Bayesian Heteroskedasticity-Robust Standard Errors

Richard Startz*

August 2012

Abstract

Use of heteroskedasticity-robust standard errors has become common in frequentist regressions. I offer here a Bayesian analog. The Bayesian version is derived by first focusing on the likelihood function for the sample values of the identifying moment conditions of least squares and then formulating a convenient prior for the variances of the error terms. The first step introduces a sandwich estimator into the posterior calculations, while the second step allows the investigator to set the sandwich for either heteroskedastic or homoskedastic error variances. If desired, the Bayesian estimator can be made to look very similar to the usual heteroskedasticity-robust frequentist estimator. Bayesian estimation is easily accomplished by a standard MCMC procedure.

* Department of Economics, 2127 North Hall, University of California, Santa Barbara, CA 93106, email: startz@econ.ucsb.edu. Advice from Doug Steigerwald is gratefully acknowledged.

Introduction

Estimation of heteroskedasticity-consistent (aka "robust") standard errors following the work of White (1980), Eicker (1967), and Huber (1967) has become routine in the frequentist literature. (For a recent retrospective see MacKinnon (2012). Freedman (2006) is also of interest.) Indeed, White (1980) was *the* most cited article in economics between 1980 and 2005 (Kim (2006)). While a number of authors have proposed Bayesian approaches to heteroskedasticity, no one has presented the direct Bayesian analogue to the frequentist approach.[2] The Bayesian version may be useful both in estimation of models subject to heteroskedasticity and in situations where such models arise as blocks of a larger Bayesian problem.

In the frequentist framework, definition of an estimator and derivation of an estimate of the distribution of that estimator generally proceed sequentially. For Bayesians the estimate can't be separated from its distribution; there is simply a posterior. For a regression subject to heteroskedastic errors the Bayesian equivalent of GLS is straightforward, but as with frequentist GLS the presence of heteroskedasticity affects the mean of the posterior. The idea of Bayesian robust standard errors is to allow heteroskedasticity to affect the spread of the posterior without changing its mean.

It turns out that the principal "trick" to finding robust standard errors for a Bayesian regression is to focus on the likelihood function for the moment conditions that identify the coefficients, rather than the likelihood function for the data generating process. The posterior for the coefficients can be made to closely mimic the frequentist OLS/robust standard error

[2] Poirier (2008) provides an extensive analysis of Bayesian rationalizations for White's estimator using the Bayesian bootstrap. Another recent example, which focuses on the sandwich estimator, is Szprio, Rice, and Lumley (2010).

distribution. The second contribution here is to offer a prior of convenience that can be parameterized to give a posterior for the error variances, conditional on the coefficients, to mimic either a homoskedastic model or the frequentist robust variance estimator.

Given the coefficient posterior conditional on the error variances and the error variance posterior conditional on the coefficients, a Gibbs sampler is completely straightforward. As an illustration, I re-examine a hedonic housing price model that has been the subject of a number of Bayesian estimates.

To make clear notation, the problem under consideration is the classic least squares model with normal, independent, but possibly heteroskedastic errors.

$$\begin{aligned} &y = X\beta + \epsilon, \epsilon \sim N(0, \Sigma = \sigma^2\Lambda), \\ &\Sigma = \mathrm{E}(\epsilon\epsilon') = \begin{cases} \sigma^2\lambda_i, & i = j \\ 0, & i \neq j \end{cases} \end{aligned} \tag{1}$$

where X is $n \times k$. Assuming that X is nonstochastic or that the analysis is conditional on X, which we shall do henceforth, the generalized least squares estimator is given by $\beta_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$ and the ordinary least squares estimator is given by $\beta_{OLS} = (X'X)^{-1}X'y$. The frequentist distribution of the estimators is respectively $\beta_{GLS} \sim N(\beta, (X'\Sigma^{-1}X)^{-1})$ and $\beta_{OLS} \sim N(\beta, (X'X)^{-1}\Omega(X'X)^{-1})$, $\Omega \equiv X'\Sigma X = \sigma^2 X'\Lambda X$. The terms on either side of Σ in the estimation of Ω give rise to the name "sandwich estimator," which plays a role in what follows. If Σ is known, then both GLS and OLS estimators are available. The GLS estimator is likely preferred on efficiency grounds. Since equation (1) can be transformed into a homoskedastic regression by pre-multiplying both sides by the Cholesky factorization of Σ^{-1}, the Bayesian version of the generalized least squares estimator is straightforward.

