Bayesian Interpretations of Heteroskedastic Consistent Covariance Estimators Using the Informed Bayesian Bootstrap

Dale J. Poirier

University of California, Irvine

September 1, 2008

Abstract

This paper provides Bayesian rationalizations for White's *heteroskedastic consistent* (HC) covariance estimator and various modifications of it. An informed Bayesian bootstrap provides the statistical framework.

1. Introduction

Often researchers find it useful to recast frequentist procedures in Bayesian terms, so as to get a clearer understanding of how and when the procedures work. This paper takes this approach with respect to the consistent, in the face of heteroskedasticity of ***unknown*** form, covariance matrix estimator proposed by White (1980) for the case of stochastic regressors, building on the fixed regressor analysis of Eicker (1963, 1967), and the misspecified maximum likelihood analysis of Huber (1967).

2. Preliminaries

This section reviews a Bayesian conjugate analysis of a multinomial sampling distribution. Let $\boldsymbol{z}$ denote a discrete random variable belonging to one of J categories indexed (j = 1, 2, ..., J) with associated probabilities Prob($\boldsymbol{z}$ = j$|\theta$) = θ_j (j = 1, 2, ..., J), where $\theta = [\theta_1, \theta_2, ..., \theta_J]' \in \Theta^J$ and Θ^J denotes the unit simplex in $\Re^J$. n independent draws $z = [z_1, z_2, ..., z_n]'$ on $\boldsymbol{z}$ imply the multinomial likelihood function

$$\mathcal{L}(\theta; z) = \left(\frac{n!}{n_1!\, n_2! \,....\, n_J!} \right) \prod_{j=1}^{J} \theta_j^{n_j} \propto \prod_{j=1}^{J} \theta_j^{n_j}, \tag{1}$$

where $\boldsymbol{N_j} = n_j$, the number of z_j (i = 1, 2, ..., n) equal to j (j = 1, 2, ..., J), are sufficient statistics for θ. The natural conjugate prior for θ is the Dirichlet distribution $D_J(\underline{\nu})$ with density

$$f^D(\theta|\underline{\nu}) \;=\; \frac{\Gamma\left(\sum_{j=1}^{J}\underline{\nu}_j\right)}{\prod_{j=1}^{J}\Gamma(\underline{\nu}_j)} \prod_{j=1}^{J}\theta_j^{\underline{\nu}_j - 1}, \quad \theta \in \Theta^J, \tag{2}$$

given $\underline{\nu}_j > 0$ (j = 1, 2, ..., J) and $\underline{\nu} = [\underline{\nu}_1, \underline{\nu}_2, ... \underline{\nu}_J]'$.[1] Without loss of generality assume the first m categories have $\boldsymbol{N_j} = n_j > 0$. It follows from Bayes' Theorem that the posterior distribution of θ is the Dirichlet distribution $D_J(\bar{\nu})$, where $\bar{\nu} = [\bar{\nu}_1, \bar{\nu}_2, ..., \bar{\nu}_J]'$, $\bar{\nu}_j = \underline{\nu}_j + n_j$ (j = 1, 2, ..., m), and $\bar{\nu}_j = \underline{\nu}_j$ (j = m+1, m+2, ..., J). The marginal mass function of z is the multinomial-Dirichlet mass function

$$f^{MD}(z|\underline{\nu}) \;=\; \frac{\Gamma\left(\sum_{j=1}^{J}\underline{\nu}_j\right)}{\Gamma\left(n + \sum_{j=1}^{J}\underline{\nu}_j\right)} \prod_{j=1}^{m}\frac{\Gamma(\underline{\nu}_j + n_j)}{\Gamma(\underline{\nu}_j)}. \tag{3}$$

The predictive mass function for a new observation z_{N+1} is

$$\mathrm{Prob}(z_{N+1} = j|z) \;=\; E(\theta_j|z) \;=\; \frac{\bar{\nu}_j}{\sum_{i=1}^{J}\bar{\nu}_i} \quad (j = 1, 2, ..., J). \tag{4}$$

Sampling from a Dirichlet distribution can be accomplished by simulating J independent gamma random variables $g = [g_1, g_2, ..., g_J]'$ with means/variances ν_j (j = 1, 2, ..., J) and unit scale parameters, i.e., with densities

$$f_\gamma(g_j|\nu_j) \;=\; [\Gamma(\nu_j)]^{-1} g_j^{\nu_j - 1}\exp(-g_j), \quad g_j > 0,\; \nu_j > 0, \quad (j = 1, 2, ..., J), \tag{5}$$

and then forming

$$\theta_j \;=\; \frac{g_j}{g_1 + g_2 + ... + g_J} \quad (j = 1, 2, ..., J). \tag{6}$$

[1] Note: $E(\theta|\underline{\nu}) = (\iota_J'\underline{\nu})^{-1}\underline{\nu}$ and $\mathrm{Var}(\theta|\underline{\nu}) = [(\iota_J'\underline{\nu})^2(\iota_J'\underline{\nu} + 1)]^{-1}\left[(\iota_J'\underline{\nu})\,\mathrm{diag}\{\underline{\nu}\} - \underline{\nu}\,\underline{\nu}'\right]$.

