On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors"

by

David A. Freedman

Abstract

The "Huber Sandwich Estimator" can be used to estimate the variance of the MLE when the underlying model is incorrect. If the model is nearly correct, so are the usual standard errors, and robustification is unlikely to help much. On the other hand, if the model is seriously in error, the sandwich may help on the variance side, but the parameters being estimated by the MLE are likely to be meaningless—except perhaps as descriptive statistics.

Introduction

This paper gives an informal account of the so-called "Huber Sandwich Estimator," for which Peter Huber is not to be blamed. We discuss the algorithm, and mention some of the ways in which it is applied. Although the paper is mainly expository, the theoretical framework outlined here may have some elements of novelty. In brief, under rather stringent conditions, the algorithm can be used to estimate the variance of the MLE when the underlying model is incorrect. However, the algorithm ignores bias, which may be appreciable. Thus, results are liable to be misleading.

To begin the mathematical exposition, let i index observations whose values are y_i. Let $\theta \in R^p$ be a $p \times 1$ parameter vector. Let $y \to f_i(y|\theta)$ be a positive density. If y_i takes only the values 0 or 1, which is the chief case of interest here, then $f_i(0|\theta) > 0$, $f_i(1|\theta) > 0$, and $f_i(0|\theta) + f_i(1|\theta) = 1$. Some examples involve real- or vector-valued y_i, and the notation is set up in terms of integrals rather than sums. We assume $\theta \to f_i(y|\theta)$ is smooth. (Other regularity conditions are elided.) Let Y_i be independent with density $f_i(\cdot|\theta)$. Notice that the Y_i are not identically distributed: f_i depends on the subscript i. In typical applications, the Y_i cannot be identically distributed, as will be explained below.

The data are modeled as observed values of Y_i for $i = 1, \ldots, n$. The likelihood function is $\prod_{i=1}^n f_i(Y_i|\theta)$, viewed as a function of θ. The log likelihood function is therefore

$$L(\theta) = \sum_{i=1}^n \log f_i(Y_i|\theta). \tag{1}$$

The first and second partial derivatives of L with respect to θ are given by

$$L'(\theta) = \sum_{i=1}^n g_i(Y_i|\theta), \quad L''(\theta) = \sum_{i=1}^n h_i(Y_i|\theta). \tag{2}$$

To unpack the notation in (2), let ϕ' denote the derivative of the function ϕ: differentiation is with respect to the parameter vector θ. Then

$$g_i(y|\theta) = [\log f_i(y|\theta)]' = \frac{\partial}{\partial\theta} \log f_i(y|\theta), \tag{3}$$

a $1 \times p$-vector. Similarly,

$$h_i(y|\theta) = [\log f_i(y|\theta)]'' = \frac{\partial^2}{\partial\theta^2} \log f_i(y|\theta), \tag{4}$$

a symmetric $p \times p$ matrix. The quantity $-E_\theta h(Y_i|\theta)$ is called the "Fisher information matrix." It may help to note that $-E_\theta h_i(Y_i|\theta) = E_\theta\big(g_i(Y_i|\theta)^T g_i(Y_i|\theta)\big) > 0$, where T stands for transposition.

Assume for the moment that the model is correct, and θ_0 is the true value of θ. So the Y_i are independent and the density of Y_i is $f_i(\cdot|\theta_0)$. The log likelihood function can be expanded in a Taylor series around θ_0:

$$\begin{aligned} L(\theta) = L(\theta_0) &+ L'(\theta_0)(\theta - \theta_0) \\ &+ \frac{1}{2}(\theta - \theta_0)^T L''(\theta_0)(\theta - \theta_0) + \ldots. \end{aligned} \tag{5}$$

If we ignore higher-order terms and write $\doteq$ for "nearly equal"—this is an informal exposition—the log likelihood function is essentially a quadratic, whose maximum can found by solving the likelihood equation $L'(\theta) = 0$. Essentially, the equation is

$$L'(\theta_0) + (\theta - \theta_0)^T L''(\theta_0) = 0. \tag{6}$$

So

$$\hat\theta - \theta_0 \doteq [-L''(\theta_0)]^{-1} L'(\theta_0)^T. \tag{7}$$

Then

$$\mathrm{cov}_{\theta_0}\hat\theta \doteq [-L''(\theta_0)]^{-1}[\mathrm{cov}_{\theta_0} L'(\theta_0)][-L''(\theta_0)]^{-1}, \tag{8}$$

the covariance being a symmetric $p \times p$ matrix.