If Σ^{-1} is unknown but can be consistently estimated by $\widehat{\Sigma^{-1}}$—for example by modeling λ_i as a function of a small number of parameters—then β can be estimated by feasible GLS which will be well-behaved in large samples. A text book version of the analogous "feasible GLS" Bayesian procedure is given in section 6.3 of Koop (2003). Alternatively, the size of the parameter space can be limited in a Bayesian analysis through use of a hierarchical prior. The outstanding example of this is probably Geweke (1993) who showed an equivalence between the normal heteroskedastic model and a model with Student-t errors.

Despite GLS' possible efficiency advantages, frequentists very often prefer OLS estimates with robust standard errors because use of an estimated error variance-covariance matrix can lead to bias in GLS coefficient estimates. The breakthrough that permitted robust standard errors was recognition that while GLS requires a weighted average of Σ^{-1}, robust standard errors require a weighted average of Σ. For reasons reviewed briefly below, the latter can be well-estimated with fewer restrictions.

Bayesian Analysis

For a Bayesian analysis in which the investigator has meaningful priors for λ_i^2, one can simply proceed with Bayesian GLS. After all, if one is happy with the model for drawing λ_i^2 then the draw for Σ^{-1} and $X'\Sigma^{-1}X$ follow immediately (and in an Markov Chain Monte Carlo (MCMC) context follow trivially). The more difficult situation is when one adopts priors of convenience while relying on a large number of observations so that the posterior will be dominated by the likelihood function with the influence of the convenience prior being small. The issue is not really different from the reason frequentists might choose OLS over GLS.

Suppose one observes $\epsilon_i, i = 1, \ldots, n$. Then $\epsilon_i^2 \sim (\sigma^2\lambda_i) \times \chi_1^2$ with mean $\sigma^2\lambda_i$ and variance $(\sigma^2\lambda_i)^2$. The squared errors ϵ_i^2 and ϵ_j^2.are independent. The reciprocal is $1/\epsilon_i^2 \sim I\Gamma(1/2\,, 1/2\sigma^2\lambda_i)$, where the inverse gamma density $I\Gamma(\alpha, \beta)$ follows the convention in Greenberg (2008) $f_{I\Gamma}(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-(\alpha+1)} \exp[-\beta/z]$. The reciprocal has no finite moments. It is sometimes said that the difficulty in estimating the GLS weighting matrix is that we have the same number of precision parameters as we have observations. While true, this is not quite to the point as we require only an estimate of the $k(k+1)/2$ parameters in $X'\Sigma^{-1}X$. Rather, the problem is that with no finite moments for Σ^{-1} the law of large numbers does not apply to $X'\Sigma^{-1}X$. In contrast, the moments for Σ are well-behaved. As a result while estimation of $X'\Sigma^{-1}X$ is problematic, estimation of $X'\Sigma X$ is straightforward so long as n is sufficiently large.

The frequentist approach to robust standard errors, in which an estimator is first defined and then robust standard errors are derived, is something of an unnatural act in a Bayesian framework since the posterior arises with no separation of point estimate and distribution around the estimate. It turns out that the requisite analytic trick is to recognize that when making a sandwich it's best to lay down the bread before the filling.

The coefficients in a regression are identified by the k moment conditions $\mathrm{E}(X'\epsilon) = 0$. Focus on the likelihood function for the sample moments $X'\epsilon$ (with $k(k+1)/2$ variance parameters) rather than the data generating process for the regression (with n variance parameters). (This technique which may be useful for moment-based estimators other than least squares as well.)