The resulting θ_j constitutes one draw from the Dirichlet distribution $D_J(\nu)$. When $\nu = \underline{\nu}$, the draw θ is from the prior density $f_D(\theta|\underline{\nu})$; when $\nu = \bar{\nu}$, the draw θ is from the posterior density $f_D(\theta|\bar{\nu})$.

3. Bayesian Bootstrap

Interpret the multinomial categories of $\boldsymbol{z}$ as J distinct values taken by a (K+1)×1 vector $\boldsymbol{z} = [\boldsymbol{y}, \boldsymbol{x}']'$, where $\boldsymbol{y}$ is scalar and $\boldsymbol{x}$ is K×1.[2] In other words, the J categories are reinterpreted as the support points for the joint distribution of $\boldsymbol{z}$, i.e., $\boldsymbol{z} = j$ is reinterpreted as $\boldsymbol{z} = z_j$ where $z_j = [\boldsymbol{y}_j, \boldsymbol{x}_j']'$. Suppose m of the J support points for $\boldsymbol{z}$ are observed in a sample of size $n \geq m$ and $\tilde{J}$ others are not observed. Denote the m observed points by $z^{(1)} = [y^{(1)}, X^{(1)}]'$, where $y^{(1)}$ is m×1 and $X^{(1)}$ is m×K, and denote the $\tilde{J}$ unobserved support points by $z^{(2)} = [y^{(2)}, X^{(2)}]'$, where $y^{(2)}$ is $\tilde{J}\times 1$ and $X^{(2)}$ is $\tilde{J}\times K$. Note that $N_j = 0$ ($j = m+1, m+2, ..., m+\tilde{J}$). Similarly, partition θ, g, $\underline{\nu}$, and $\bar{\nu}$ into m and $\tilde{J}$ components corresponding to $z^{(1)}$ and $z^{(2)}$: $\theta = [\theta^{(1)\prime}, \theta^{(2)\prime}]'$, $g = [g^{(1)\prime}, g^{(2)\prime}]'$, $\underline{\nu} = [\underline{\nu}^{(1)\prime}, \underline{\nu}^{(2)\prime}]'$, and $\bar{\nu} = [\bar{\nu}^{(1)\prime}, \bar{\nu}^{(2)\prime}]'$. The ***standard case*** corresponds to data that are distinct and exhaust the support: $J = n = m$, $N_1 = N_2 = ... = N_J = 1$, $\tilde{J} = 0$, and $z^{(2)}$ is null. See Chamberlain and Imbens (2003), Hahn (1977), Lancaster (2003; 2004, Section 3.4), and Ragusa (2007). The ***general case*** is any non-standard case.

The Bayesian bootstrap of Rubin (1981) begins by selecting a parameter/functional of interest, say, $\beta = \beta(\theta)$. For example, in the context of a linear model, suppose interest focuses on the moment condition $E_{\mathbf{z}|\theta}[\boldsymbol{x}(\boldsymbol{y} - \boldsymbol{x}'\beta)] = 0_K$.[3] Then $\beta = \beta(\theta)$ are the coefficients associated with the linear projection of $\boldsymbol{y}$ on $\boldsymbol{x}$: $\beta = [E_{\mathbf{z}|\theta}(\boldsymbol{x}'\boldsymbol{x})]^{-1} E_{\mathbf{z}|\theta}(\boldsymbol{x}'\boldsymbol{y})$, where

$$E_{\mathbf{z}|\theta}(\boldsymbol{x}'\boldsymbol{x}) = X'\mathrm{diag}\{\theta\}X = X^{(1)\prime}\mathrm{diag}\{\theta^{(1)}\}X^{(1)} + X^{(2)\prime}\mathrm{diag}\{\theta^{(2)}\}X^{(2)}, \qquad (7)$$

$$E_{\mathbf{z}|\theta}(\boldsymbol{x}'\boldsymbol{y}) = X'\mathrm{diag}\{\theta\}y = X^{(1)\prime}\mathrm{diag}\{\theta^{(1)}\}y^{(1)} + X^{(2)\prime}\mathrm{diag}\{\theta^{(2)}\}y^{(2)}, \qquad (8)$$

with $X = [X^{(1)\prime}, X^{(2)\prime}]'$, $y = [y^{(1)\prime}, y^{(2)\prime}]'$, and diag{·} denotes a diagonal matrix with diagonal

[2] Chamberlain and Imbens (2003, p. 12) note that because J can be arbitrarily large and our data are measured with finite precision, the finite support assumption is arguably not restrictive.

