A NOTE ON BAYESIAN SCORE TESTS FOR HETEROSCEDASTICITY

Cinzia Carota

Dipartimento di Statistica e Matematica Applicata, Università di Torino
Via Maria Vittoria, 38 – 10134 Torino – Italia
cinzia.carota@unito.it

Abstract

We consider two different approaches to the heteroscedasticity problem and in both cases derive a Bayesian diagnostic tool based on the marginal score function (Carota 2005), then we compare the obtained results and the Goldfeld and Quandt test, on the one hand, and the Lagrange multiplier test, discussed by Breusch and Pagan (1979), Godfrey (1978), Cook and Weisberg (1983), on the other hand. Finally, some natural extensions of our results are described and an application to a well known real data set is provided.

Keywords: Bayesian diagnostic, Classical Core test, Goldfeld-Quandt test, Heteroscedasticity, Score function.

1. INRODUCTION

A general Bayesian diagnostic tool for model criticism has been introduced in Carota, Parmigiani and Polson (1996) where a useful linear version of such a tool is also proposed and widely discussed . In this paper, in order to test for heteroscedasticity in a multiple regression model, we will employ a "symmetric" version of such a diagnostic (Carota 2005) defined as follows

$$\bar{\Delta} = \frac{1}{2}[KL_1\{p(\mathbf{l}|\mathbf{y}) : p(\mathbf{l})\} + KL_2\{p(\mathbf{l}) : p(\mathbf{l}|\mathbf{y})\}],$$

where $\mathbf{y}$ denotes the sample data and KL_1 (KL_2) represents the Kullback-Leibler divergence between posterior and prior (prior and posterior) distributions of a random parameter $\mathbf{l}$ which is assumed to express the heteroscedasticity of the distur-

bances in the multiple regression model, as will be explained in the next section. We assume that a) there is a unique value of $\mathbf{l}$, say $\mathbf{l}_0$ or *null value,* such that the homoscedasticity assumption holds, and b) the prior distribution of $\mathbf{l}$ is sharply peaked around $\mathbf{l}_0$ (a plausible circumstance, given that the currently entertained model assumes the homoscedasticity of the disturbances). These assumptions allow us to consider a linear approximation, $\bar{\Delta}_L$, of $\bar{\Delta}$ obtained from a Taylor expansion of the integrand in $KL_i,\ i=1,2$, around $\mathbf{l}_0$,

$$\bar{\Delta}_L = \frac{1}{2}[E(\mathbf{l}|\mathbf{y}) - E(\mathbf{l})]\frac{\partial \mathscr{L}(\mathbf{l})}{\partial \mathbf{l}}|_{\mathbf{l}=\mathbf{l}_0}.$$

In this formula $\mathscr{L}$ denotes the the log-likelihood of $\mathbf{l}$, $\frac{\partial \mathscr{L}(\mathbf{l})}{\partial \mathbf{l}}|_{\mathbf{l}=\mathbf{l}_0}$ is the score function evaluated at the null value and in square brackets we have the difference between posterior and prior expectations of $\mathbf{l}$. Small values of $\bar{\Delta}$ ($\bar{\Delta}_L$) indicate closeness of prior and posterior distributions of $\mathbf{l}$ and, by virtue of assumption b), they are interpreted as sample evidence in favor of the null hypothesis of homoscedasticity of the disturbances.
In the next section we consider two different approaches to the heteroscedasticity problem and in both cases derive the Bayesian diagnostic $\bar{\Delta}_L$ based on the score function; we then compare the obtained results and the Goldfeld-Quandt test (Goldfeld and Quandt, 1965) on the one hand, and the Lagrange multiplier test (see Breusch and Pagan, 1979, Godfrey 1978, Cook and Weisberg, 1983) on the other hand. Finally, we discuss some natural extensions of previuos results and, in section 3, we provide an application to a well known real data set.

Before going into formal details it is worth noting that in the context we have described $\mathbf{l}$ represents the parameter of interest, while all other parameters, including the regression parameters, play, in this preliminary phase of model criticism, the role of "nuisance" parameteres. Therefore their priors are chosen to be "default" priors and the log-likelihood $\mathscr{L}(\mathbf{l})$ has to be interpreted as the log of the likelihood integrated with respect to the nuisance parameters.