Consider pre-multiplying equation (1) by X' to start the sandwich.

$$X'y = X'X\beta + X'\epsilon, X'\epsilon \sim N(0, \Omega) \tag{2}$$

Conditional on Ω (i. e. σ^2 and Λ), equation (2) is simply a regression model with correlated errors $X'\epsilon$. If one assumes a normal prior for β, $\beta \sim N(\beta_0, V_0)$, independent of σ^2 and Λ, then the conditional posterior follows from the standard formula for Bayesian GLS.[3]

$$\begin{aligned} &\beta|y, \sigma^2, \Lambda \sim N\left(\bar{\beta}, \bar{V}\right) \\ &\bar{V} = \left(V_0^{-1} + (X'X)'\Omega^{-1}(X'X)\right)^{-1} \\ &\bar{\beta} = \bar{V}(V_0^{-1}\beta_0 + (X'X)'\Omega^{-1}[(X'y)]) \end{aligned} \tag{3}$$

Equations (2) and (3) appear odd on the surface, since the smörgås'ed mean equation has only k observations. Note, however, that $X'X$, $X'y$, and Ω are all composed of summations with n terms, so the right hand terms in the posterior expressions all converge in probability.. Thus, unlike $(X'\Sigma^{-1}X)^{-1}$, Ω^{-1} is well behaved in large samples.

Consider what happens to the conditional posterior in equation (3) as the prior precision V_0^{-1} approaches zero. Very conveniently, the posterior variance $\bar{V} \rightarrow \left((X'X)'\Omega^{-1}(X'X)\right)^{-1} = (X'X)^{-1}\Omega(X'X)^{-1}$ and the posterior mean $\bar{\beta} \rightarrow \left((X'X)'\Omega^{-1}(X'X)\right)^{-1}\left((X'X)'\Omega^{-1}(X'y)\right) = (X'X)^{-1}((X'X)'\Omega^{-1})^{-1}(X'X)'\Omega^{-1}(X'y) = (X'X)^{-1}X'y$. In other words, as the prior precision becomes small relative to the information from the data, the posterior for β approaches the classical least squares distribution with robust standard errors.

[3] The straightforward regression analog to the third line of equation (3) is a regression with $X'y$ as the dependent variable, $X'X$ as the matrix of independent variables, and Ω as the variance-covariance matrix of the errors. Thus the conditional posterior mean is $\bar{\beta} = \bar{V}\left(\underline{V}^{-1}\beta + (X'X)'\Omega^{-1}(X'X)[(X'X)'\Omega^{-1}(X'X)]^{-1}[(X'X)'\Omega^{-1}(X'y)]\right)$, which simplifies to the expression given in equation (3).

Returning to equation (1), draws of σ^2 are straightforward. Conditional on β and Λ, the standardized errors $\epsilon_i/\sqrt{\lambda_i}$ are observed. Thus the draw for σ^2 is as from a standard regression model. If we assume the prior for σ^2 is $I\Gamma\left(\frac{\nu_0+2}{2}, \frac{\sigma_0^2\nu_0}{2}\right)$ then the conditional posterior is

$$\sigma^2|y, \beta, \Lambda \sim I\Gamma\left(\frac{\alpha_1}{2}, \frac{\delta_1}{2}\right)$$

$$\alpha_1 = \nu_0 + 2 + n$$

$$\hat{e}_i = \frac{(y_i - X_i\beta)}{\sqrt{\lambda_\mathrm{i}}}$$

$$\delta_1 = \sigma_0^2\nu_0 + \hat{e}'\hat{e}$$

In specifying prior parameters it may be useful to note that $\mathrm{E}(\sigma^2) = \sigma_0^2, \nu_0 > 0$, $\mathrm{var}(\sigma^2) = \frac{2\left(\sigma_0^2\right)^2}{\nu_0 - 2}, \nu_0 > 2$.