[3] Chamberlain and Imbens (2003, p. 13) discuss how this approach can be extended to the overidentified case in which the number of moment conditions exceeds the dimension of β.

elements equal to the given argument. β has the ***weighted least squares (WLS)*** representation [4,5]

$$\beta = \beta(\theta) = \left[X'\operatorname{diag}\{\theta\}X\right]^{-1}X'\operatorname{diag}\{\theta\}\,y. \qquad (9)$$

Note that the “parameter” β depends on both $z^{(1)} = [y^{(1)}, X^{(1)`}]$ and $z^{(2)} = [y^{(2)}, X^{(2)`}]$, as well as θ.

The prior and posterior distributions for θ in Section 2 induce prior and posterior distributions for β via (9), albeit not particularly tractable, because β is a nonlinear function of the θ. The limiting (as $\underline{v} \to 0_J$) “noninformative” improper prior together with the standard case ($\tilde{J}$ = 0 and $z^{(2)}$ null) have received most of the attention. This is unfortunate. Because the number of hyperparameters in $\underline{v}$ is J, which is greater than or equal to the sample size n, an ***informative*** prior is warranted. Unfortunately, this necessitates eliciting a large number of prior hyperpararmeters.

This paper introduces an informative prior for θ with two properties: all elements in $\underline{v}$ are positive implying prior density (2) is proper, and unobserved support points (fictitious observations) $z^{(2)} = [y^{(2)}, X^{(2)}]'$ (together with $\underline{v}^{(2)}$) are introduced to reflect prior beliefs about β. The resulting combination is defined to be an ***informed Bayesian bootstrap***.

The multinomial probabilities θ are not the parameters of interest. The functional $\beta(\theta)$ is the focus of attention. Choosing $\underline{v}$, so that the implied prior for β is sensible, is a challenging task. An obvious reference case is the ***symmetric prior*** where all elements in $\underline{v}$ are the same. Then θ_j (j = 1, 2, ..., J) are identically distributed, and the prior and posterior for β are centered over $b = (X'X)^{-1}X'y$.

For example, consider the standard case and ***symmetric prior*** in which θ_j (j = 1, 2, ..., J) are identically distributed with

$$\underline{v}_j = c \qquad (j = 1, 2, ..., J), \qquad c > 0. \qquad (10)$$

Then $E(\theta_j) = J^{-1}$ (j = 1, 2, ..., J), regardless of c, and $\mathrm{Var}(\theta_j) = (J - 1)[[J^2(Jc + 1)]^{-1}$ which is decreasing in c. As $c \to \infty$, the elements of θ_j (j = 1, 2, ..., J) become degenerate random variables equal to J^{-1}.

[4] Since the denominator in (6) is common across j, Lancaster (2003) notes that β also has the WLS representation $\beta = \beta(g) = [X'\operatorname{diag}\{g\}X]^{-1}X'\operatorname{diag}\{g\}y$. This leads to a slightly different approximation than discussed in the text.

[5] Gilstein and Learner (1983) characterize the set of WLS estimates as the union of the bounded convex polytopes formed by the hyperplanes on which residuals are zero.

Because $\bar{\nu} = (c+1)\iota_J$, $\iota_J{}'\bar{\nu} = J(c+1)$, the posterior mean of θ is the same as the prior mean J^{-1}, however, the posterior variance $\mathrm{Var}(\theta_j) = (J - 1)[[J^2(J\{c+1\} + 1)]^{-1}$ is smaller than the prior variance. In other words, the hyperparameter c only affects the prior and posterior dispersions. Setting c = .005 corresponds to a fairly diffuse proper prior, whereas, the uniform case c = 1 corresponds to a fairly tight prior. Jeffreys' prior corresponds to c = .50.

Note the standard case assumption in the preceding example is important for the posterior analysis. Suppose only m of the possible J support points are observed. Let $n^{(1)} = [n_1, n_2, ..., n_m]'$. Then $\bar{\nu} = [\bar{\nu}^{(1)\prime}, \bar{\nu}^{(2)\prime}]'$, where $\bar{\nu}^{(1)} = \underline{\nu}^{(1)} + n^{(1)}$ and $\bar{\nu}^{(2)} = \underline{\nu}^{(2)}$. In this case, the posterior mean of θ consists of two different subvectors $E(\theta^{(1)}|z) = \dfrac{\bar{\nu}^{(1)}}{\iota_m{}'\bar{\nu}^{(1)} + \iota_{\tilde{J}}{}'\underline{\nu}^{(2)}} = \dfrac{c\iota_m + n^{(1)}}{cJ + n}$ and $E(\theta^{(2)}|z) = \dfrac{\bar{\nu}^{(2)}}{\iota_m{}'\bar{\nu}^{(1)} + \iota_{\tilde{J}}{}'\underline{\nu}^{(2)}} = \dfrac{c\iota_{\tilde{J}}}{cJ + n}$. In other words, the symmetry of the prior does ***not*** carry over to the posterior. Heterogeneous priors are introduced in Section 6.

