Heteroscedasticity and Complex Variation

HARVEY GOLDSTEIN

Volume 2, pp. 790–795

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Heteroscedasticity and Complex Variation

Introduction

Consider the simple linear regression model with normally distributed residuals

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad e_i \sim N(0, \sigma_e^2) \qquad (1)$$

where β_0, β_1 are the intercept and slope parameters respectively, i indexes the observation, and e_i is an error term (*see* **Multiple Linear Regression**). In standard applications, such a model for a data set typically would be elaborated by adding further continuous or categorical explanatory variables and interactions until a suitable model describing the observed data is found (*see* **Model Selection**). A common diagnostic procedure is to study whether the constant residual variance (homoscedasticity) assumption in (1) is satisfied. If not, a variety of actions have been suggested in the literature, most of them concerned with finding a suitable nonlinear transformation of the response variable so that the homoscedasticity assumption is more closely approximated (*see* **Transformation**). In some cases, however, this may not be possible, and it will also in general change the nature of any regression relationship. An alternative is to attempt to model the heteroscedasticity explicitly, as a function of explanatory variables. For example, for many kinds of behavioral and social variables males have a larger variance than females, and rather than attempting to find a transformation to equalize these variances, which would in this case be rather difficult, we could fit a model that had separate variance parameters for each gender. This would have the advantage not only of a better fitting model, but also of providing information about variance differences that is potentially of interest in its own right.

This article discusses general procedures for modeling the variance as a function of explanatory variables. It shows how efficient estimates can be obtained and indicates how to extend the case of linear models such as (1) to handle multilevel data (*see* **Linear Multilevel Models**) [2]. We will first describe, through a data example using a simple linear model, a model fitting separate gender variances and then discuss general procedures.

An Example Data Set of Examination Scores

The data have been selected from a very much larger data set of examination results from six inner London Education Authorities (school boards). A key aim of the original analysis was to establish whether some schools were more 'effective' than others in promoting students' learning and development, taking account of variations in the characteristics of students when they started Secondary school. For a full account of that analysis, see Goldstein et al. [5].

The variables we shall be using are an approximately normally distributed examination score for 16-year-olds as the response variable, with a standardized reading test score for the same students at age 11 and gender as the explanatory variables.

The means and variances for boys and girls are given in Table 1.

We observe, as expected, that the variance for girls is lower than for the boys.

We first fit a simple model which has a separate mean for boys and girls and which we write as

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad e_i \sim N(0, \sigma_e^2)$$
$$x_{1i} = 1 \text{ if a boy, } 0 \text{ if a girl, } x_{2i} = 1 - x_{1i} \qquad (2)$$

There is no intercept in this model since we have a dummy variable for both boys and girls. Note that these data in fact have a two-level structure with significant variation between schools. Nevertheless, for illustrative purposes here we ignore that, but see Browne et al. [1] for a full multilevel analysis of this data set.

If we fit this model to the data using ordinary least squares (OLS) regression (*see* **Least Squares Estimation**; **Multiple Linear Regression**), we obtain the estimates in Table 2.

Note that the fixed coefficient estimates are the same as the means in Table 1, so that in this simple case the estimates of the means do not depend on the homoscedasticity assumption. We refer to the explanatory variable coefficients as 'fixed' since they

Table 1 Exam scores by gender

	Boy	Girl	Total
N	1623	2436	4059
Mean	−0.140	0.093	−0.000114
Variance	1.051	0.940	0.99

Table 2 OLS estimates from separate gender means model (2)

	Coefficient	Standard error
Fixed		
Boy (β_1)	−0.140	0.024
Girl (β_2)	0.093	0.032
Random		
Residual variance (σ_e^2)	0.99	0.023
−2 log-likelihood	11455.7	

have a fixed underlying population value, and the residual variance is under the heading 'random' since it is associated with the random part of the model (residual term).

