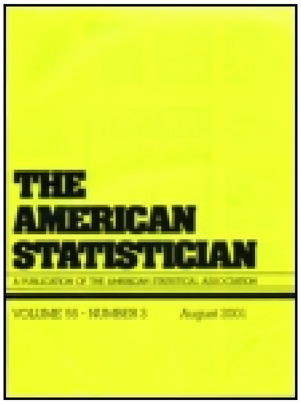

The American Statistician
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/utas20

The Mean Value Theorem and Taylor's Expansion in Statistics

Changyong Feng [a], Hongyue Wang [a], Yu Han [a], Yinglin Xia [a] & Xin M. Tu [a]
[a] Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, 14642, USA

Accepted author version posted online: 02 Oct 2013.Published online: 18 Nov 2013.

To cite this article: Changyong Feng, Hongyue Wang, Yu Han, Yinglin Xia & Xin M. Tu (2013) The Mean Value Theorem and Taylor's Expansion in Statistics, The American Statistician, 67:4, 245-248, DOI: 10.1080/00031305.2013.844203

To link to this article: http://dx.doi.org/10.1080/00031305.2013.844203

The Mean Value Theorem and Taylor's Expansion in Statistics

Changyong FENG, Hongyue WANG, Yu HAN, Yinglin XIA, and Xin M. TU

The mean value theorem and Taylor's expansion are powerful tools in statistics that are used to derive estimators from nonlinear estimating equations and to study the asymptotic properties of the resulting estimators. However, the mean value theorem for a vector-valued differentiable function does not exist. Our survey shows that this nonexistent theorem has been used for a long time in statistical literature to derive the asymptotic properties of estimators and is still being used. We review several frequently cited papers and monographs that have misused this "theorem" and discuss the flaws in these applications. We also offer methods to fix such errors.

KEY WORDS: Asymptotic normality; Consistent estimator; Estimating equation.

1. INTRODUCTION

Many estimators in statistics are obtained through estimating equations. For example, the maximum likelihood estimator (MLE) is the solution of the score equation (Cramér, 1946); the least square estimator in linear regression is the solution of normal equation (Seber and Lee 2003); the generalized estimating equations (GEE) estimator (Liang and Zeger 1986), is the solution of a set of equations. Under mild regularity conditions, the estimators from such estimating equations are consistent and asymptotically normally distributed. Their asymptotic variance can be easily obtained from some robust procedures such as the "sandwich variance estimator" (Huber 1967; White 1982; Liang and Zeger 1986).

Estimating equations are generally nonlinear. Highly efficient numerical methods are available to solve those equations. However, the proof of consistency and asymptotic normality is not so straightforward. Taylor's expansion and the mean value theorem (MVT), which are useful tools in mathematics, have been widely used in statistics to study the asymptotic properties of estimators obtained from nonlinear estimating equations. In calculus we know that the mean value theorem exists for multivariate real-valued differentiable functions. However, a similar mean value theorem does not exist for vector-valued differentiable functions. Our survey shows that this fact is not well appreciated and many frequently cited papers and books in statistics have used this nonexistent theorem in their respective contexts.

In this article, we show how the nonexistent mean value theorem for vector-valued differentiable function has been used in some highly cited journal papers and monographs. We also discuss the reason for such flaws and methods to fix them.

The article is organized as follows. In Section 2 we briefly summarize the Taylor's expansion and MVT, especially for the case of vector-valued multivariate differentiable functions. In Section 3 we discuss some published works that have misused the MVT and Taylor's expansion. In Section 4 we show how to fix the flaws, followed by the conclusion in Section 5.

2. MEAN VALUE THEOREM AND TAYLOR'S EXPANSION OF VECTOR-VALUED FUNCTIONS

We now summarize some well-known results on the mean value theorem and Taylor's expansion in mathematical analysis. All these results can be obtained from any standard book on mathematical analysis such as Rudin (1976).