4. HC Estimation

Lancaster (2003) showed the equivalence to order J^{-1} of the covariance matrix of the bootstrap distribution of the OLS and the Eicker/White/Huber heteroskedasticity robust covariance matrix estimate. Furthermore, he showed that the covariance matrices of the Efron (1979) bootstrap and the Bayesian bootstrap of Rubin (1981) are identical to $O(J^{-1})$.

Lancaster uses a limiting (as $\underline{\nu} \to 0_N$) "noninformative" improper prior in the standard case ($\tilde{J} = 0$ and $z^{(2)}$ is null). Extending Lancaster (2003) to the ***general case*** in which $\tilde{J} > 0$ and $z^{(2)}$ is non-null, this paper approximates the posterior distribution of $\beta = \beta(\theta)$ in (9) by taking a Taylor series expansion around the posterior mean $\bar{\theta} \equiv E(\theta|z) = \bar{\nu}/(\iota_J{}'\bar{\nu})$, where ι_J is a J×1 vector of ones. This leads to the approximation

$$\begin{aligned}
\beta^*(\theta) &= \beta(\bar{\theta}) + \left[\frac{\partial\,\beta(\theta)}{\partial\,\theta'}\right]_{\theta=\bar{\theta}} (\theta - \bar{\theta}) \\
&= \beta(\bar{\theta}) + \left[[y - X\beta(\bar{\theta})]' \otimes (X'\mathrm{diag}\left\{\bar{\theta}\right\}X)^{-1}X'\right]\left[\frac{\partial\,\mathrm{vec}(\mathrm{diag}\{\theta\})}{\partial\,\theta'}\right]_{\theta=\bar{\theta}} (\theta - \bar{\theta}) \qquad (11) \\
&= \beta(\bar{\theta}) + \left[u(\bar{\theta})' \otimes Q(\bar{\theta})'\right]\begin{bmatrix} e_1 e_1{}' \\ \vdots \\ e_J e_J{}' \end{bmatrix}(\theta - \bar{\theta}) \\
&= \beta(\bar{\theta}) + R(\bar{\theta})'(\theta - \bar{\theta}),
\end{aligned}$$

where e_j is a J×1 vector with the j^{th} element equal to unity and all other elements equal to zero, $Q(\bar{\theta}) = X(X'\text{diag}\{\bar{\theta}\}X)^{-1}$, $u(\bar{\theta}) = y - X\beta(\bar{\theta})$, and

$$R(\bar{\theta}) \;=\; [e_1e_1', ..., e_Je_J']\left[u(\bar{\theta})\otimes Q(\bar{\theta})\right] \;=\; \text{diag}\left\{u(\bar{\theta})\right\}Q(\bar{\theta}). \tag{12}$$

The posterior $\theta|z \sim D_J(\bar{\nu})$ implies that the exact posterior mean and variance of (11) are

$$E[\beta^*(\theta)|z] \;=\; \beta(\bar{\theta}), \tag{13}$$

$$\begin{aligned} V^{HC} &= \text{Var}[\beta^*(\theta)|z] \\ &= R(\bar{\theta})'\,\text{Var}(\theta|z)\,R(\bar{\theta}) \\ &= (\iota_J'\bar{\nu}+1)^{-1}Q(\bar{\theta})'\,\text{diag}\left\{u(\bar{\theta})\right\}\left[\text{diag}\left\{\bar{\theta}\right\}-\bar{\theta}\bar{\theta}'\right]\text{diag}\left\{u(\bar{\theta})\right\}Q(\bar{\theta}). \end{aligned} \tag{14a}$$

Since the first-order condition for WLS implies $(X'\text{diag}\{\bar{\theta}\}X)^{-1}X'\text{diag}\{\bar{\theta}\}u(\bar{\theta}) = 0_K$, and because

$$\begin{aligned} \text{diag}\left\{u(\bar{\theta})\right\}\left[\text{diag}\left\{\bar{\theta}\right\}-\bar{\theta}\bar{\theta}'\right]\text{diag}\left\{u(\bar{\theta})\right\} \;=\; &\text{diag}\left\{\left[\bar{\theta}_1[u(\bar{\theta})]^2, ..., \bar{\theta}_J[u(\bar{\theta})]^2\right]'\right\} \\ &- \text{diag}\left\{\bar{\theta}\right\}u(\bar{\theta})u(\bar{\theta})'\,\text{diag}\left\{\bar{\theta}\right\}. \end{aligned} \tag{15}$$

