2$, that violates the identity Equation (1). That is, the violation could happen whenever the domain and codomain of U_n are subsets of $\mathbb{R}^p$ with $p \geq 2$. Therefore, a correction on this expansion is important to correctly justify standard asymptotics for $\hat{\theta}_n$ without changing the main ingredients of the standard technique.

Due to the well-known Cramér–Wold's theorem (see the Appendix), Equation (1) is not necessary to study the first-order asymptotics in statistics. The following true identity suffices:

$$\lambda^\top U_n(\hat{\theta}_n) = \lambda^\top U_n(\theta) + \lambda^\top U_n'(\tilde{\theta}_{n,\lambda})(\hat{\theta}_n - \theta),$$

CONTACT Alexandre Galvão Patriota patriota@ime.usp.br Departamento de Estatística, IME, Universidade de São Paulo, Rua do Matão, 1010, São Paulo/SP, 05508-090, Brazil.

for $\lambda \in \mathbb{R}^p$, where $\tilde{\theta}_{n,\lambda} = h_\lambda \hat{\theta}_n + (1-h_\lambda)\theta$ for some $h_\lambda \in (0,1)$. Since the codomain of the function $\lambda^\top U_n(\theta)$ is the real set, by the usual mean theorem value, the above equation is always true (see Rudin 1976, for instance); moreover, as $U_n(\hat{\theta}_n) = 0$, it reduces to

$$\lambda^\top U_n(\theta) = -\lambda^\top U_n'(\tilde{\theta}_{n,\lambda})(\hat{\theta}_n - \theta). \tag{3}$$

The Cramér–Wold's theorem, the identity Equation (3) and some extra conditions on $U_n'(\theta)$ guarantee the asymptotic relation between $U_n(\theta)$ and $\hat{\theta}_n$ presented in Equation (2).

Theorem 1. In addition to the assumptions (A) and (B), assume also that

(i) $\| \hat{\theta}_n - \theta \| = o_p(1)$,
(ii) there exists a positive definite matrix I_θ such that $\| \frac{1}{n} U_n'(\theta) + I_\theta \| = o_p(1)$,
(iii) $\sup_{k\in(0,1)} \left\| \frac{1}{n} U_n'\big(\theta + k(\hat{\theta}_n - \theta)\big) - \frac{1}{n} U_n'(\theta) \right\| = o_p(1)$, and
(iv) $\frac{1}{\sqrt{n}} U_n(\theta)$ converges in distribution to a random vector

for each $\theta \in \Theta$. Then, Equation (2) holds for each $\theta \in \Theta$.

The matrix I_θ is the Fisher information under the standard asymptotic theory for maximum likelihood estimators. Notice also that if $U_n'(\theta)$ is a type of Lipschitz continuous function,[1] then item iii holds from item *i*. The proof of the above theorem is presented in the Appendix; it makes use of Equation (3) and the Cramér–Wold's theorem, but it could also have been proved directly from the invalid Equation (1) without applying the Cramér–Wold's theorem. Feng et al. (2013) proposed two ways to circumvent the problem, but both of them require different techniques and more conditions to verify. Here, we just need to multiply the terms of the wrong Equation (1) by $\lambda^\top$ to obtain a valid expansion and apply the Cramér–Wold's theorem to justify the asymptotic distribution of z_n.

3. Examples in the Literature

In this section, we present three examples in the literature where the invalid MVT (1) is employed to justify the asymptotics of proposed estimators. We also suggest how to correct the arguments without changing the main assumptions of the involved theorems.

3.1. Nonlinear Regression Model

Barnett (1976) studied the asymptotic distribution of the maximum likelihood estimators under the following nonlinear regression model: $y_t = g(x_t, \gamma) + \epsilon_t$, where the random vectors ϵ_t, $t = 1, \ldots, T$, are independent and identically distributed (iid) as $N_q(0, \Omega)$ with Ω being a $(q \times q)$ positive definite matrix and g is a continuous and differentiable nonlinear function with respect to γ. Let $\theta = (\gamma^\top, \text{vech}(\Omega)^\top)^\top$ be the p-dimensional parameter vector, where $p = q + q(q+1)/2$ and the vech operator stacks the columns of a symmetric matrix without its redundant elements. The vector-valued estimating function is the derivative of the log-likelihood function with respect to θ, namely, $U_T(\theta) = \frac{\partial \log L(\theta|y,x)}{\partial \theta}$, where $L(\theta|y,x)$ is the likelihood function. Barnett's (1976) theorem 2 states that $z_T = \sqrt{T}(\hat{\theta}_T - \theta)$ converges to a normal distribution, where $\hat{\theta}_T$ is the maximum likelihood estimator. In the appendix, at the left bottom of page 358, Barnett (1976) gave a proof of theorem 2 in which the invalid MVT formula (1) is explicitly employed as we show in the box below:

