Bayesian Heteroskedasticity-Robust Regression

Richard Startz[*]

revised February 2015

Abstract

I offer here a method for Bayesian heteroskedasticity-robust regression. 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-robust or homoskedastic error variances. Bayesian estimation is easily accomplished by a standard MCMC procedure. I illustrate the method with an efficient-market regression of daily S&P500 returns on lagged returns. The heteroskedasticity-robust posterior shows considerably more uncertainty than does the posterior from a regression which assumes homoskedasticity.

keywords: heteroskedasticity, Bayesian regression, robust standard errors

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

Introduction

Estimation of heteroskedasticity-consistent (aka “robust”) regression 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)). I present here a Bayesian approach to heteroskedasticity-robust regression. 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.

A number of recent papers have connected heteroskedastic consistent covariance estimators to the Bayesian approach using a variety of nonparametric and Bayesian bootstrap techniques. Notably, see Lancaster (2003), Müller (2009), Szpiro, Rice, and Lumley (2010), Poirier (2011), and Norets (2012). The estimator offered below is particularly simple and easily allows the investigator to nest homoskedastic and heteroskedastic models.

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 regression is to allow heteroskedasticity to affect the spread of the posterior without changing its mean.

Fundamentally, the approach offered below allows Bayesian estimation of regression parameters without requiring a complete probability statement with regard to the error variances. On the one hand, this allows for the introduction of prior information about the regression parameters without requiring much to be said about the error variances. This may be useful. On the other hand if information is available about the error variances, Bayesian GLS might be preferred.

It turns out that the principal “trick” to heteroskedasticity-robust 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. A smaller, but useful, second piece of the puzzle is to offer a prior of convenience that can be parameterized to give a posterior for the error variances, conditional on the coefficients, to force either a homoskedastic model or to allow for heteroskedastic-robust estimation.

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 estimate an “efficient market” model where I regress the return on the S&P500 on the lagged return and ask how close the lag coefficient is to zero.

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, distribution of the ordinary least squares estimator is given by $\beta_{OLS} \sim N(\beta, (X'X)^{-1}\Omega(X'X)^{-1})$, $\Omega \equiv X'\Sigma X = \sigma^2 X'\Lambda X$. The terms on either side of Σ in the matrix 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 in both Bayesian and frequentist versions.

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.[2] A text book version of the "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, 2005) who showed that a model with Student-*t* errors can be expressed as a heteroskedastic model with the conditional posterior for the regression parameters following a Bayesian weighted least squares regression.[3]

Despite GLS' possible efficiency advantages, OLS estimates with robust standard errors are often preferred because use of a mispecified error variance-covariance matrix can lead to bias in GLS coefficient estimates. The breakthrough that permitted robust standard errors was recognition that while $\beta_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \sim N(\beta, (X'\Sigma^{-1}X)^{-1})$ 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

[2] For an application of heteroskedastic variance estimation in a nonparametric setting see Chib and Greenberg (2013).

[3] See also Albert and Chib (1993) section 3.2 for the connection between *t*- errors and a GLS conditional posterior.

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.

What makes robust estimation work is that $X'\Sigma X$ is identified even though $X'\Sigma^{-1}X$ is not. To see this, imagine a Gedanken experiment in which the investigator is handed $\epsilon_i, i = 1, \dots, n$. 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 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) $f_{I\Gamma}(z; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-(\alpha+1)} \exp[-\beta/z]$. In contrast to the squared error, the reciprocal of the squared error 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 $\widehat{\Sigma^{-1}}$ the law of large numbers does not apply to $X'\widehat{\Sigma^{-1}}X$. In contrast, the moments for $\hat{\Sigma}$ 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 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). Note that focusing on the moment conditions is not entirely free. The structural regression implies the moment conditions, but the moment conditions do not necessarily imply the structural regression. Focusing on the moment conditions is sufficient for finding the regression parameters, which after all are identified by these moment conditions, but not sufficient for identifying observation-specific error variances for which we return to equation (1).