it follows that (14a) simplifies to

$$\begin{aligned} V^{HC} &= (\iota_J'\bar{\nu}+1)^{-1}Q(\bar{\theta})'\,\text{diag}\left\{\left[\bar{\theta}_1[u(\bar{\theta})]^2, ..., \bar{\theta}_J[u(\bar{\theta})]^2\right]'\right\}Q(\bar{\theta}) \\ &= (\iota_J'\bar{\nu}+1)^{-1}(X'\text{diag}\left\{\bar{\theta}\right\}X)^{-1}\left[X'\,\text{diag}\left\{\bar{\theta}_1[u(\bar{\theta})]^2, ..., \bar{\theta}_J[u(\bar{\theta})]^2\right\}X\right](X'\text{diag}\left\{\bar{\theta}\right\}X)^{-1} \\ &= \frac{\iota_J'\bar{\nu}}{(\iota_J'\bar{\nu}+1)}(X'\,\text{diag}\left\{\bar{\nu}\right\}X)^{-1}\left[X'U^{HC}X\right](X'\,\text{diag}\left\{\bar{\nu}\right\}X)^{-1}, \end{aligned} \tag{14b}$$

where $U^{HC} \equiv U^{HC}(\bar{\nu}) \equiv \text{diag}\left\{\bar{\nu}_1[u_1(\bar{\nu})]^2, ..., \bar{\nu}_J[u_J(\bar{\nu})]^2\right\}$. To the extent that $\beta^*(\theta)$ in (11) is a good approximation to $\beta(\theta)$ in (9), (13) and (14b) should be good approximations to the posterior mean and variance of β.

To add some transparency to (14b), consider symmetric prior (10) and the standard case. Then, the posterior mean is $\bar{\theta} \equiv E(\theta|z) = J^{-1}\iota_J$, $\beta(\bar{\theta}) = b$, and (14b) simplifies to

$$V^{HC} \;=\; \left(\frac{J}{J(c+1)+1}\right)V^{HC0}, \tag{16}$$

where V^{HC0} is the White/Eicker robust covariance matrix

$$V^{HC0} \;=\; (X'X)^{-1}X'U^{HC0}X(X'X)^{-1}, \tag{17}$$

$U^{HC0} \equiv \text{diag}\left\{(u_1^{OLS})^2, \ldots, (u_J^{OLS})^2\right\}$ and $u^{OLS} = y - Xb$. As $c \to 0$ (Lancaster's case), $V^{HC} \to [J/(J+1)]V^{HC0}$. Therefore, the noninformative symmetric prior provides a Bayesian justification for using the OLS estimate b together with V^{HC0} in approximate Bayesian inference for β.

5. Choice of $z^{(2)} = [y^{(2)}, X^{(2)}]$

$z^{(j)} = [y^{(j)}, X^{(j)}]$ $(j = 1, 2)$ play similar roles in defining $\beta = \beta(\theta)$ in (9) and transforming prior beliefs about θ into prior beliefs about β. Approximate the ***prior*** distribution of $\beta(\theta)$ by taking a Taylor series expansion around the prior mean $\underline{\theta} \equiv E(\theta) = \underline{\nu}/(\iota_J{}'\underline{\nu})$. Then analogous to (13), the prior mean of β can be approximated by

$$\begin{aligned}
\beta(\underline{\theta}) &= \beta(\underline{\nu}) = [X'\text{diag}\{\underline{\theta}\}X]^{-1}X'\text{diag}\{\underline{\theta}\}y \\
&= [X'\text{diag}\{\underline{\nu}\}X]^{-1}X'\text{diag}\{\underline{\nu}\}y \qquad (18)\\
&= \left[X^{(1)\prime}\text{diag}\{\underline{\nu}^{(1)}\}X^{(1)} + X^{(2)\prime}\text{diag}\{\underline{\nu}^{(2)}\}X^{(2)}\right]^{-1}\left[X^{(1)\prime}\text{diag}\{\underline{\nu}^{(1)}\}y^{(1)} + X^{(2)\prime}\text{diag}\{\underline{\nu}^{(2)}\}y^{(2)}\right].
\end{aligned}$$

One goal in choosing $z^{(2)} = [y^{(2)}, X^{(2)}]$ may be to locate (18) at an *a priori* likely value. For example, suppose $\tilde{J} = n$ and once again consider symmetric prior (10). Then all the diagonal matrices in (18) cancel. Often an interesting prior for public consumption centers all the slopes in (18) over zero. Choose the first column of $X^{(2)}$ equal to $-\iota_{\tilde{J}}$, the remaining columns of $X^{(2)}$ equal to the corresponding columns of $X^{(1)}$, and $y^{(2)} = -y^{(1)}$. Then $\beta(\underline{\theta}) = [\bar{y}, 0_{K-1}{}']'$.