Suppose that O is an open interval and $f : O \to \mathbb{R}$ is differentiable. Then for any $[a, b] \subset O$, there exists $\theta \in (a, b)$ such that

$$f(b) - f(a) = f'(\theta)(b - a). \tag{1}$$

This is the well-known *mean value theorem* in calculus. This result can be easily generalized to multivariate real-valued differentiable functions. Suppose G is an open convex subset of $\mathbb{R}^p$ $(p > 1)$ and $f : G \to \mathbb{R}$ is a differentiable function. Then for any $\mathbf{a}, \mathbf{b} \in G$, there exists θ between $\mathbf{a}$ and $\mathbf{b}$ such that

$$f(\mathbf{b}) - f(\mathbf{a}) = \nabla f(\theta)(\mathbf{b} - \mathbf{a}), \tag{2}$$

where $\nabla f(\theta)$ (a row vector) is the gradient of f at θ.

However, the MVT for vector-valued differentiable functions does not exist. To see this, consider a vector-valued function f. For an open convex subset $G \subset \mathbb{R}^p$ $(p \geq 1)$, the function $f : G \to \mathbb{R}^q$ $(q > 1)$ is said to be differentiable at $\mathbf{x} \in G$ if there

C. Feng is Associate Professor (E-mail: *feng@bst.rochester.edu*), H. Wang is Research Assistant Professor (E-mail: *cookie@bst.rochester.edu*), Y. Han is Ph.D. candidate (E-mail: *yu_han@urmc.rochester.edu*), Y. Xia is Professor (E-mail: *yinglin_xia@urmc.rochester.edu*) and X. M. Tu is Research Assistant Professor (E-mail: *xin_tu@urmc.rochester.edu*), Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA. This research was supported by a research grant from the Clinical and Translational Sciences Institute at the University of Rochester Medical Center. The authors thank the editor, associate editor, and two referees for their very insightful comments which improved this article.

© 2013 American Statistical Association DOI: 10.1080/00031305.2013.844203

exists a linear operator $Df(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^q$ such that

$$\lim_{\mathbf{x}+\mathbf{h}\in G,\, \mathbf{h}\to\mathbf{0}} \frac{\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = 0. \tag{3}$$

The function f is differentiable in G if it is differentiable at each point of G. From (3)

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + o(\|\mathbf{h}\|). \tag{4}$$

If f is differentiable in G, for any $\mathbf{a} \in G$, the first-order Taylor's expansion of f around $\mathbf{a}$ is

$$f(\mathbf{x}) = f(\mathbf{a}) + Df(\mathbf{a})(\mathbf{x}-\mathbf{a}) + o(\|\mathbf{x}-\mathbf{a}\|). \tag{5}$$

Although (5) is quite similar to (2) and the remainder term in (5) has higher order than the linear part, it does not mean that for any $\mathbf{a}, \mathbf{b} \in G$ there exists θ on the line segment between $\mathbf{a}$ and $\mathbf{b}$ such that

$$f(\mathbf{b}) - f(\mathbf{a}) = Df(\theta)(\mathbf{b}-\mathbf{a}). \tag{6}$$

We call (6) the *nonexistent mean value theorem* (NEMVT) for vector-valued functions.

Here is an example to show why (6) is not true in general. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$f(x, y) = \begin{bmatrix} x + \sin y \\ x + \cos y \end{bmatrix}.$$

This function is continuously differentiable in $\mathbb{R}^2$ with derivative

$$Df(x, y) = \begin{bmatrix} 1 & \cos y \\ 1 & -\sin y \end{bmatrix}.$$

Note that $f(0, 0) = f(0, 2\pi) = (0, 1)^\top$. Assume that there exists (x^*, y^*) on the line segment between $(0, 0)$ and $(0, 2\pi)$ such that (6) holds, then we must have

$$f(0, 2\pi) - f(0, 0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & \cos y^* \\ 1 & -\sin y^* \end{bmatrix} \left\{ \begin{bmatrix} 0 \\ 2\pi \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

This means that $\cos y^* = \sin y^* = 0$, which is impossible.

We give an argument to show why (6) is generally wrong. Our argument is not novel to mathematicians, but may be new to some statisticians.