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

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

We can write the likelihood function as

$$p(X'y|\beta, \Omega; X'X) = (2\pi)^{-\frac{k}{2}} |\Omega|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(X'y - X'X\beta)'\Omega^{-1}(X'y - X'X\beta)\right] \tag{3}$$

Conditional on Ω $(\mathrm{i.e.}\,\sigma^2$ and $\Lambda)$, 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.

$$\begin{aligned} &\beta|y, \sigma^2, \Lambda \sim N(\bar{\beta}, \bar{V}) \\ &\bar{V} = \left(V_0^{-1} + (X'X)'\Omega^{-1}(X'X)\right)^{-1} \\ &\bar{\beta} = \bar{V}\left(V_0^{-1}\beta_0 + (X'X)'\Omega^{-1}(X'y)\right) \end{aligned} \tag{4}$$

Equations (2) and (4) may appear odd on the surface, as 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 (4) as the prior precision V_0^{-1} approaches zero. Very conveniently, the posterior variance $\bar{V} \to \left((X'X)'\Omega^{-1}(X'X)\right)^{-1} = (X'X)^{-1}\Omega(X'X)^{-1}$ and the posterior mean $\bar{\beta} \to \left((X'X)'\Omega^{-1}(X'X)\right)^{-1}\left((X'X)'\Omega^{-1}(X'y)\right) =$

$(X'X)^{-1}((X'X)'\Omega^{-1})^{-1}(X'X)'\Omega^{-1}(X'y) = (X'X)^{-1}X'y$. Note that as the prior precision becomes small relative to the information from the data, the conditional posterior for β approaches the classical least squares distribution with robust standard errors.

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

$$\begin{aligned}
&\lambda_i|\epsilon_i^2 \sim I\Gamma\left(a+\frac{1}{2},\frac{\epsilon_i^2}{2\sigma^2}+a-1\right)\\
&\mathrm{E}(\lambda_i|\epsilon_i^2)=\frac{\frac{\epsilon_i^2}{2\sigma^2}+a-1}{a-\frac{1}{2}},a>1\\
&\mathrm{var}(\lambda_i|\epsilon_i^2)=\frac{\left(\frac{\epsilon_i^2}{2\sigma^2}+a-1\right)^2}{\left(a-\frac{1}{2}\right)^2(a-1.5)},a>1.5
\end{aligned} \tag{5}$$

Judicious choice of a allows equation (5) to represent either heteroskedastic or homoskedastic errors. Note that as the prior parameter $a \rightarrow 1$, then $\mathrm{E}(\lambda_i|\epsilon_i^2) \rightarrow \epsilon_i^2/\sigma^2$ so that the conditional posterior mean of λ is standardized. As $a \rightarrow \infty$, the prior and posterior both converge in probability to 1. This forces a homoskedastic model so σ^2 is identified separately from Λ, which is fixed by the prior. As $a \rightarrow \infty$, the parameters for the conditional posterior for β converge to $\bar{V} = (V_0^{-1} + (X'X)/\sigma^2)^{-1}$ and $\bar{\beta} = \bar{V}(V_0^{-1}\beta_0 + X'y/\sigma^2)$, which is the standard Bayesian conditional posterior under homoskedasticity. 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. Note in general that while conditional posteriors are given for all n elements of Λ, all that we make use of is the $k \times k$ matrix $X'\sigma^2\Lambda X$.

In summary, a diffuse prior that is robust to heteroskedasticity consists in setting V_0 large, ν_0 just above 2, and a just above 1.

Comparison with frequentist results

Some comparisons with frequentist estimators when heteroskedasticity is present may be of interest, both for purposes of intuition and to ask what are the frequentist properties of the proposed Bayesian estimator when diffuse priors are used. First, as noted earlier, the posterior for β conditional on Ω mimics the frequentist heteroskedasticity-robust distribution for the least squares estimator. Further, when $a \rightarrow \infty$ the Bayesian distribution mimics the frequentist distribution based on homoskedastic standard errors.

Bayesians generally have less concern than do frequentists with the asymptotic behavior of estimators. Nonetheless, a look at equation (4) 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 (5) 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.

While the conditional posterior of β under heteroskedasticity looks like the frequentist distribution, the marginal posterior is more spread out due to the sampling of λ. This is illustrated in a simple Monte Carlo. I generated 2,000 realizations of the model $y_i = \beta x_i + \varepsilon_i, i = 1, \ldots, n$ for $= 120, \beta = 0$, x generated from $U(0,1)$ and then normalized so $\sum x_i^2 = 1$, and $\varepsilon_i \sim N\left(0, n\sqrt{x_i^2}\right)$. Priors were $\beta_0 = 0, V_0^{-1} = 0.01$, and $\nu_0 = 3$. The Gibbs sampler was run for 1,000 burn-ins followed by 10,000 retained draws, first with $a = 1.001$ and then with $a = 1{,}000$. Results are shown in Table 1.

	least squares	Bayesian heteroskedastic $(a = 1.001)$	Bayesian homoskedastic $(a = 1{,}000)$
mean β_{ols} or mean of posterior means	0.0125	0.0121	0.0120
$\mathrm{var}(\beta_{ols})$ or variance of posterior means	1.759	1.677	1.725
mean reported OLS variance or mean posterior variance assuming homoskedasticity	1.000		0.989
mean reported OLS variance or mean posterior variance assuming heteroskedasticity	1.778	2.354	