Next consider approximate posterior mean (13). Because

$$X'\text{diag}\{\bar{\nu}\}X = X'\text{diag}\{\underline{\nu}\}X + X^{(1)\prime}\text{diag}\{n^{(1)}\}X^{(1)}, \qquad (19)$$

$$X'\text{diag}\{\bar{\nu}\}y = X'\text{diag}\{\underline{\nu}\}y + X^{(1)\prime}\text{diag}\{n^{(1)}\}y^{(1)}, \qquad (20)$$

(13) has the representation

$$\begin{aligned}
\beta(\bar{\theta}) &= [X'\text{diag}\{\bar{\theta}\}X]^{-1}X'\text{diag}\{\bar{\theta}\}y \\
&= [X'\text{diag}\{\bar{\nu}\}X]^{-1}X'\text{diag}\{\bar{\nu}\}y \\
&= \left[X'\text{diag}\{\underline{\nu}\}X + X^{(1)\prime}\text{diag}\{n^{(1)}\}X^{(1)}\right]^{-1}\left[X'\text{diag}\{\underline{\nu}\}y + X^{(1)\prime}\text{diag}\{n^{(1)}\}y^{(1)}\right] \qquad (21)\\
&= \left[X'\text{diag}\{\underline{\nu}\}X + X^{(1)\prime}\text{diag}\{n^{(1)}\}X^{(1)}\right]^{-1} \\
&\qquad \left[\left(X'\text{diag}\{\underline{\nu}\}X\right)\beta(\underline{\theta}) + \left(X^{(1)\prime}\text{diag}\{n^{(1)}\}X^{(1)}\right)\beta^{(1)}(n^{(1)})\right],
\end{aligned}$$

which is a matrix-weighted average of approximate prior mean (18) and the WLS estimate based on the observed $z^{(1)} = [y^{(1)}, X^{(1)}]$:

$$\beta^{(1)}(n^{(1)}) = \left[X^{(1)\prime}\,\mathrm{diag}\left\{n^{(1)}\right\}X^{(1)}\right]^{-1}X^{(1)\prime}\,\mathrm{diag}\left\{n^{(1)}\right\}y^{(1)}. \tag{22}$$

These fictitious observations enter approximate prior mean (18) and posterior mean (21) through the K(K+1)/2 unique elements in $X^{(2)\prime}\mathrm{diag}\left\{\underline{\nu}^{(2)}\right\}X^{(2)}$, and the K×1 location vector $X^{(2)\prime}\mathrm{diag}\left\{\underline{\nu}^{(2)}\right\}y^{(2)}$ [or $\beta^{(2)}(\underline{\nu}^{(2)})$]. The actual $\tilde{J}$ (K+1) elements in $z^{(2)}$ are ***not*** required to compute the approximation (18) or (21). This is no more onerous than the customary linear regression case.

6. Choice of $\underline{\nu}$

Symmetric prior (10) centers on OLS. When θ_j (j = 1, 2, ..., J) are ***not*** identically distributed, then attention switches to WLS. But unfortunately, the not identically distributed case requires choice of J hyperparameters: $\underline{\nu}_j$ (j = 1, 2, ..., J). This section considers extensions of White/Eicker/ Huber, motivated by sampling properties of V^{HC0}, that suggest choices for $\underline{\nu}_j$ (j = 1, 2, ..., J).

The White/Eicker robust covariance matrix (17) has been criticized by numerous authors for its sampling distribution (e.g., it tends to underestimate variances and test statistics have the wrong size).[6] Numerous suggestions have been made for adding weights to the diagonal elements of U^{HC0}. These weights can be interpreted as a choice for $\bar{\nu}_j$(j = 1, 2, ..., J) in $U^{HC}(\bar{\nu})$, from which values of prior hyperparameters $\underline{\nu}_j$(j = 1, 2, ..., J) can be backed out. The resulting $\underline{\nu}$ results in ***proper*** priors and a well-defined marginal mass function (3).

An early critic of (17), predating White (1980), was Hinkley (1977) who suggested using b and scaling-up V^{HC0} by J/(J - K):

$$V^{HC1} = [J/(J-K)]V^{HC0}. \tag{23}$$

While this reminiscent of symmetric prior for the standard case, it is different in an important way: unlike (23), (16) scales V^{HC0} down, not up. Hence, a Bayesian justification of (23) is unclear.

Other modifications of V^{HC0} involve diagonal elements h_j (j = 1, 2, ..., J) of the so-called "hat

[6] See Chesher and Jewitt(1987), Cribari-Neto et al. (2000), Cribari-Neto and Zarkos (1999, 2001), Godfrey (2006), and MacKinnon and White (1985).

matrix" $X(X'X)^{-1}X'$. Cribari-Neto and Zarkos (2001) showed that the presence of high leverage points, as measured by h_j (j = 1, 2, ..., J), in the design matrix is more decisive for the finite-sample behavior of different HC robust covariance matrix estimators than the degree of heteroskedasticity.