In the conventional textbook development, $L''(\theta_0)$ and $\mathrm{cov}_{\theta_0} L'(\theta_0)$ are computed—approximately or exactly—using Fisher information. Thus, $-L''(\theta_0) \doteq -\sum_{i=1}^n E_{\theta_0} h_i(Y_i)$. Furthermore, $\mathrm{cov}_{\theta_0} L'(\theta_0) = -\sum_{i=1}^n E_{\theta_0} h_i(Y_i)$. The sandwich idea is to estimate $L''(\theta_0)$ directly from the sample data, as $L''(\hat\theta)$. Similarly, $\mathrm{cov}_{\theta_0} L'(\theta_0)$ is estimated as

$$\sum_{i=1}^n g_i(Y_i|\hat\theta)^T g_i(Y_i|\hat\theta).$$

So (8) is estimated as

$$\hat V = (-A)^{-1} B (-A)^{-1} \tag{9a}$$

where

$$A = L''(\hat\theta) \text{ and } B = \sum_{i=1}^n g_i(Y_i|\hat\theta)^T g_i(Y_i|\hat\theta) \tag{9b}$$

The $\hat V$ in (9) is the "Huber sandwich estimator." The square roots of the diagonal elements of $\hat V$ are "robust standard errors" or "Huber-White standard errors." The middle factor B in (9) is not centered in any way. No centering is needed, because

$$\begin{aligned} E_\theta[g_i(Y_i|\theta)] &= 0, \\ \mathrm{cov}_\theta\big[g_i(Y_i|\theta)\big] &= E_\theta\big[g_i(Y_i|\theta)^T g_i(Y_i|\theta)\big]. \end{aligned} \tag{10}$$

Indeed,

$$\begin{aligned} E_\theta[g_i(Y_i|\theta)] &= \int g_i(y|\theta)\, f_i(y|\theta)\, dy \\ &= \int \frac{\partial}{\partial\theta} f_i(y|\theta)\, dy \\ &= \frac{\partial}{\partial\theta} \int f_i(y|\theta)\, dy \\ &= \frac{\partial}{\partial\theta} 1 \\ &= 0. \end{aligned} \tag{11}$$

A derivative was passed through the integral sign in (11). Regularity conditions are needed to justify such maneuvers, but we finesse these mathematical issues.

If the motivation for the middle factor in (9) is still obscure, try this recipe. Let U_i be independent $1 \times p$-vectors, with $E(U_i) = 0$. Now $\text{cov}(\sum U_i) = \sum \text{cov}(U_i) = \sum E(U_i^T U_i)$. Estimate $E(U_i^T U_i)$ by $U_i^T U_i$. Take $U_i = g_i(Y_i|\theta_0)$. Finally, substitute $\hat\theta$ for θ_0.

The middle factor B in (9) is quadratic. It does not vanish, although

$$\sum_{i=1}^n g_i(Y_i|\hat\theta) = 0. \tag{12}$$

Remember, $\hat\theta$ was chosen to solve the likelihood equation $L'(\theta) = \sum_{i=1}^n g_i(Y_i|\theta) = 0$, explaining (12).

In textbook examples, the middle factor B in (9) will be of order n, being the sum of n terms. Similarly, $-L''(\theta_0) = -\sum_{i=1}^n h_i(Y_i|\theta_0)$ will be of order n: see (2). Thus, (9) will be of order $1/n$. Under suitable regularity conditions, the strong law of large numbers will apply to $-L''(\theta_0)$, so $-L''(\theta_0)/n$ converges to a positive constant; the central limit theorem will apply to $L'(\theta_0)$, so $\sqrt{n}L'(\theta_0)$ converges in law to a multivariate normal distribution with mean 0. In particular, the randomness in L' is of order $\sqrt{n}$. So is the randomness in $-L''$, but that can safely be ignored when computing the asymptotic distribution of $[-L''(\theta_0)]^{-1}L'(\theta_0)^T$, because $-L''(\theta_0)$ is of order n.

Robust standard errors

We turn now to the case where the model is wrong. We continue to assume the Y_i are independent. The density of Y_i, however, is φ_i—which is not in our parametric family. In other words, there is specification error in the model, so the likelihood function is in error too. The sandwich estimator (9) is held to provide standard errors that are "robust to specification error." To make sense of the claim, we need the

Key Assumption. There is a common θ_0 such that $f_i(\cdot|\theta_0)$ is closest—in the Kullback-Leibler sense of relative entropy, defined in (14) below—to φ_i.