Modeling Separate Variances

Now let us extend (2) to incorporate separate variances for boys and girls. We write

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + e_{1i} x_{1i} + e_{2i} x_{2i}$$
$$e_{1i} \sim N(0, \sigma_{e1}^2), \; e_{2i} \sim N(0, \sigma_{e2}^2)$$
$$x_{1i} = 1 \text{ if a boy, } 0 \text{ if a girl, } x_{2i} = 1 - x_{1i} \quad (3)$$

so that we have separate residuals, with their own variances for boys and girls. Fitting this model, using the software package MLwiN [6], we obtain the results in Table 3.

We obtain, of course, the same values as in Table 1 since this model is just fitting a separate mean and variance for each gender. (Strictly speaking they will not be exactly identical because we have used **maximum likelihood estimation** for our model estimates, whereas Table 1 uses unbiased estimates for the variances; if restricted maximum likelihood (REML) model estimates are used, then they will be identical (*see* **Maximum Likelihood Estimation**)). Note that the difference in the −2 log-likelihood values is 6.2, which judged against a chi squared distribution on 1 degree of freedom (because we are adding just 1 parameter to the model) is significant at approximately the 1% level.

Table 3 Estimates from separate gender means model (3)

	Coefficient	Standard error
Fixed		
Boy (β_1)	−0.140	0.025
Girl (β_2)	0.093	0.020
Random		
Residual variance Boys (σ_{e1}^2)	1.051	0.037
Residual variance Girls (σ_{e2}^2)	0.940	0.027
−2 log-likelihood	11449.5	

Now let us rewrite (3) in a form that will allow us to generalize to more complex variance functions.

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$
$$e_i = e_{0i} + e_{1i} x_{1i}$$
$$\text{var}(e_i) = \sigma_{e0}^2 + 2\sigma_{e01} x_{1i} + \sigma_{e1}^2 x_{1i}^2, \quad \sigma_{e1}^2 \equiv 0$$
$$x_{1i} = 1 \text{ if a boy, } 0 \text{ if a girl} \quad (4)$$

Model (4) is equivalent to (3) with

$$\beta_2^* \equiv \beta_0, \quad \beta_1^* \equiv \beta_0 + \beta_1$$
$$\sigma_{e2}^{2*} \equiv \sigma_{e0}^2, \quad \sigma_{e1}^{2*} \equiv \sigma_{e0}^2 + 2\sigma_{e01} \quad (5)$$

where the * superscript refers to the parameters in (3).

In (4), for convenience, we have used a standard notation for variances and the term σ_{e01} is written as if it were a covariance term. We have written the residual variance in (4) as $\text{var}(e_i) = \sigma_{e0}^2 + 2\sigma_{e01} x_{1i} + \sigma_{e1}^2 x_{1i}^2$, $\sigma_{e1}^2 \equiv 0$, which implies a covariance matrix with one of the variances equal to zero but a nonzero covariance. Such a formulation is not useful and the variance in (4) should be thought of simply as a reparameterization of the residual variance as a function of gender. The notation in (4) in fact derives from that used in the general multilevel case [2], and in the next section we shall move to a more straightforward notation that avoids any possible confusion with covariance matrices.

Modeling the Variance in General

Suppose now that instead of gender the explanatory variable in (4) is continuous, for example, the reading test score in our data set, which we will now denote by x_{3i}. We can now write a slightly extended form of (4) as

$$y_i = \beta_0 + \beta_3 x_{3i} + e_i$$
$$e_i = e_{0i} + e_{3i} x_{3i}$$
$$\text{var}(e_i) = \sigma_{e0}^2 + 2\sigma_{e03} x_{3i} + \sigma_{e3}^2 x_{3i}^2 \quad (6)$$

Table 4 Estimates from fitting reading score as an explanatory variable with a quadratic variance function

	Coefficient	Standard error
Fixed		
Intercept (β_0)	−0.002	
Reading (β_3)	0.596	0.013
Random		
Intercept variance (σ^2_{e0})	0.638	0.017
Covariance (σ_{e03})	0.002	0.007
Reading variance (σ^2_{e3})	0.010	0.011
−2 log-likelihood	9759.6	

This time we can allow the variance to be a quadratic function of the reading score; in the case of gender, since there are really only two parameters (variances) one of the parameters in the variance function (σ^2_{e1}) was redundant. If we fit (6), we obtain the results in Table 4.