Table 1 summarizes the effects of adding weights to the diagonal elements of U^{HC0}, using a common notation HC0-HC4 in the literature. Besides the original White/Eicker/Huber case (HC0) and Hinkley's (HC1), Table 1 contains the modifications by Horn, Horn, et al. (1975) (HC2), MacKinnon and White (1985) (HC3), and Cribari-Neto (2004) (HC4). In addition Table 1contains three Bayesian analogs (denoted HC2a, HC3a, and HC4a) that give three different choices for $\underline{\nu}_j$ (j = 1, 2, ..., J) that imply the same diagonal weights for U^{HC0} suggested by HC2-HC4. HC2a-HC4a, however, center inferences over $\beta(\bar{\nu})$ (instead of b) and use $u(\bar{\theta})$ instead of the OLS residuals.

7. Summary

The Bayesian bootstrap provides a Bayesian semiparametric analysis of the linear model not requiring a distributional assumption other than the multinomial sampling. Using the Bayesian bootstrap and a symmetric improper prior in the standard case, Lancaster (2003) showed the White/Eicker/Huber robust standard errors are reasonable asymptotic approximations to the posterior standard deviations around OLS. Using a proper (possibly heterogeneous) prior in the general case, this paper has extended Lancaster's results and linked the White/Eicker/Huber robust standard errors to posterior Bayesian analysis, but with two important differences: the posterior location for $\beta = \beta(\theta)$ is WLS not OLS, and the posterior covariance matrix (14b), while of a similar form to (16), is evaluated at WLS residuals instead of OLS residuals.[7]

Choice of $z^{(2)} = [y^{(2)}, X^{(2)}]'$ and $\underline{\nu}$ is critical. Section 5 suggested choosing $z^{(2)}$ to locate the prior over zero slopes. Section 6 drew on various modifications of the White/Eicker/Huber estimator to suggest choices for $\underline{\nu}$.

[7] Leamer (1991?) calls the treatment of heteroskedasticity that alters the size but not the location of confidence sets ***"White-washing."*** He argues an ***increase*** in the standard deviations after White-correction signals a potentially ***large*** change in the location of the confidence sets. On the other hand, if the White-correction leaves unchanged or ***reduces*** the standard errors, he argues the estimates are likely to be ***insensitive*** to reweighting.

References

Chamberlain, G. and G. W. Imbens, 2003, “Nonparametric Applications of Bayesian Inference,” *Journal of Business & Economic Statistics* 21, 12-18.

Chesher, A. and I. Jewitt, 1987, “The Bias of a Heteroskedasticity Consistent Covariance Matrix Estimator,” *Econometrica* 55, 1217-1222.

Cribari-Neto, F., 2004, “Asymptotic Inference Under Heteroskedasticity of Unknown Form,” *Computational Statistics & Data Analysis* 45, 215-233.

Cribari-Neto, F., S. L. P. Ferrari, and G. M. Cordeiro, 2000, “Improved Heteroscedasticity-Consistent Covariance Matrix Estimators,” *Biometrika* 87, 907–918.

Cribari-Neto, F., S. G. Zarkos, 1999, “Bootstrap Methods for Heteroskedastic Regression Models: Evidence on Estimation and Testing,” *Econometric Reviews* 18, 211–228.

Cribari-Neto, F., S. G. Zarkos, 2001, “Heteroskedasticity-Consistent Covariance Matrix Estimation: White’s Estimator and the Bootstrap,” *Journal of Statistical Computing and Simulation* 68, 391-411.

Cribari-Neto, F. and N. Galvão, 2003, “A Class of Improved Heteroskedasticity-Consistent Covariance Matrix Estimators,” *Communications in Statistics: Theory and Methods* 32, 1951-1980.

Efron, B., 1979, “Bootstrap Methods: Another Look at the Jackknife,” *Annals of Statistics* 7, 1-26.

Eicker, F., 1963, “Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions,” *Annals of Mathematical Statistics* 34, 447-456.

Eicker, F., 1967, “Limit Theorems for Regressions with Unequal and Dependent Errors,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press, 59-82.

Godfrey, L. G., 2006, “Tests for Regression Models with Heteroskedasticity of Unknown Form,” *Computational Statistics & Data Analysis* 50, 2715-2733,

Hahn, J.,1977, “Bayesian Bootstrap of the Quantile Regression Estimator: A Large Sample Study,” *International Economic Review* 38, 206–213.

Hinkley, D. V., 1977, “Jackknifing in Unbalanced Situations,” *Technometrics* 19, 285-292.

Horn, S. D., Horn, R. A., Duncan, D. B., 1975, “Estimating Heteroskedastic Variances in Linear Models,” *Journal of the American Statistical Association* 70, 380-385.

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 (Berkeley: University of California Press), 221–233.

Kauemann, G. and R. J. Carroll, 2001, “A Note on the Efficiency of Sandwich Covariance Matrix Estimation,” *Journal of the American Statistical Association* 96, 1387-1396.