2. DIFFERENT APPROACHES TO HETEROSCEDASTICITY

Consider a partitioned linear model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u}, \tag{1}$$

where $\mathbf{y}' = (\mathbf{y}_1' \vdots \mathbf{y}_2')$, $\mathbf{X}' = (\mathbf{X}_1' \vdots \mathbf{X}_2')$ and $\mathbf{u}' = (\mathbf{u}_1' \vdots \mathbf{u}_2')$ respectively represent two vectors of observations on the dependent variable y, whose dimensions are $(n_1 \times$

1) and $(n_2 \times 1)$, $n_1 + n_2 = n$; two matrices of predetermined values of k independent variables, both with rank k; and two vectors of uncorrelated Gaussian errors with zero means; while $\mathbf{b}$ is a common vector of random parameters. We take up the situation in which the variances in the two groups are unequal and both unknown: $\sigma_1^2 \neq \sigma_2^2$. The likelihood function is

$$L(\mathbf{b}, \sigma_1, \sigma_2 | \mathbf{X}, \mathbf{y}) \propto \frac{1}{\sigma_1^{n_1} \sigma_2^{n_2}} \times$$

$$exp[-\frac{1}{2\sigma_1^2}(\mathbf{y}_1 - \mathbf{X}_1\mathbf{b})'(\mathbf{y}_1 - \mathbf{X}_1\mathbf{b}) - \frac{1}{2\sigma_2^2}(\mathbf{y}_2 - \mathbf{X}_2\mathbf{b})'(\mathbf{y}_2 - \mathbf{X}_2\mathbf{b})],$$

and, changing variables from $(\mathbf{b}, \sigma_1^2, \sigma_2^2)$ to $(\mathbf{b}, \sigma_1^2, l = \sigma_1^2/\sigma_2^2)$ with $l \in (0, \infty)$, (see Zellner, 1971), it can be written as

$$\frac{l^{n_2/2}}{\sigma_1^{n_1+n_2}} exp\{-\frac{1}{2\sigma_1^2}[(\mathbf{y}_1 - \mathbf{X}_1\mathbf{b})'(\mathbf{y}_1 - \mathbf{X}_1\mathbf{b}) + l(\mathbf{y}_2 - \mathbf{X}_2\mathbf{b})'(\mathbf{y}_2 - \mathbf{X}_2\mathbf{b})]\}.$$

In this way the parameter of interest turns out to be l ($\mathbf{l}$ reduces to a scalar) whose *null value* is $l = l_0 = 1$, while $\mathbf{b}$ and σ_1 are the nuisance parameters. Assuming that $p(\mathbf{b}, \sigma_1, l) = p(\mathbf{b}, \sigma_1)p(l) \propto \frac{1}{\sigma_1} p(l)$ and omitting additive constants, we obtain that the log of the integrated likelihood with respect to $\mathbf{b}$ and σ_1 is

$$\mathscr{L}(l|\mathbf{X}, \mathbf{y}) \equiv log \int L(\mathbf{b}, \sigma_1, l | \mathbf{X}, \mathbf{y}) p(\mathbf{b}, \sigma_1) d\mathbf{b} d\sigma_1 \propto$$

$$\frac{n_2}{2} logl - \frac{1}{2} log|\mathbf{X}_1'\mathbf{X}_1 + l\mathbf{X}_2'\mathbf{X}_2| - \frac{n_1 + n_2 - k}{2} \times$$

$$log\{\mathbf{y}_1'\mathbf{y}_1 + l\mathbf{y}_2'\mathbf{y}_2 - (\mathbf{X}_1'\mathbf{y}_1 + l\mathbf{X}_2'\mathbf{y}_2)'(\mathbf{X}_1'\mathbf{X}_1 + l\mathbf{X}_2'\mathbf{X}_2)^{-1}(\mathbf{X}_1'\mathbf{y}_1 + l\mathbf{X}_2'\mathbf{y}_2)\},$$

whose first derivative evaluated at the null value $l_0 = 1$ provides the score function, hereafter denoted by S,

$$S = \tfrac{n_2}{2} - \tfrac{1}{2} tr\{((\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_1'\mathbf{X}_1 + I)^{-1}\} - \tfrac{n_1+n_2-k}{2} \, \tfrac{\hat{\mathbf{u}}_2'\hat{\mathbf{u}}_2}{\hat{\mathbf{u}}'\hat{\mathbf{u}}},$$

with $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$, $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\mathbf{u}}_2 = \mathbf{y}_2 - \mathbf{X}_2\hat{\mathbf{b}}$. Dividing by $\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1$ both numerator and denominator of the last term in S, we obtain that $\hat{\mathbf{u}}_2'\hat{\mathbf{u}}_2/\hat{\mathbf{u}}'\hat{\mathbf{u}} = GQ/(GQ+1)$, where $GQ = \hat{\mathbf{u}}_2'\hat{\mathbf{u}}_2/\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1$ mimics the Goldfeld-Quandt test statistic for heteroscedasticity, the difference being that here we do not fit two separate