(A possible extension will be mentioned, below.) Equation (11) may look questionable in this new context. But

$$E_0\big[g_i(Y_i|\theta)\big] = \int \Big(\frac{\partial}{\partial\theta} f_i(y|\theta)\Big) \frac{1}{f_i(y|\theta)} \varphi_i(y)\, dx$$
$$= 0 \text{ at } \theta = \theta_0. \tag{13}$$

This is because θ_0 minimizes the Kullback-Leibler relative entropy,

$$\theta \to \int \log\left[\frac{\varphi_i(y)}{f_i(y|\theta)}\right] \varphi_i(y)\, dy. \tag{14}$$

By the key assumption, we get the same θ_0 for every i.

Under suitable conditions, the MLE will converge to θ_0. Furthermore, $\hat\theta - \theta_0$ will be asymptotically normal, with mean 0 and covariance $\hat V$ given by (9), that is,

$$\hat V^{-1/2}(\hat\theta - \theta_0) \to N(0_p, I_{p\times p}). \tag{15}$$

By definition, $\hat\theta$ is the θ that maximizes $\theta \to \prod_i f_i(Y_i|\theta)$—although it is granted that Y_i does not have the density $f_i(\cdot|\theta)$. In short, it is a pseudo-likelihood that is being maximized, not a true likelihood. The asymptotics in (15) therefore describe convergence to parameters of an incorrect model that is fitted to the data.

For some rigorous theory in the independent but not identically distributed case, see Amemiya (1985, Section 9.2.2) or Fahrmeir and Kaufmann (1985). For the more familiar IID (independent and identically distributed) case, see Rao (1973, Chapter 6), or Lehmann and Casella (2003, Chapter 6). Lehmann (1998, Chapter 7) and van der Vaart (1998) are less formal, more approachable. These references all use Fisher information rather than (9), and consider true likelihood functions rather than pseudo-likelihoods.

Why not assume IID variables?

The sandwich estimator is commonly used in logit, probit, or cloglog specifications. See, for instance, Gartner and Segura (2000), Jacobs and Carmichael (2002), Gould, Lavy, and Passerman (2004), Lassen (2005), or Schonlau (2006). Calculations are made conditional on the explanatory variables, which are left implicit here. Different subjects have different values for the explanatory variables. Therefore, the response variables have different conditional distributions. Thus, according to the model specification itself, the Y_i are not IID. If the Y_i are not IID, then θ_0 exists only by virtue of the key assumption.

Even if the key assumption holds, bias should be of greater interest than variance, especially when the sample is large and causal inferences are based on a model that is incorrectly specified. Variances will be small, and bias may be large. Specifically, inferences will be based on the incorrect density $f_i(\cdot|\hat\theta) \doteq f_i(\cdot|\theta_0)$, rather than the correct density φ_i. Why do we care about $f_i(\cdot|\theta_0)$? If the model were correct, or nearly correct—that is, $f_i(\cdot|\theta_0) = \varphi_i$ or $f_i(\cdot|\theta_0) \doteq \varphi_i$—there would be no reason to use robust standard errors.

A possible extension

Suppose the Y_i are independent but not identically distributed, and there is no common θ_0 such that $f_i(\cdot|\theta_0)$ is closest to φ_i. One idea is to choose θ_n to minimize the total relative entropy, that is, to minimize

$$\sum_{i=1}^{n} \int \log\left[\frac{\varphi_i(y)}{f_i(y|\theta)}\right] \varphi_i(y)\, dy. \tag{16}$$

Of course, θ_n would depend on n, and the MLE would have to be viewed as estimating this moving parameter. Many technical details remain to be worked out. For discussion along these lines, see White (1994, pp. 28–30, pp. 192–195).

Cluster samples

The sandwich estimator is often used for cluster samples. The idea is that clusters are independent, but subjects within a cluster are dependent. The procedure is to group the terms in (9), with one group for each cluster. If we denote cluster j by c_j, the middle factor in (9) would be replaced by

$$\sum_j \left[\sum_{i\in c_j} g_i(Y_i|\hat\theta)\right]^T \left[\sum_{i\in c_j} g_i(Y_i|\hat\theta)\right]. \tag{17}$$

The two outside factors in (9) would remain the same. The results of the calculation are sometimes called "survey-corrected" variances, or variances "adjusted for clustering."