Table 1

The first line of Table 1 shows that the posterior means are unbiased in repeated samples, just as the least squares estimator is. The second line shows that the distribution of posterior means has approximately the same variance as does the least squares estimator, remembering that essentially uninformative priors are being imposed. Line three gives the mean reported variances assuming homoskedasticity. The frequentist and Bayesian values are approximately equal. Unsurprisingly given the heteroskedastic data generating process, they are equally wrong. The last line shows, as suggested above, that the variance of the heteroskedastic marginal posterior is typically somewhat larger than the frequentist variance across repeated samples.

Illustrative Example

As an illustration, I consider an estimate of weak form stock market efficiency. The equation of interest is

$$r_t = \beta_0 + \beta_1 r_{t-1} + \epsilon_t \tag{6}$$

where r_t is the daily return on the S&P 500.[4] If the stock market is efficient, then β_1 should be very close to zero. Large deviations from zero imply the possibility of trading rules with large profits.

I choose relatively diffuse priors. Most importantly, the prior for β_1 is centered on zero with a standard deviation of 3.0. $|\beta_1| \geq 1$ implies an explosive process for the stock market return, which is not tenable. So a standard deviation of 3.0 is very diffuse. The prior for the intercept is also centered at zero, with a prior standard deviation set at three times the mean sample return. The prior for σ^2 is also diffuse, with $\sigma_0^2 = \mathrm{var}(r_t)$ and $\nu_0 = 3$. For the homoskedastic model I set $a = 1{,}000$; for the heteroskedastic model $a = 1.001$. Results for Gibbs sampling with 100,000 draws retained after a burn-in of 10,000 draws are reported in Table 2.[5]

The substantive question of interest is the size of the coefficient on lagged returns. The posterior for β_1 is shown in Figure 1, with descriptive statistics given in Table 2. The posterior for the lag coefficient is centered at approximately 0.026 for both the homoskedastic and the heteroskedastic models. As one would expect if heteroskedasticity is present, the heteroskedastic posterior is much more spread out than the homoskedastic posterior. The posterior standard deviation of the former is more than twice the posterior standard deviation of the latter. The 95 percent highest posterior density intervals are $(-0.015, 0.068)$ and $(0.010, 0.043)$, respectively. It is noteworthy the HPD estimated by the homoskedastic model

[4] The underlying price series, p_t, is SP500 from the FRED database at the Federal Reserve Bank of St. Louis. The return is defined as $r_t = 365 \times (\log p_t - \log p_{t-1})$. The sample covers 1/2/1957 through 4/22/2014 with $n = 13{,}393$ valid observations.

[5] Estimation required 1.98 milliseconds per draw on a fast vintage 2011 PC running Matlab, so computation time is not an issue.

excludes $\beta_1 = 0$ while the heteroskedastic HPD, recognizing that there is greater uncertainty, includes $\beta_1 = 0$. Substantively, both estimates suggest fairly small lag coefficients, but the estimate allowing for heteroskedasticity demonstrates considerably more uncertainty as to the true coefficient.