Lancaster, T., 2003, “A Note on Bootstraps and Robustness,” unpublished manuscript, Brown University.

Lancaster, T., 2004, *An Introduction to Modern Bayesian Econometrics* (Oxford: Blackwell Publishing).

MacKinnon, J. and H. White (1985), “Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties,” *Journal of Econometrics* 29, 305-325.

Ragusa, G., 2007, “Bayesian Likelihoods for Moment Condition Models,” unpublished manuscript, University of California, Irvine.

Rubin, D., 1981, “The Bayesian Bootstrap,” *Annals of Statistics* 9, 130-134.

White, H., 1980, “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica* 48, 817-838.

Table 1: Choice of $\underline{\nu}$

	Authors	$\underline{\nu}_j$	$E[\beta^*(g)\|z]$	U^{HC}	V^{HC}
HC0	White/Eicker/Huber	$\to 0$	b	$\mathrm{diag}\left\{[u_1^{OLS}]^2,\ ...,\ [u_J^{OLS}]^2\right\}$	$(X'X)^{-1}X'U^{HC0}X(X'X)^{-1}$
HC1	Hinkley		b	$\left(\frac{J}{J-K}\right)\mathrm{diag}\left\{[u_1^{OLS}]^2,\ ...,\ [u_J^{OLS}]^2\right\}$	$\left(\frac{J}{J-K}\right)V^{HC0}$
HC2	Horn, et al.		b	$\mathrm{diag}\left\{\frac{[u_1^{OLS}]^2}{1-h_1},\ ...,\ \frac{[u_J^{OLS)}]^2}{1-h_J}\right\}$	$(X'X)^{-1}X'U^{HC2}X(X'X)^{-1}$
HC2a		$\frac{h_j}{1-h_j}$	$\beta(\bar{\nu})$	$\mathrm{diag}\left\{\frac{[u_1(\bar{\nu})]^2}{1-h_1},\ ...,\ \frac{[u_J(\bar{\nu})]^2}{1-h_J}\right\}$	$\kappa(X'\,\mathrm{diag}\{\bar{\nu}\}X)^{-1}X'U^{HC2a}X\,(X'\,\mathrm{diag}\{\bar{\nu}\}X)^{-1}$
HC3	MacKinnon/White		b	$\mathrm{diag}\left\{\frac{[u_1^{OLS}]^2}{(1-h_1)^2},\ ...,\ \frac{[u_J^{OLS}]^2}{(1-h_J)^2}\right\}$	$(X'X)^{-1}X'U^{HC2}X(X'X)^{-1}$
HC3a		$\frac{h_j(2-h_j)}{(1-h_j)^2}$	$\beta(\bar{\nu})$	$\mathrm{diag}\left\{\frac{[u_1(\bar{\nu})]^2}{(1-h_1)^2},\ ...,\ \frac{[u_J(\bar{\nu})]^2}{(1-h_J)^2}\right\}$	$\kappa(X'\,\mathrm{diag}\{\bar{\nu}\}X)^{-1}X'U^{HC3a}X\,(X'\,\mathrm{diag}\{\bar{\nu}\}X)^{-1}$
HC4	Cribari-Neto		b	$\mathrm{diag}\left\{\frac{[u_1^{OLS}]^2}{(1-h_1)^{\delta_1}},\ ...,\ \frac{[u_J^{OLS}]^2}{(1-h_J)^{\delta_J}}\right\}$	$(X'X)^{-1}X'U^{HC3}X(X'X)^{-1}$
HC4a		$\frac{1-(1-h_j)^{\delta_j}}{(1-h_j)^{\delta_j}}$	$\beta(\bar{\nu})$	$\mathrm{diag}\left\{\frac{[u_1(\bar{\nu})]^2}{(1-h_1)^{\delta_1}},\ ...,\ \frac{[u_J(\bar{\nu})]^2}{(1-h_J)^{\delta_J}}\right\}$	$\kappa(X'\,\mathrm{diag}\{\bar{\nu}\}X)^{-1}X'U^{HC4a}X\,(X'\,\mathrm{diag}\{\bar{\nu}\}X)^{-1}$

Note: $\beta(\bar{\nu}) = [X'\,\mathrm{diag}\{\bar{\nu}\}X]^{-1}X'\,\mathrm{diag}\{\bar{\nu}\}y$, $\bar{\nu}_j = \underline{\nu}_j + n_j$ (j = 1, 2, ..., m), $\bar{\nu}_j = \underline{\nu}_j$ (j = m+1, m+2, ..., J), $\delta_j = \min\left\{4,\ \frac{h_j}{K/N}\right\}$ $(j = 1, 2, ..., J)$, and $\kappa = \iota_J'\bar{\nu}/(\iota_J'\bar{\nu} + 1)$.