There is undoubtedly a statistical model for which the calculation gives sensible answers, because the quantity in (17) should estimate the variance of $\sum_j \left[\sum_{i\in c_j} g_i(Y_i|\hat\theta)\right]$—if clusters are independent and $\hat\theta$ is nearly constant. (Details remain to be elucidated.) It is quite another thing to say what is being estimated by solving the non-likelihood equation $\sum_{i=1}^{n} g_i(Y_i|\theta) = 0$. This is a non-likelihood equation because $\prod_i f_i(\cdot|\theta)$ does not describe the behavior of the individuals comprising the population. If it did, we would not be bothering with robust standard errors in the first place. The sandwich estimator for cluster samples presents exactly the same conceptual difficulty as before.

The linear case

The sandwich estimator is often conflated with the correction for heteroscedasticity in White (1980). Suppose $Y = X\beta + \epsilon$. We condition on X, assumed to be of full rank. Suppose the ϵ_i are independent with expectation 0, but not identically distributed. The OLS estimator is $\hat\beta_{\text{OLS}} = (X'X)^{-1}X'Y$. White proposed that the covariance matrix of $\hat\beta_{\text{OLS}}$ should be estimated as $(X'X)^{-1}X'\hat G X(X'X)^{-1}$, where $e = Y - X\hat\beta_{\text{OLS}}$ is the vector of residuals, $\hat G_{ij} = e_i^2$ if $i = j$, and $\hat G_{ij} = 0$ if $i \neq j$. Similar ideas can be used if the ϵ_i are independent in blocks. White's method often gives good results, although $\hat G$ can be so variable that t-statistics are surprisingly non-t-like. Compare Beck, Katz, Alvarez, Garrett, and Lange (1993).

The linear model is much nicer than other models, because $\hat\beta_{\text{OLS}}$ is unbiased even in the case we are considering, although OLS may of course be inefficient, and—more important—the usual SEs may be wrong. White's correction tries to fix the SEs.

An example

Suppose there is one real-valued explanatory variable, x, with values x_i spread fairly uniformly over the interval from 0 to 10. Given the x_i, the response variables Y_i are independent, and

$$\text{logit}\, P(Y_i = 1) = \alpha + \beta x_i + \gamma x_i^2, \tag{18}$$

where logit $p = \log[p/(1-p)]$. Equation (18) is a logit model with a quadratic response. The sample size is moderately large. However, an unwitting statistician fits a linear logit model,

$$\text{logit}\, P(Y_i = 1) = a + b x_i. \tag{19}$$

If γ is nearly 0, for example, then $\hat{a} \doteq \alpha$, $\hat{b} \doteq \beta$, and all is well—with or without the robust SEs. Suppose, however, that $\alpha = 0$, $\beta = -3$, and $\gamma = .5$. (The parameters are chosen so the quadratic has a minimum at 3, and the probabilities spread out through the unit interval.) The unwitting statistician will get $\hat{a} \doteq -5$ and $\hat{b} \doteq 1$, concluding that on the logit scale, a unit increase in x makes the probability that $Y = 1$ go by one, across the whole range of x. The only difference between the usual SEs and the robust SEs is the confidence one has in this absurd conclusion.

In truth, for x near 0, an unit increase in x makes the probability of a response go *down*, by 3 (probabilities are measured here on the logit scale). For x near 3, increasing x makes no difference. For x near 10, a unit increase in x makes the probability go up by 7.

Could the specification error be detected by some kind of regression diagnostics? Perhaps, especially if we knew what kind of specification errors to look for. Keep in mind, however, that the robust SEs are designed for use when there is *undetected* specification error.

What about Huber?