> "*Proof of Theorem 2:* By the multivariate mean value theorem, it follows that
>
> $$(1/T)\frac{\partial \log L(\theta|y,x)}{\partial \theta_i}\bigg|_{\theta=\hat{\theta}} = (1/T)a_{Ti}(\theta_0) - (1/T)b_{Ti}'(\theta^*)(\hat{\theta} - \theta_0),$$
>
> where b_{Ti}' is the ith row of $B_T(\theta)$, and θ^* is on the line between $\hat{\theta}$ and θ_0. So by definition of $\hat{\theta}$, we can determine that
>
> $$0 = (1/T)a_T(\theta_0) - (1/T)B_T'(\theta^*)(\hat{\theta} - \theta_0),"$$

where T is the sample size, $a_T(\theta)$ is the vector-valued estimating function $U_T(\theta)$ and $B_T'(\theta)$ is the matrix $U_T'(\theta)$. Clearly, in the first equation of the box, θ^* should be indexed by i or, alternatively, in the second equation, both sides should be multiplied by $\lambda^\top$, where λ is a vector with the same length as $a_T(\theta)$.

Wu (1981) studied the asymptotic distribution of the least squares estimators under this nonlinear model with homoscedastic non normal errors (constant variances). McCullagh (1983) examined the connection between quasi-likelihood functions, exponential models and nonlinear least squares. Although the shape of the estimating function in these two papers is not the derivative of the log-likelihood function, the authors justify the asymptotics through the invalid MTV (1); see formula (4.6) in the former and formula (12) in the latter.

3.2. Cox's Regression Model

Andersen and Gill (1982) developed an asymptotic theory for the maximum likelihood estimators in a Cox's regression model for censored survival data. Let $T_1, \ldots, T_n$ be n possible right censored survival times and $z_1, \ldots, z_n$ the corresponding covariate vectors, where z_i is observed on $[0, T_i]$. The partial likelihood function is

$$L(\beta, T, z) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta^\top z_i(T_i)}}{\sum_{j\in\mathcal{R}_i} e^{\beta^\top z_j(T_i)}} \right\}^{\delta_i},$$

where $\mathcal{R}_i = \{j : T_j \geq T_i\}$ and $1 - \delta_i$ is the censoring indicator. The estimator $\hat{\beta}_n$ that maximizes L is a root of the derivative of the logarithm of the partial likelihood function with respect to β, namely $U_n(\beta, 1)$. In their Section 2.2, Formula (2.5), this vector-valued estimating function is expanded as follows:

[1] That is, it satisfies the following condition: $\left\| \frac{1}{n} U_n'(\theta_1) - \frac{1}{n} U_n'(\theta_2) \right\| \leq q_n \| \theta_1 - \theta_2 \|$, where $q_n = O_p(1)$.

> "Taylor expanding $U(\beta, 1)$ around β_0, we get
>
> $$U(\beta, 1) - U(\beta_0, 1) = -\mathcal{I}(\beta^*, 1)(\beta - \beta_0),$$
>
> where β^* is on the line segment between β and β_0 [and the positive semidefinite matrix $\mathcal{I}(\beta, 1)$ is minus the derivative of $U(\beta, 1)$.]"

As $U(\beta, 1)$ is a vector-valued estimating function, the above expansion is the application of the invalid MTV formula (1). To correct this expansion, one could multiply by $\lambda^\top$ in both sides of the above equation, where λ is a vector with the same length as $U(\beta, 1)$.

3.3. General Estimating Functions

Jacod and Sørensen (2018) provided a review of asymptotic theory of estimating functions. They considered a vector-valued estimating function G_n and an estimator $\hat{\theta}_n$ such that $G_n(\hat{\theta}_n) = 0$. The authors stated many regularity conditions on G_n to show that $z_n = \sqrt{n}(\hat{\theta}_n - \bar{\theta})$ converges to a normal distribution (which is a consequence of their Theorem 2.11). In the proof of their theorem 2.11, the authors also employed the invalid MVT formula (1) as we can see in the following:

> "[T]he mean value theorem yields that on C_n
>
> $$G_n(\hat{\theta}_n) - G_n(\bar{\theta}) = \partial_\theta \tilde{G}_n(\hat{\theta}_n - \bar{\theta}).$$
>
> Here, $\partial_\theta \tilde{G}_n$ is the $p \times p$—matrix whose jkth entry is $\partial_\theta \tilde{G}_n(\theta_n^{(j)})_{jk}$, where each $\theta_n^{(j)}$ is a (random) convex combination of $\hat{\theta}_n$ and $\bar{\theta}$."

Clearly, the above MVT is the invalid MVT (1). This expression may be corrected by multiplying both sides of the above equation by $\lambda^\top$, where λ is a vector with the same length as $G_n(\beta)$.

In order to correct the asymptotic arguments in all the three examples, we only need to replace the invalid MVT (1) with the valid MVT (3) and apply the Cramér–Wold device. Alternatively, Theorem 1 could also be applied directly.