regressions for the two sets of observations. This is not necessary because we do not need to known the distributional properties of GQ and no costs in terms of degrees of freedom are therefore involved. In order to obtain $\bar{\Delta}_L$ the score function S must be multiplied by the difference between prior and posterior expectations of the parameter l. Though not necessary, it is quite natural to assume that the prior expectation coincides with the null value, while to obtain the posterior expectation (the posterior distribution of l is $p(l|\mathbf{X},\mathbf{y}) \propto p(l)exp\{\mathscr{L}(l|\mathbf{X},\mathbf{y})\}$) numerical integration techniques are required. Note that if the regression coefficient vectors were not identical in the two data sets, as in the Goldfeld and Quandt approach, and if, for large m and r, the prior would be as follows:

$$p(l) \propto \begin{cases} l^m & if\ 0 < l \leq 1 \\ l^{-r} & if\ 1 < l < \infty \end{cases}$$

then the posterior expectation could be immediately evaluated by suitably referring to the standard F distribution.

More generally, in model (1) we can assume that $\mathbf{u}$ follows a multivariate normal distribution with mean zero and diagonal covariance matrix $\sigma_2\mathbf{W}$, whose diagonal entries are $w_1, w_2, ..., w_n$, with all $w_i > 0$ and $w_i = w(\mathbf{z}_i, \mathbf{l})$. Now $\mathbf{l}$ is a q-dimensional vector of unknown parameters and $\mathbf{z}_i = (z_{i1}, ..., z_{iq})'$ denotes a known vector, while the unknown constant σ^2 represents the part of $Var(y_i)$ common to all the observations. If w_i is a twice differentiable function of $\mathbf{l}$ and there is a unique value $\mathbf{l}_0$ of $\mathbf{l}$ such that $w_i = w(\mathbf{z}_i, \mathbf{l}_0) = 1$, then, after some algebra, the corresponding score function, S_W, can be expressed as

$$S_W = \frac{1}{2}[\mathbf{d} + \mathbf{D}'(\mathbf{v} - \mathbf{1})]$$

where $\mathbf{d}$ is a q by 1 vector with typical element $d_s = tr\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ddot{\mathbf{W}}_s\mathbf{X}\}$ and $\ddot{\mathbf{W}}_s$ denotes the diagonal matrix of the derivatives of w_i, $i = 1, .., n$, with respect to l_s; $\mathbf{D}$ is defined as follows

$$D = \begin{bmatrix} \frac{\partial w_1}{\partial l_1} & \cdots & \frac{\partial w_1}{\partial l_q} \\ \vdots & \vdots & \vdots \\ \frac{\partial w_n}{\partial l_1} & \cdots & \frac{\partial w_n}{\partial l_q} \end{bmatrix};$$

and $\mathbf{v} - \mathbf{1}$ is a q- dimensional vector with typical element $(\hat{u}_s^2/\tilde{\sigma}^2 - 1)$. Here $\tilde{\sigma}^2 = \sum \hat{u}_s^2/(n-k)$, i. e. the unbiased estimation of the common variance σ^2

instead of the maximum likelihood estimation appearing in the classical score function (see, for instance, Godfrey, 1978, Cook and Weisberg, 1983)
As to the difference between prior and posterior expectation, for a given prior distribution $p(\mathbf{l})$, the posterior is

$$p(\mathbf{l}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{l}) \prod_{1}^{n} w_i^{-1/2} |\mathbf{X}_w'\mathbf{X}_w|^{-1/2} \{\mathbf{y}_w'[I - \mathbf{X}_w(\mathbf{X}_w'\mathbf{X}_w)^{-1}\mathbf{X}_w']\mathbf{y}_w\}^{-(n-k)/2}$$

where $X_w = \mathbf{W}^{-1/2}\mathbf{X}$ and $\mathbf{y}_w = \mathbf{W}^{-1/2}\mathbf{y}$. For further progress an explicit form for w_i must be chosen. Specific families are: $w_i = (\sum l_s z_{is})^t$, with t a pre-specified integer; $w_i = exp\{\sum l_s z_{is}^{a_s}\}$, which, if by convention we take $z^a = logz$ when $a = 0$, yields important sub-families for $\mathbf{a} = \mathbf{0}$ and $\mathbf{a} = \mathbf{1}$ (see Cook and Weisberg, 1983); and $w(\mathbf{z}_i, \mathbf{l}) = w(\mathbf{x}_i'\mathbf{b}, l)$, in which $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$ and $\mathbf{l}$ reduces to l, so that the variance is constrained to depend on the expected response.