Let $\mathbf{f} = (f_1, \ldots, f_q)^\top : \mathbb{R}^p \to \mathbb{R}^q$, $p \geq 1, q > 1$ be differentiable. For $k = 1, \ldots, q$, $f_k : \mathbb{R}^p \to \mathbb{R}$ are multivariate real-valued differentiable functions. According to (2), for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, there exists θ_k between $\mathbf{a}$ and $\mathbf{b}$ such that

$$f_k(\mathbf{b}) - f_k(\mathbf{a}) = \nabla f_k(\theta_k)(\mathbf{b}-\mathbf{a}), \quad k = 1, \ldots, q. \tag{7}$$

Note that θ_k may be different for different k. This means in general we cannot find a universal θ between $\mathbf{a}$ and $\mathbf{b}$ such that (7) holds for all k, even when $Df(x)$ is invertible for all $x \in \mathbb{R}^p$. All examples discussed in Section 3 simply assume that θ_k is the same for all k. Although

$$Df(\theta) = \begin{bmatrix} \nabla f_1(\theta) \\ \vdots \\ \nabla f_q(\theta) \end{bmatrix}, \quad \theta \in \mathbb{R}^p,$$

from (7) we can only have for vectors $\theta_1, \ldots, \theta_q$,

$$f(\mathbf{b}) - f(\mathbf{a}) = \begin{bmatrix} \nabla f_1(\theta_1) \\ \vdots \\ \nabla f_q(\theta_q) \end{bmatrix} (\mathbf{b}-\mathbf{a}),$$

which is totally different.

Sometimes the MVT and Taylor's expansion are used interchangeably in the statistical literature. See Example 1 in the next section, in which the Taylor's expansion is actually the NEMVT.

3. SOME MISUSES OF THE MEAN VALUE THEOREM IN STATISTICS

Our survey of the literature shows that the NEMVT (6) has been used quite extensively for decades in the statistical literature. In this section we focus on five highly cited publications (two journal article and three published monographs) to illustrate the extent of the problem. To make our argument precise, we try to use the same notation as in the original works and point out the exact formula where mistakes occur. Please refer to the original works for more details.

Example 1. Wei, Lin, and Weissfeld (1989).

This is a highly cited article in survival analysis, especially in the literature of multivariate survival data and recurrent event failure time data. In this article, the marginal hazard functions of failure times are assumed to be of the form

$$\lambda_k(t) = \lambda_{k0}(t) \exp(\beta_k' Z(t)),$$

where β_k is a $p \times 1$ vector. The marginal working independence partial likelihood was used to estimate the parameter β_k. The score function $U_k(\beta_k, \infty)$ is also a $p \times 1$ vector. To study the asymptotic properties of $\hat{\beta}_k$, Wei, Lin, and Weissfeld (1989) stated "By the Taylor series expansion of $U_k(\beta_k, \infty)$ around β_k, we have"

$$n^{-1/2} U_k(\beta_k, \infty) = \hat{A}_k(\beta_k^*) n^{1/2} (\hat{\beta}_k - \beta_k), \tag{8}$$

where "β_k^* is on the line segment between $\hat{\beta}_k$ and β_k". See formula (A.1) and the expression of $\hat{A}_k(\beta^*)$ in Wei, Lin, and Weissfeld (1989). In fact, $\hat{A}_k(\beta^*)$ is exactly $-DU_k(\beta_k^*, \infty)/n$.

It is obvious that Equation (8) [formula (A.1) in Wei, Lin, and Weissfeld (1989)] is not obtained from the Taylor series expansion but from the application of the NEMVT on the vector-valued function $U_k(\beta_k, \infty)$.

Example 2. Gross and Huber (1987).

Gross and Huber (1987) studied the asymptotic properties of semiparametric estimators of regression parameters in the hazard functions of matched-pair survival data. In their setup, the regression parameter β is a $p \times 1$ vector. The partial likelihood score function $DL_n(\beta)$ (formula 4 of Gross and Huber 1987) is a vector-valued function with p components. To study the asymptotic normality of the maximum partial likelihood estimator $\tilde{\beta}_n$, they used the following expansion of the score function DL_n around the true value β_0:

$$DL_n(\tilde{\beta}_n) = DL_n(\beta_0) + \left(\tilde{\beta}_n - \beta_0\right) D^2 L_n(\beta_n^*), \tag{9}$$

where "D^2 denotes the matrix of second partial derivatives with respect to β, and β_n^* lies between β_0 and $\tilde{\beta}_n$" [see lines 3–6 on p31 of Gross and Huber (1987)].