The deviance (−2 log-likelihood) for a model that assumes a simple residual variance is 9760.5, so that there is no evidence here that complex variation exists in terms of the reading score. This is also indicated by the standard errors for the random parameters, although care should be taken in interpreting these (and more elaborate Wald tests) using Normal theory since the distribution of variance estimates will often be far from Normal.

Model (6) can be extended by introducing several explanatory variables with 'random coefficients' e_{hi}. Thus, we could have a model where the variance is a function of gender (with x_{2i} as the dummy variable for a girl) and reading score, that is,

$$y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$
$$\text{var}(e_i) = \sigma^2_{e_i} = \alpha_0 + \alpha_2 x_{2i} + \alpha_3 x_{3i} \qquad (7)$$

We have changed the notation here so that the residual variance is modeled simply as a linear function of explanatory variables (Table 5).

The addition of the gender term in the variance is associated only with a small reduction in deviance (1.6 with 1 degree of freedom), so that including the reading score as an explanatory variable in the model appears to remove the heterogeneous variation associated with gender. Before we come to such a conclusion, however, we look at a more elaborate model where we allow for the variance to depend on the interaction between gender and the reading score,

Table 5 Estimates from fitting reading score and gender (girl = 1) as explanatory variables with linear variance function

	Coefficient	Standard error
Fixed		
Intercept (β_0)	−0.103	
Girl (β_2)	0.170	0.026
Reading (β_3)	0.590	0.013
Random		
Intercept (α_0)	0.665	0.023
Girl (α_2)	−0.038	0.030
Reading (α_3)	0.006	0.014
−2 log-likelihood	9715.3	

that is,

$$y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$
$$\text{var}(e_i) = \sigma^2_{e_i} = \alpha_0 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{2i} x_{3i} \qquad (8)$$

Table 6 shows that the fixed effects are effectively unchanged after fitting the interaction term, but that the latter is significant with a reduction in deviance of 6.2 with 1 degree of freedom. The variance function for boys is given by $0.661 - 0.040x_3$ and for girls by $0.627 + 0.032x_3$. In other words, the residual variance decreases with an increasing reading score for boys but increases for girls, and is the same for boys and girls at a reading score of about 0.5 standardized units. Thus, the original finding that boys have more variability than girls needs to be modified: initially low achieving boys (in terms of reading) have higher variance, but the girls have higher variance if they are initially high achievers. It is interesting to note that if we fit an interaction term between reading and gender in the fixed part of the model, we obtain a very small and nonsignificant coefficient whose inclusion does not affect the estimates for the remaining parameters. This term therefore, is omitted from Table 6.

One potential difficulty with linear models for the variance is that they have no constraint that requires them to be positive, and in some data sets the function may become negative within the range of the data or provide negative variance predictions that are unreasonable outside the range. An alternative formulation that avoids this difficulty is to formulate a nonlinear model, for example, for the logarithm of the variance having the general

Table 6 Estimates from fitting reading score and gender (girl = 1) as explanatory variables with linear variance function including interaction

	Coefficient	Standard error
Fixed		
Intercept (β_0)	−0.103	
Girl (β_2)	0.170	0.026
Reading (β_3)	0.590	0.013
Random		
Intercept (α_0)	0.661	0.023
Girl (α_2)	−0.034	0.030
Reading (α_3)	−0.040	0.022
Interaction (α_4)	0.072	0.028
−2 *log-likelihood*	9709.1	

form

$$\log[\mathrm{var}(e_i)] = \sum_h \alpha_h x_{hi}, \quad x_{0i} \equiv 1 \tag{9}$$

We shall look at estimation algorithms suitable for either the linear or nonlinear formulations below.