The final step is to find Λ conditional on β, σ^2. One approach, consistent with Geweke (1993), is to assume independent inverse gamma priors for λ_i. A very convenient parameterization is $\lambda_i \sim I\Gamma(a, a-1), a > 1$. Note this gives prior expectation $\mathrm{E}(\lambda_i) = 1$ and, for $a > 2$, $\mathrm{var}(\lambda_i) = 1/(a-2)$. Conditional on β, ϵ_i and therefore ϵ_i^2 is observable from equation (1). The likelihood for $\epsilon_i^2|\sigma^2, \lambda_i$ is $\Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2\lambda_i}\right)$, where the gamma density $\Gamma(\alpha, \beta)$ is $f_\Gamma(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} \exp(-\beta z)$. It follows immediately that the conditional posterior is characterized by

$$\lambda_i|\epsilon_i^2 \sim I\Gamma\left(a+\frac{1}{2}, \frac{\epsilon_i^2}{2\sigma^2}+a-1\right)$$
$$\mathrm{E}(\lambda_i|\epsilon_i^2) = \frac{\frac{\epsilon_i^2}{2\sigma^2}+a-1}{a-\frac{1}{2}}, a>1 \qquad (4)$$
$$\mathrm{var}(\lambda_i|\epsilon_i^2) = \frac{\left(\frac{\epsilon_i^2}{2\sigma^2}+a-1\right)^2}{\left(a-\frac{1}{2}\right)^2(a-1.5)}, a>1.5$$

Judicious choice of a allows equation (4) to represent either heteroskedastic or homoskedastic errors. Note that as the prior parameter $a \to 1$, then $\mathrm{E}(\lambda_i|\epsilon_i^2) \to \epsilon_i^2/\sigma^2$ so that the conditional posterior mean is the same as the frequentist estimate for the heteroskedastic variance. As $a \to \infty$, the prior and posterior both converge in probability to 1, indicating a homoskedastic model. In the latter case, σ^2 is identified separately from Λ by the homoskedastic prior. Note also that while conditional posteriors are given for all n elements of Λ, just as in the frequentist case all that we make use of is the $k \times k$ matrix $X'\sigma^2\Lambda X$. The investigator can choose intermediate values of a to allow for a limited amount of heteroskedasticity. See Geweke (1993) or the textbooks by Greenberg (section 4.5) or Koop (section 6.4) for discussion of hierarchal priors for a.

In summary, a diffuse prior corresponding to the frequentist robust standard error model consists in setting V_0 large, ν_0 just above 2, and a just above 1.

Bayesians generally have less concern than do frequentists with the asymptotic behavior of estimators. Nonetheless, a look at equation (3) shows that the conditional posterior for β converges in distribution to the frequentist, robust distribution. Hence, the estimator for

β is consistent. In contrast, examination of equation (4) shows that the distribution of the heteroskedastic variance terms does not collapse with a large sample. In other words the results are the same as White's and others, the variance of the regression coefficients is well-identified in a large sample but the individual error variances are not.

Illustrative Example

As an illustration I re-examine Koop's (2003) version of Anglin and Gençay's (1996) hedonic regression of house prices in Windsor, Canada, as Koop provides five Bayesian estimates of this model using different assumptions.[4] For our purposes I provide least squares results with both homoskedastic and robust standard errors and, for comparison, results of the estimator presented above with relatively noninformative priors and with both a set to 10,000 to enforce a homoskedasticity assumption and a set to 1.001 to allow for heteroskedastic errors. As a further example, I provide a heteroskedastic version with mildly informative priors suggested by Koop that can be compared to Koop's implementation of Geweke's model. The substantive model regresses sales price (CDN$) on a constant, lot size (sq. feet), the number of bedrooms, the number of bathrooms, and the number of stories in the house. There are 546 observations. Results for Gibbs sampling with 10,000 draws retained after a burn-in of 5,000 draws are reported in Table 1.[5]

[4] The data is available at http://www.wiley.com/legacy/wileychi/koopbayesian/datasets.html and at http://qed.econ.queensu.ca/jae/1996-v11.6/anglin-gencay/.

[5] Estimation required 0.26 milliseconds per draw on a fast vintage 2011 PC running Matlab, so computation time is not an issue. Results for 100,000 draws after 50,000 burn-ins generally agreed with the results in the table to two significant digits.