To study the asymptotic normality of MLE $\hat{\beta}_n$, they expanded the score function DL_n^F around β^* in the following way

$$n^{-1/2}DL_n^F(\hat{\beta}_n) = 0 = n^{-1/2}DL_n^F(\beta^*) + n^{1/2}(\hat{\beta}_n - \beta^*)n^{-1}D^2L_n^F(\beta_n^{**}), \quad (10)$$

"with $\beta_n^{**} \in (\beta^*, \hat{\beta}_n)$" (see Gross and Huber 1987, sec. 3.2).

It is again clear that these two expansions of the vector-valued function used by Gross and Huber (1987) are the application of formula (6).

Example 3. Kalbfleisch and Prentice (2002).

This book is a revised and expanded version of Kalbfleisch and Prentice (1980). Since the publication of the first edition, the book has become an "authority on censoring and likelihood and the hazard function approach to models for several types of failure" (Andersen et al. 1993, p. 7).

In the proof of the asymptotic normality of the Cox partial likelihood estimator $\hat{\beta}$ of parameter vector β (formula 5.59 in Kalbfleisch and Prentice 2002, p. 176), the following formula

$$0 = n^{-1/2}U(\hat{\beta}, \tau) = n^{-1/2}U(\beta_0, \tau) - [n^{1/2}(\hat{\beta} - \beta_0)]'[n^{-1}I(\beta^*, \tau)] \quad (11)$$

is used, "where β^* is between $\hat{\beta}$ and β_0" (Kalbfleisch and Prentice 2002, p. 177).

Example 4. Fleming and Harrington (1991).

This is the first book devoted entirely to the counting process method in survival data. It is a beautifully written book and is accessible to researchers in survival analysis with a solid background in probability theory.

In the proof of the asymptotic normality of the Cox partial likelihood estimator $\hat{\beta}$ of parameter vector β in Fleming and Harrington (1991, p. 299), they used a similar expansion to Kalbfleisch and Prentice (2002) for the score function U,

$$U(\hat{\beta}, \tau) = U(\beta_0, \tau) - \mathcal{I}(\beta^*, \tau)(\hat{\beta} - \beta_0), \quad (12)$$

"where β^* is on a line segment between $\hat{\beta}$ and β_0" (Fleming and Harrington 1991, p. 299). Here $\mathcal{I}(\beta, \tau) = -\partial U(\beta, \tau)/\partial\beta^\top$.

Example 5. Tsiatis (2006).

Tsiatis (2006) is a very nice introduction to semiparametric methods in statistics and is more accessible than some classical books in semiparametric statistical methods such as Bickel et al. (1993), which may be too theoretical for graduate students in statistics and applied researchers.

Unfortunately, the NEMVT was used numerous times in Tsiatis (2006) to derive certain estimators. For example, in the discussion of M-estimator, Tsiatis (2006, p. 30) has the following expansion of estimating equation

$$0 = \sum_{i=1}^n m(Z_i, \hat{\theta}_n) = \sum_{i=1} m(Z_i, \theta_0) + \left\{ \sum_{i=1}^n \frac{\partial m(Z_i, \theta_n^*)}{\partial\theta^\top} \right\}^{p\times p} (\hat{\theta}_n - \theta_0), \quad (13)$$

"where θ_n^* is an intermediate value between $\hat{\theta}_n$ and θ_0" (Tsiatis 2006, p. 30).

4. CAN WE FIX THE FLAWS?

In Section 3, we showed some misuses of the MVT in published papers and monographs, although the conclusions usually remain unchanged in large samples. Here we use the idea of Andersen et al. (1993) to illustrate how to fix the mistake. Consider the problem mentioned earlier in Kalbfleisch and Prentice (2002). Assume that $U(\beta) = (U_1(\beta), \ldots, U_p(\beta))'$, $U(\hat{\beta}) = 0$, and $\hat{\beta}$ is a consistent estimator of β. From MVT (2) we have that for $k = 1, \ldots, p$,

$$0 = n^{-1/2}U_k(\hat{\beta}) = n^{-1/2}U_k(\beta) + n^{-1}\nabla U_k\big(\beta^{(k*)}\big) \cdot \sqrt{n}(\hat{\beta} - \beta),$$

where $\beta^{(k*)}$ is between $\hat{\beta}$ and β. Suppose we can prove that $-n^{-1}\nabla U_k(\beta^{(k*)}) \to \Omega_k$, and $\Omega = (\Omega_1', \ldots, \Omega_p')'$ is nonsingular, then we can prove

$$\sqrt{n}(\hat{\beta} - \beta) = \Omega^{-1} \cdot n^{-1/2}U_k(\beta) + o_p(1).$$