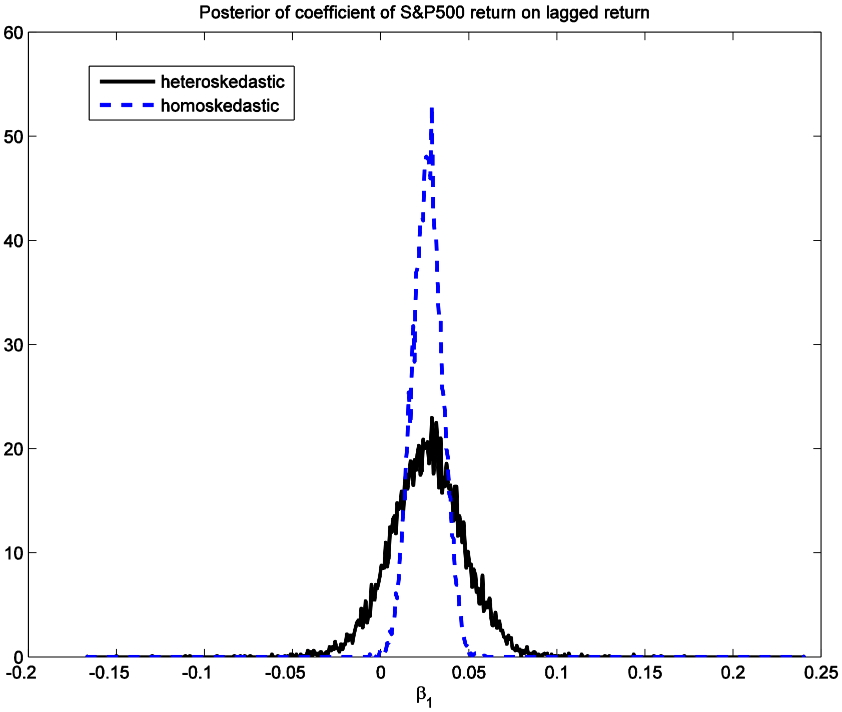

Figure 1

Prior and posteriors for S&P500 lagged return model			
		β_0	β_1
prior	mean	0	0
	standard deviation	0.255	3.00
Posterior (homoskedastic, $a = 1{,}000$)	mean	0.082	0.026
	median	0.082	0.026
	standard deviation	0.031	0.009
	95 percent HPD	$(0.021{,}0.143)$	$(0.010{,}0.043)$
Posterior (heteroskedastic, $a = 1.001$)	mean	0.081	0.027
	median	0.081	0.027
	standard deviation	0.040	0.022
	95 percent HPD	$(0.004{,}0.158)$	$(-0.015{,}0.069)$

Table 2

Conclusion

Bayesian heteroskedasticity-robust regression is straightforward. The first step is to model the sample moment conditions. This works because the required sandwich is identified even though the variances of the original errors are not. The second step is to use convenient priors 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

Albert, J.H. and S. Chib, 1993. Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association,* Vol. 88, No. 22, 669-679.

Chib, S. and E. Greenberg, 2013. On conditional variance estimation in nonparametric regression, *Statistics and Computing,* Vol. 23, Issue 2, 261-270.

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.

_____, 2005. *Contemporary Bayesian Econometrics and Statistics,* John Wiley & Sons, Hoboken, N.J.

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., A. Morse, and L. Zingales, 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.

Lancaster, T., 2003. A note on bootstraps and robustness, working paper, Brown University.

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

Müller, U. K., 2009. Risk of bayesian inference in misspecified models, and the sandwich covariance matrix, working paper, Princeton University.

Norets, A., 2012. Bayesian regression with nonparametric heteroskedasticity, working paper, Princeton University.

Poirier, D., 2011. Bayesian interpretations of heteroskedastic consistent covariance estimators using the informed bayesian bootstrap, *Econometric Reviews*, 30, 457-468.

Szpiro, A.A., K.M. Rice, and T. Lumley, 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.

Appendix – Not for Publication

This appendix provides details of the algebra connecting equations (2) and (4).

Consider a generic GLS regression of a left-hand side variable L on right-hand side variables R with error variance-covariance C, $L = R\beta + e, e \sim N(0, C)$. The textbook (e.g. Koop (2003), section 6.3) solution given a normal prior for β, $\beta \sim N(\beta_0, V_0)$, independent of C is

$$\beta | L, R, C \sim N(\bar{\beta}, \bar{V})$$

$$\bar{V} = (V_0^{-1} + R'C^{-1}R)^{-1}$$

$$\bar{\beta} = \bar{V}(V_0^{-1}\beta_0 + R'C^{-1}L)$$

In equation (2) we have $L = X'y$, $R = X'X$, and $C = \Omega$. Substituting we get

$$\bar{V} = (V_0^{-1} + (X'X)'\Omega^{-1}(X'X))^{-1}$$

$$\bar{\beta} = \bar{V}(V_0^{-1}\beta_0 + (X'X)'\Omega^{-1}X'y)$$

which is equation (4).

Alternatively, pre-multiply the original regression specification by $(X'X)^{-1}X'$ rather than by X'. Now $L = (X'X)^{-1}X'y$, $R = (X'X)^{-1}X'X = I$, and $C = (X'X)^{-1}\Omega(X'X)^{-1}$. Substituting these into equation (2) we get

$$\bar{V} = (V_0^{-1} + I'(((X'X)^{-1}\Omega(X'X)^{-1})^{-1}I)^{-1}$$

$$= \left(V_0^{-1} + (X'X)'\Omega^{-1}(X'X)\right)^{-1}$$

$$\bar{\beta} = \bar{V}(V_0^{-1}\beta_0 + I'((X'X)^{-1}\Omega(X'X)^{-1})^{-1}(X'X)^{-1}X'y)$$

$$= \bar{V}(V_0^{-1}\beta_0 + (X'X)'\Omega^{-1}(X'X)(X'X)^{-1}X'y)$$

$$= \bar{V}(V_0^{-1}\beta_0 + (X'X)'\Omega^{-1}X'y)$$

Note that this is again equation (4).