	Least squares	Robust Bayesian Estimate (noninformative priors)		Robust Bayesian Estimate (Koop priors)	Student-t model
		$a = 1.001$	$a = 10{,}000$	$a = 1.001$	
Intercept	-4,010 (3,603) [3,651]	-3,986 [4,963]	-3,999 (3,608)	-3,953 [4,063]	-413 (2,898)
Lot size	5.43 (0.37) [0.46]	5.43 [0.58]	5.43 (0.37)	5.49 [0.55]	5.24 (0.36)
Bedrooms	2,825 (1,215) [1,257]	2,817 [1,681]	2,819 (1,214)	3,461 [1,317]	2,118 (972)
Bathrooms	17,105 (1,734) [2,253]	17,111 [2,849]	17,093 (1,746)	15,146 [2,353]	14,910 (1,666)
Stories	7,635 (1,008) [913]	7,634 [1,312]	7,638 (1,006)	7,795 [1,219]	8,109 (956)
Notes: mean estimates with standard deviations in parentheses and robust standard errors in brackets.					

Table 1

The leftmost column of Table 1 give least squares results with both classical and robust standard errors. The next two columns give Bayesian results with relatively noninformative priors. Specifically, $\beta_0 = 0, V_0 = 10^{16} \times I_5, \nu_0 = 2.1$, and $\sigma_0^2 = \text{var}(y)$. In column (2) $a = 1.001$, so heteroskedasticity is allowed for, while the estimates in column (3) force homoskedasticity with $a = 10{,}000$. As expected with $n = 546$, mean coefficient estimates are essentially the same as from least squares. The numerical standard deviations from the Bayesian homoskedastic estimate are also essentially the same as the frequentist standard

errors. The numerical standard deviations from the Bayesian heteroskedastic model are relatively close to, but somewhat larger than, the frequentist robust OLS standard errors.

The final two columns in Table 1 allow for a comparison to Geweke's method. The penultimate column shows the heteroskedasticity-robust Bayesian results using mildly informative priors for β suggested by Koop. The right-most column, taken from Koop (2003) Table 6.2, gives Koop's estimate of Geweke's student-t model with the same priors. The differences are not terribly large, with the exception of the intercept, suggesting that for this application the choice of prior is at least as important as how heteroskedasticity is handled.

Conclusion

The Bayesian analog to the now classical frequentist approach to robust standard errors in the regression model is straightforward for linear regression. The first step is to model the sample moment conditions. This works because the GLS estimator for the moments, Bayesian or frequentist, is essentially the same as the regression procedure with robust standard errors. In addition to linear regression, this step is likely to apply to models as well. The second step is to model the sandwich estimator for the coefficient variance-covariance matrix, which is straightforward for heteroskedastic errors. MCMC estimation is simple to implement and the illustrative example gives the expected results.

References

Anglin, P., Gençay, R. 1996. Semiparametric estimation of a hedonic price function, *Journal of Applied Econometrics,* 11, 633-648.

Eicker, F., 1967. Limit theorems for regression with unequal and dependent errors, *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics, Vol. 1,* Berkeley: University of California Press.

Freedman, D., 2006. On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors," *The American Statistician*, 60:4.

Geweke, J., 1993. Bayesian treatment of the independent student-t linear model', *Journal of Applied Econometrics*, vol. 8, no. S1, pp. S19-S40.

Greenberg, E., 2008. *Introduction to Bayesian Econometrics,* Cambridge, Cambridge University Press.

Huber, P., 1967. The behavior of maximum likelihood estimates under nonstandard conditions, *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics, Vol. 1,* Berkeley: University of California Press.

Kim, E.H., Morse, A., Zingales, L. 2006. What has mattered to economics since 1970, *Journal of Economic Perspectives* 20, No. 4.

Koop, G., 2003. *Bayesian Econometrics,* Chichester, John Wiley and Sons.

MacKinnon, J., 2012. Thirty years of heteroskedasticity-robust inference, Queen's Economics Department Working Paper No. 1268.

Poirier, D. 2008., Bayesian interpretations of heteroskedastic consistent covariance estimators using the informed Bayesian bootstrap, University of California, Irvine working paper.

Spiro, A.A., Rice, K.M., Lumley, T. 2010, Model-robust regression and a bayesian ‘sandwich’ estimator, *Annals of Applied Statistics,* 4, No. 4.

White, H., 1980. A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica,* 48, 1980.