Covariance Modeling and Multilevel Structures

Consider the repeated measures model where the response is, for example, a growth measure at successive occasions on a sample of individuals as a polynomial function of time (t)

$$y_{ij} = \sum_{h=0}^{p} \beta_h t_{ij}^h + e_{ij}$$
$$\mathrm{cov}(\mathbf{e}_j) = \Omega_e \quad \mathbf{e}_j = \{e_{ij}\} \tag{10}$$

where $\mathbf{e}_j$ is the vector of residuals for the jth individual and i indexes the occasion. The residual covariance matrix between measurements at different occasions (Ω_e) is nondiagonal since the same individuals are measured at each occasion and typically there would be a relatively large between-individual variation. The covariance between the residuals, however, might be expected to vary as a function of their distances apart so that a simple model might be as follows

$$\mathrm{cov}(e_{tj}, e_{t-s,j}) = \sigma_e^2 \exp(-\alpha s) \tag{11}$$

which resolves to a first-order autoregressive structure (*see* **Time Series Analysis**) where the time intervals are equal.

The standard formulation for a repeated measures model is as a two-level structure where individual random effects are included to account for the covariance structure with correlated residuals. A simple such model with a random intercept u_{0j} and random 'slope' u_{1j} can be written as follows

$$y_{ij} = \sum_{h=0}^{p} \beta_h t_{ij}^h + u_{0j} + u_{1j} t_{ij} + e_{ij}$$
$$\mathrm{cov}(\mathbf{e}_j) = \sigma_e^2 I, \ \mathrm{cov}(\mathbf{u}_j) = \Omega_u, \ \mathbf{u}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \tag{12}$$

This model incorporates the standard assumption that the covariance matrix of the level 1 residuals is diagonal, but we can allow it to have a more complex structure as in (11). In general, we can fit complex variance and covariance structures to the level 1 residual terms in any multilevel model. Furthermore, we can fit such structures at any level of a data hierarchy. A general discussion can be found in Goldstein [2, Chapter 3] and an application modeling the level 2 variance in a multilevel generalized linear model (*see* **Generalized Linear Mixed Models**) is given by Goldstein and Noden [4]; in the case of generalized linear models, the level 1 variance is heterogeneous by virtue of its dependence on the linear part of the model through the (nonlinear) link function.

Estimation

For normally distributed variables, the likelihood equations can be solved, iteratively, in a variety of ways. Goldstein et al. [3] describe an iterative generalized least squares procedure (*see* **Least Squares Estimation**) that will handle either linear models such as (7) or nonlinear ones such as (9) for both variances and covariances. Bayesian estimation can be carried out readily using Monte Carlo Markov Chain (MCMC) methods (*see* **Markov Chain Monte Carlo and Bayesian Statistics**), and a detailed comparison of likelihood and Bayesian estimation for models with complex variance structures is given in Browne et al. [1]. These authors also compare the fitting of linear and loglinear models for the variance.

Conclusions

This article has shown how to specify and fit a model that expresses the residual variance in a linear

model as a function of explanatory variables. These variables may or may not also enter the fixed, regression part of the model. It indicates how this can be extended to the case of multilevel models and to the general modeling of a covariance matrix. The example chosen shows how such models can uncover differences between groups and according to the values of a continuous variable. The finding that an interaction exists in the model for the variance underlines the need to apply considerations of model adequacy and fit for the variance modeling. The relationships exposed by modeling the variance will often be of interest in their own right, as well as better specifying the model under consideration.

References

[1] Browne, W., Draper, D., Goldstein, H. & Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level 1 variation, *Computational Statistics and Data Analysis* **39**, 203–225.

[2] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Edition, Edward Arnold, London.

[3] Goldstein, H., Healy, M.J.R. & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data, *Statistics in Medicine* **13**, 1643–1655.

[4] Goldstein, H. & Noden, P. (2003). Modelling social segregation, *Oxford Review of Education* **29**, 225–237.

[5] Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. & Thomas, S. (1993). A multilevel analysis of school examination results, *Oxford Review of Education* **19**, 425–433.

[6] Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langfprd, I. & Lewis, T. (2000). *A User's Guide to MlwiN*, 2nd Edition, Institute of Education, London.

(*See also* **Cross-classified and Multiple Membership Models**)

HARVEY GOLDSTEIN