A further possibility, here not considered in details, is embedding the problem of heteroscedasticity in the framework of the mixed linear model. Since Hildreth and Houch (1968), many papers by Swamy (for a survey see Swamy, 1973) and Hsiao (1975), a number of authors have suggested that parameter heterogeneity can be reasonably viewed as due to stochastic variation. The underlying idea can be summarized as follows. The model (1) is extended by incorporating a vector of random effects, $\mathbf{e}$ and the corresponding design matrix H,

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + H\mathbf{e} + \mathbf{u}. \tag{2}$$

Then, we assume that $\mathbf{e}$ can be partitioned into a series of sub-vectors

$$\mathbf{e} = (\mathbf{e}_1' \vdots \mathbf{e}_2' \vdots ... \vdots \mathbf{e}_\mathbf{r}')',$$

with $\mathbf{e_i} \sim N(\mathbf{0}, \sigma^2 I_{q_i}) \ \ \forall\ i$ (q_i is the number of elements in $\mathbf{e}_i$, i.e., of levels of the factor corresponding to $\mathbf{e_i}$ that are represented in the data) and $cov(\mathbf{e}_i, \mathbf{e}_j') = \mathbf{0}$ $\forall\ \ i \neq j$, and that $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 I_n)$ with $cov(\mathbf{e}, \mathbf{u}') = \mathbf{0}$. Therefore, partitioning H conformably with $\mathbf{e}$ and writing (2) as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{i=1}^{r} H_i \mathbf{e}_i + \mathbf{u},$$

give $Var(\mathbf{y}) = \sum_{i=1}^{r} \sigma_i^2 H_i H_i' + \sigma^2 I_n$ and the problem of testing for heterogeneity of the parameters reduces to testing for heteroscedasticity, with $\mathbf{l} = (\sigma_1, \sigma_2, .., \sigma_r)'$ and $\mathbf{l_0} = (0, 0, .., 0)'$. A Bayesian multivariate generalization of such a model is

analyzed by Jelenkowska and Press (1996) and, in section 6, they provide approximate posterior means for the elements of all covariance matrices involved in their multivariate mixed model, which are denoted by $(\Sigma, \Sigma_1, \Sigma_2, .., \Sigma_r)'$ and correspond to the previous variance vector $(\sigma, \sigma_1, \sigma_2, .., \sigma_r)'$. Such a result allows us to suitably extend the Bayesian diagnostic $\bar{\Delta}_L$ to this more general context, bypassing the computational difficulties due to the fact that, on the one hand, the joint posterior distribution of $(\Sigma, \Sigma_1, \Sigma_2, .., \Sigma_r)'$ is analytically intractable and, on the other hand, the marginal posteriors for Σ and the Σ_i are unknown. Instead, further work is necessary to extend $\bar{\Delta}_L$ to the homogeneity testing problem in the generalized mixed linear model . Key results about the corresponding classical score test can be found in Jacqmin-Gadda and Commenges (1995) and Commenges and Jacqmin-Gadda (1997).

A second aspect which merits attention is investigating the invariance property of $\bar{\Delta}_L$ for assessing heteroscedasticity in the linear model with classical measurement error (i.e. an alternative to the standard regression model, in which it is assumed that the independent variables X are subject to error, often referred to as "measurement error model" or "error in variables model"), given that Cheng and Tsai (2004) proved the invariance of the Lagrange multiplier test derived by Breusch and Pagan (1979) and Cook and Weisberg (1983) in this different context. This is a crucial result for two reasons: the great additional complexity originating from the presence of errors in the variables and the severe consequences of ignoring them on certain inferences, such as bias and inconsistency in parameter estimation.