4. Concluding Remarks

We conclude that, despite the "Taylor's expansion" presented in Equation (1) for an estimating function is incorrect, the resulting first-order asymptotic theory remains correct. Therefore, although Feng et al.'s (2013) criticism is important and should be considered, it does not affect very much the asymptotic theory of estimators attained from estimating equations.

Appendix A. Appendix: Proof of the Theorem

We use the following lemma in the proof of Theorem 1.

Lemma A.1 (Cramér–Wold's theorem). Let $\{W_n\}_{n\geq 0}$ be a sequence of random p-vectors. Then,

$$W_n \xrightarrow{\mathcal{D}} W_0 \qquad \Longleftrightarrow \qquad \lambda^\top W_n \xrightarrow{\mathcal{D}} \lambda^\top W_0,$$

for all $\lambda \in \mathbb{R}^p$, $\lambda \neq 0$, where "$\xrightarrow{\mathcal{D}}$" stands for "convergence in distribution to."

We begin the proof of the Theorem 1, from identity Equation (3), that

$$\begin{aligned}\lambda^\top \frac{1}{n} U_n(\theta) &= -\lambda^\top \frac{1}{n} U_n'(\theta)(\hat{\theta}_n - \theta) \\ &\quad -\lambda^\top \frac{1}{n}\big[U_n'(\tilde{\theta}_{n,\lambda}) - U_n'(\theta)\big](\hat{\theta}_n - \theta) \qquad (4)\end{aligned}$$

where $\tilde{\theta}_{n,\lambda} = \theta + h_\lambda(\hat{\theta}_n - \theta)$ for some $h_\lambda \in (0, 1)$. Thus, by ii, iii, and iv,

$$O_p(n^{-1/2}) = \lambda^\top I_\theta(\hat{\theta}_n - \theta) + o_p(\| \hat{\theta}_n - \theta \|).$$

As $\| \hat{\theta}_n - \theta \| = o_p(1)$, we have that $\| \hat{\theta}_n - \theta \| = O_p(n^{-1/2})$ and, consequently,

$$\sqrt{n} \left\| \frac{1}{n}\big[U_n'(\tilde{\theta}_{n,\lambda}) - U_n'(\theta)\big] \right\| \| \hat{\theta}_n - \theta \| = o_p(1). \qquad (5)$$

Multiplying the terms in Equation (4) by $n^{-1/2}$, plugging Equation (5) into Equation (4) and employing assumption ii, we obtain

$$\lambda^\top \frac{1}{\sqrt{n}} U_n(\theta) = \lambda^\top \sqrt{n} I_\theta(\hat{\theta}_n - \theta) + o_p(1).$$

As the above equation is valid for each $\lambda \in \mathbb{R}^p$, by Lemma A.1 (the Cramér–Wold's theorem),

$$\frac{1}{\sqrt{n}} U_n(\theta) = \sqrt{n} I_\theta(\hat{\theta}_n - \theta) + o_p(1)$$

and the result follows by multiplying both sides of the above equation by the inverse of I_θ.

References

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120. [1,2]

Barndorff-Nielsen, O. E., and Cox, D. R. (1996), *Inference and Asymptotics* (Texts in Statistical Science), Boca Raton, FL: Chapman & Hall/CRC, p. 87. [1]

Barnett, W. A. (1976). "Maximum Likelihood and Iterated Aitken Estimation of Nonlinear Systems of Equations," *Journal of the American Statistical Association*, 71, 354–360. [1,2]

Feng, C., Wang, H., Han, Y. Xia, Y., and Tu, X. M. (2013), "The Mean Value Theorem and Taylor's Expansion in Statistics," *The American Statistician*, 67, 245–248. [1,2,3]

Ferguson, T. S. (1996), *A Course in Large Sample Theory* (Texts in Statistical Science), Boca Raton, FL: Chapman & Hall/CRC, p. 20, rule 3. [1]

Jacod, J., and Sørensen, M. (2018), "A Review of Asymptotic Theory of Estimating Functions," *Statistical Inference and Stochastic Process*, 21, 415–434. [1,3]

McCullagh, P. (1983), "Quasi-Likelihood Functions," *The Annals of Statistics*, 11, 59–67. [1,2]

Rudin, W. (1976), *Principles of Mathematical Analysis* (3rd ed.), New York: McGraw-Hill. [2]

Shao, J. (2003), *Mathematical Statistics* (2nd ed.), New York: Springer-Verlag. [1]

Serfling, R. J. (2002), *Approximation Theorems of Mathematical Statistics* (Wiley Series in Probability and Statistics), Hoboken, NJ: Wiley. [1]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: University Press. [1]

Wang, J.-L. (1999), "Asymptotic Properties of M-estimators Based on Estimating Equations and Censored Data," *The Scandinavian Journal of Statistics*, 26, 297–318. [1]

Wu, C.-F. (1981), "Asymptotic Theory of Nonlinear Least Squares Estimation," *The Annals of Statistics*, 9, 501–513. [1,2]