Another way to correct the mistake is to use the integral form of the mean value theorem for vector-valued functions, which states that if $f : \mathbb{R}^p \to \mathbb{R}^q$ is differentiable and Df is continuous in a neighborhood $\{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| < r\}$ of $\mathbf{x}_0$, then for all $\mathbf{t}$ with $\|\mathbf{t}\| < r$,

$$f(\mathbf{x}_0 + \mathbf{t}) - f(\mathbf{x}_0) = \int_0^1 Df(x_0 + u\mathbf{t})du \cdot \mathbf{t}. \quad (14)$$

The proof of this result can be found in Lang (1993, p. 341) and Ferguson (1996, p. 20).

The key to using this result is that if Df is continuous in the open area, then it is uniformly bounded in any compact region contained in the neighborhood of $\mathbf{x}_0$. In this case,

$$\left\| \int_0^1 Df(\mathbf{x}_0 + u\mathbf{t})du - Df(\mathbf{x}_0) \right\| \leq \max_{\mathbf{x}:\|\mathbf{x}-x_0\|\leq\|\mathbf{t}\|} \|Df(\mathbf{x}) - Df(\mathbf{x}_0)\| \to 0$$

as $\mathbf{t} \to 0$. A direct result of this fact is

$$f(\mathbf{x}_0 + \mathbf{t}) - f(\mathbf{x}_0) = Df(\mathbf{x}_0) \cdot \mathbf{t} + o(\|\mathbf{t}\|), \quad (15)$$

which is exactly Equation (4), equivalent to the definition of differentiability.

Ferguson (1996, pp. 45, 122, and 138) used (14) to prove the multivariate δ-method, the asymptotic normality of the MLE, and the equivalence of the one-step estimator to the MLE. Shao (2003, p. 191) also used (14) to prove the asymptotic normality of the MLE.

5. CONCLUSION

Our limited survey establishes misuse of the MVT for vector-valued differentiable functions in the statistical literature for more than two decades. All the examples discussed in Section 3 have been cited numerous times in many statistics/biostatistics publications. Some books discussed in Section 3 have been widely used as textbooks in graduate education or as reference books for researchers. Our article reminds statisticians that such an "MVT" does not exist and appropriate results should be used to ensure the validity of estimators and asymptotic properties derived from estimating equations. As one reviewer pointed out "Our statistical literature should not be perpetuating this error."

[Received March 2013. Revised September 2013.]

REFERENCES

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer. [247]

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Inference in Semiparametric Models*, Baltimore, MD: Johns Hopkins University Press. [247]

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, NJ: Princeton University Press. [245]

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, New York: Chapman & Hall. [247]

Fleming, T. R., and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, Hoboken, NJ: Wiley. [247]

Gross, S. T., and Huber, C. (1987), "Matched Pair Experiments: Cox and Maximum Likelihood Estimation," *Scandinavian Journal of Statistics*, 14, 27–41. [246,247]

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimation Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 221–233. [245]

Lang, S. (1993), *Real and Functional Analysis*, New York: Springer. [247]

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Model," *Biometrika*, 73, 13–22. [245]

Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley. [247]

Kalbfleisch, J. D., and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data* (2nd ed.), Hoboken, NJ: Wiley. [247]

Rudin, W. (1976), *Principles of Mathematical Analysis* (3rd ed.), New York: McGraw-Hill. [245]

Seber, G. A. F., and Lee, A. J. (2003), *Linear Regression Analysis* (2nd ed.), Hoboken, NJ: Wiley. [245]

Shao, J. (2003), *Mathematical Statistics* (2nd ed.), New York: Springer. [247]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [247]

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," *Journal of the American Statistical Association*, 84, 1065–1073. [246]

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [245]