3. APPLICATION

We reconsider the data set analyzed by Greene (1993, chapter14). Per capita expenditure on public schools and per capita income by state (in 1979 in U.S.) are given in Table 14.1, p.385, and the residuals of the least square regression of spending on a constant, per capita income and the square per capita income are plotted against income for the 15 highest and 15 lowest values of income in Fig. 14.1, on p.386. This visual inspection suggests that the disturbances of the regression model are heteroscedastic with variances depending on income. Consequently, the observations ranked on the basis of this variable are assigned to group 1 (low variance) or 2 (high variance) and two separate regressions are carried out in order to apply a Goldfeld-Quandt test. Surprisingly, for many splitting of the sample into two groups (from a 13/13 split to a 25/25 split), such a test never leads to rejection of the null hypothesis of homoscedasticity. Vice versa here is what happens when we apply the diagnostic $\bar{\Delta}_L$ to this data set. The score

function S takes a value of -3.35 and if we adopt a Gamma prior, $Ga(a,b)$ with a=b (this ensures that the prior expectation $E(l)$ coincides with the null value of l), the corresponding difference between posterior and prior expectations is such that $\bar{\Delta}_L > 0.66$ for all values of the hyperparameter a which are less than 10.50. The value 0.66 is taken as a reference value since, according to the calibration by McCulloch (1989) based on the comparison of two Bernoulli distributions $B(\theta)$, it represents the value of the symmetrized Kulback-Leibler divergence between $B(\theta = 0.5)$ and $B(\theta = 0.95)$, that is to say between two extremely different distributions. Therefore, the previous result can be more clearly described saying that $\bar{\Delta}_L$ applied to the expenditure data strongly leads us to reject the null hypothesis of homoscedasticity unless the prior variance of l (here interpretable as a measure of the subjective degree of belief in the null hypothesis) is less than 0.09.

REFERENCES

BREUSCH T.S., PAGAN A.R. (1979), A simple test for heteroscedasticity and random coefficient variation, *Econometrica*, **47**, n. 5, 1287-1294.

CAROTA, C., PARMIGIANI, G., POLSON, N.G. (1996), Diagnostic Measures for Model Criticism, *Journal of the American Statistical Association*, **91**, 753-762.

CAROTA C. (2005), Symmetric diagnostics for the analysis of the residuals in regression models *Biometrika*, **92**, n. 4, 787-99.

CHENG C-L., TSAI C-L. (2004), The invariance of some score tests in the linear model with classical measurement error. *Journal of the American Statistical Association*, **99** , 805-09.

COMMENGES D., JACQMIN-GADDA H. (1997), Generalized score tests of homogeneity based on correlated random effects models, *Journal of the Royal Statistical Society*, **59**, 157-151.

COOK D., WEISBERG S. (1983), Diagnostics for heteroscedasticity in regression, *Biometrika*, **71**, n. 1, 1-10.

GODFREY L. (1978), Testing for multiplicative heteroskedasticity, *Journal of Econometrics*, **8**, 227-236.

GOLDFELD S. QUANDT R., (1965), Some Tests for Homoscedasticity, *Journal of the American Statistical Association*, **60**, 539-547.

GREENE W.H. (1992), *Econometric Analysis*, Macmillan Publishing Company, New York.

HILDRETH C., HOUCH C. (1968), Some Estimators for Linear Models with Random Coefficients, *Journal of the American Statistical Association*, **63**, 584-595.

HSIAO C. (1975) Some Estimation Methods for a Random Coefficient Model, *Econometrica*, **43**, 305-325.

JACQMIN-GADDA H., COMMENGES D. (1995), Tests of homogeneity for generalized linear models. *Journal of the American Statistical Association*, **90**, 1237-1246.

JELENKOWSKA T. H., PRESS S.J. (1996), Bayesian mixed linear models. In Bayesian analysis in Statistics and Econometrics, Essays in Honor of Arnold Zellner, Berry Chaloner and Geweke Eds. Wiley, New York.

MCCULLOCH R. E. (1989), Local model influence, *Journal of the American Statistical Association*, **84**, 473-78.

SWAMY P. (1973), Criteria, Constraints and Multicollinearity in Random Coefficient Regression Models, *Annals of Economics and Social Measurements*, **2**, 429-450.

ZELLNER A. (1971), *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.

UNA NOTA SUI TEST DI ETEROSCHEDASTICITÀ

Riassunto

In questo articolo si considerano due approcci distinti al problema dell'eteroschedasticità degli errori in un modello di regressione multipla e in ciascun caso si ricavano delle diagnostiche bayesiane imperniate sulla funzione score *marginale, ossia la derivata rispetto ai parametri di interesse del logaritmo della funzione di verosimiglianza integrata rispetto ai parametri di disturbo. I risultati ottenuti vengono confrontati con il test di Goldfeldt and Quandt (1965) da un lato e il test dei moltiplicatori di Lagrange (Breusch e Pagan, 1979, Godfrey, 1978, Cook e Weisberg, 1983), dall'altro. Infine vengono delineati alcuni sviluppi nell'ambito del modelli lineari con effetti aleatori e viene presentata una applicazione a dati reali già studiati in letteratura.*