The usual applications of the so-called "Huber sandwich estimator" go far beyond the mathematics in Huber (1967), and our critical comments do not apply to his work. In free translation—this is no substitute for reading the paper—he assumes the Y_i are IID, so $f_i \equiv f$, and $g_i \equiv g$, and $h_i \equiv h$. He considers the asymptotics when the true density is f_0, not in the parametric family. Let $A = \int h(y|\theta_0) f_0(y)\, dy$, and $B = \int g(y|\theta_0)^T g(y|\theta_0) f_0(y)\, dy$. Both are $p \times p$ symmetric matrices. Plainly, $L'(\theta_0) = \frac{1}{n} \sum_{i=1}^n g(Y_i|\theta_0)$. Under regularity conditions discussed in the paper,

(i) $\hat{\theta} \to \theta_0$, which minimizes the "distance" between $f(\cdot|\theta)$ and f_0.
(ii) $\frac{1}{n} L''(\theta_0) = \frac{1}{n} \sum_{i=1}^n h(X_i|\theta_0) \to A$.
(iii) $n^{1/2} B^{-1/2} L'(\theta_0) \to N(0_p, I_{p\times p})$.

Asymptotic normality of the MLE follows:

$$C_n^{-1/2}(\hat{\theta} - \theta_0) \to N(0_{p\times 1}, I_{p\times p}), \tag{20a}$$

where

$$C_n = n^{-1}(-A)^{-1} B (-A)^{-1}. \tag{20b}$$

Thus, Huber's paper answers a question that (for a mathematical statistician) seems quite natural: what is the asymptotic behavior of the MLE when the model is wrong? Applying the algorithm to

data, while ignoring the assumptions of the theorems and the errors in the models—that is not Peter Huber.

Summary and conclusions

The sandwich algorithm, under stringent regularity conditions, yields variances for the MLE that are asymptotically correct even when the specification—and hence the likelihood function—are incorrect. However, it is quite another thing to ignore bias. It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter.

More particularly, inferences are based on a model that is admittedly incorrect. (If the model were correct, or nearly correct, there would be no need for sandwiches.) The chief issue, then, is the difference between the incorrect model that is fitted to the data and the process that generated the data. This is bias due to specification error. The algorithm does not take bias into account. Applied papers that use sandwiches rarely mention bias. There is room for improvement here.

See Koenker (2005) for additional discussion. On White's correction, see Greene (2003, p. 220). For a more general discussion of independence assumptions, see Berk and Freedman (2003) or Freedman (2005). The latter reference also discusses model-based causal inference in the social sciences.

References

Amemiya, T. (1985). *Advanced Econometrics*, Harvard University Press.

Beck, N., Katz, J. N., Alvarez, R. M., Garrett, G., and Lange, P. (1993). "Government Partisanship, Labor Organization, and Macroeconomic Performance," *American Political Science Review*, 87, 945–48.

Fahrmeir, L., Kaufmann, H. (1985). "Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models," *The Annals of Statistics*, 13, 342–68.

Berk, R. A. and Freedman, D. A. (2003). "Statistical Assumptions as Empirical Commitments," in T. G. Blomberg and S. Cohen, eds., *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd ed., Aldine de Gruyter, New York, pp. 235–54.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*, Cambridge University Press.

Gartner, S. S. and Segura, G. M. (2000). "Race, Casualties, and Opinion in the Vietnam War," *Journal of Politics*, 62, 115–146.

Gould, E. D., Lavy, V., and Passerman, M. D. (2004). "Immigrating to Opportunity: Estimating the Effect of School Quality Using a Natural Experiment on Ethiopians in Israel," *Quarterly Journal of Economics*, vol. 119, pp. 489–526.

Greene, W. H. (2003). *Econometric Analysis*, 5th ed., Prentice Hall.

Huber, P. J. (1967). "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, pp. 221–33.

Jacobs, D. and Carmichael, J. T. (2002). "The Political Sociology of the Death Penalty," *American Sociological Review*, 67, 109–31.

Koenker, R. (2005). "Maximum Likelihood Asymptotics under Nonstandard Conditions: A Heuristic Introduction to Sandwiches," www.econ.uiuc.edu/~roger/courses/476/lectures/L10.pdf

Lassen, D. D. (2005). "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment," *American Journal of Political Science*, 49, 103–18.

Lehmann, E. L. (1998). *Elements of Large-Sample Theory*, Springer.

Lehmann, E. L. and Casella, G. (2003). *Theory of Point Estimation*, 2nd ed., Springer.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed., Wiley.

van der Vaart, A. (1998). *Asymptotic Statistics*, Cambridge University Press.

White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–38.

White, H. (1994). *Estimation, Inference, and Specification Analysis*, Cambridge University Press.

Author's footnote. Dick Berk (Penn), Peter Westfall (Texas Tech), and Paul Ruud (Berkeley) made helpful comments.

Department of Statistics
UC Berkeley, CA 94720-3860
September, 2006