From ordinary to generalized least squares: a worked example

Ettore Marubini[1], Annalisa Orenti[2]

[1] Prof. Emerito di Statistica Medica - Università di Milano
[2] Sezione di Statistica medica, Biometria e Bionformatica-Fondazione IRCCS-INT, Milano

Corresponding Author:
Annalisa Orenti
Unità di Statistica Medica e Biometria
Fondazione IRCCS Istituto Nazionale dei Tumori, Via Venezian, 1 - 20133 Milano
e-mail: annalisa.orenti@istitutotumori.mi.it

1. Generation of data

In this example the relationship between the dependent variable (response) (Y) and the independent variable (regressor) (X) is postulated to be: $Y = X$.
It is assumed that the regressor X has six values (i=1,2,3,4,5,6) and, for each X value, there are four replicates of the response Y (j=1,2,3,4); furthermore the error variance (σ_i^2) increases together with the X values. By using small letters to specify the empirical realizations of random variables, the model used to generate the data results to be:

$$y_{ij} = x_i + \sigma_i \varepsilon_{ij} \qquad [1]$$

where the random errors $\varepsilon_{ij} \sim N(0,1)$ were obtained by means of the *rnorm* function of software R.
The computer clock was used to specify the seed.
It is easy to see that

$$E(Y_{ij}) = \mu_i = X_i \qquad [2]$$

Table 1 specifies both the values of X_i and $\sigma_i = \frac{1}{2} X_i$ used to generate the set of 24 data to be processed in this exercise.

The range of σ_i values is in accord with the rule of thumb suggested by Carroll and Ruppert ((1), p.16): if the standard deviations s_i (estimates of σ_i) differ by a factor of 3:1 or more, then weighting will generally be called for; whereas if the s_i do not vary by a factor of 1.5:1, then weighting would not be necessary. Values of ε_{ij} obtained by the *rnorm* function, multiplied by σ_i, and the corresponding y_{ij} obtained according to [1] are reported in table 2, third and fourth column respectively; the fifth column gives the means ($\bar{y}_i$) for every X_i together with the corresponding sample variances s_i^2 (sixth column). Values of σ_i^2 are reported in the seventh column.
Comparing columns 5 with 2, the mean estimates appear to differ slightly from the corresponding μ_i, except for the fifth subset, in which $\bar{y}_i$ tends to underestimate μ_i. On the contrary, comparing columns 6 with 7 it appears that σ_i^2 are poorly estimated by s_i^2, which are all based on 3 degrees of freedom (d.f.) only. As a matter of fact Carroll and Cline (2) assert that the number of replications for each group (i=1,2,…,I) should be at least 10, to get reasonable estimates of σ_i^2; however, in many biological analyses, such a number of replications could be prevented by practical (economical) reasons.

Table 1. Values of X_i and σ_i used to generate the set of 24 data.

i	1	2	3	4	5	6
X_i	2	3	4	5	6	7
σ_i	1	1.5	2	2.5	3	3.5

Table 2. Simulated data with some pertinent statistics.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	$X_i = \mu_i$	$\sigma_i \varepsilon_{ij}$	y_{ij}	$\bar{y}_i$	s_i^2	σ_i^2
1	2	0.6902	2.6902	1.8182	0.9229	1
2		0.5748	2.5748			
3		-0.7617	1.2383			
4		-1.2304	0.7696			
5	3	-0.9403	2.0597	3.1321	0.8455	2.25
6		0.5550	3.5550			
7		-0.2491	2.7509			
8		1.1628	4.1628			
9	4	-2.5433	1.4567	3.6793	8.0203	4
10		2.4690	6.4690			
11		-2.9684	1.0316			
12		1.7598	5.7598			
13	5	-1.8841	3.1159	5.2536	5.8360	6.25
14		3.6322	8.6322			
15		0.2468	5.2468			
16		-0.9807	4.0193			
17	6	-0.9776	5.0224	5.0261	6.1422	9
18		-0.0568	5.9432			
19		1.4993	7.4993			
20		-4.3604	1.6396			
21	7	-4.1160	2.8840	6.7267	24.6650	12.25
22		-4.6688	2.3312			
23		5.5740	12.5740			
24		2.1178	9.1178			

2. Regression Models and Results

2.1 Ordinary Least Squares (OLS)

All the above information concerning the genesis of data is not available to the analyst interested in estimating the simple linear relationship between Y and X. The first step of the analysis consists in drawing a scatter plot of the data as reported in figure 1, panel (a). The latter points out a linear relationship between Y and X and, in the meantime, raises doubts about the assumption of homoschedasticity. Nevertheless it seems sensible to start the analysis under this assumption in order to compute the diagnostics suitable to investigate whether the assumption is tenable or not. Thus the following model is fitted:

$$y_{ij} = \beta_0 + \beta_1 x_i + e_{ij} = \beta_0 + \beta_1 x_i + \sigma\varepsilon_{ij} \quad [3]$$

where σ is a scale factor common to every i.

The estimates of β_0 and β_1 are obtained by minimizing (with respect to β_0 and β_1) the residual sum of squares:

$$RSS = \sum_{i=1}^{I}\sum_{j=1}^{m_i} e_{ij}^2 = \sum_{i=1}^{I}\sum_{j=1}^{m_i} (y_{ij} - \beta_0 - \beta_1 x_i)^2 \quad [4]$$

The estimate of σ^2 is the residual mean square (RMS):

$$\frac{RSS}{\text{residual d.f.}}$$

The estimated coefficients of the model are: $\beta_0 = 0.1842$, $\beta_1 = 0.9085$.

They were obtained by means of the function *lm* of software R.

Note that $0.1842 \rightarrow \beta_0 = 0$ and $0.9085 \rightarrow \beta_1 = 1$, where $\rightarrow$ means "is estimate of".

The RMS = 6.4742, based on 22 d. f. is an estimate of $\bar{\sigma}^2 = \sum_{i=1}^{6} \sigma_i^2 / 6 = 5.79$.

The residuals e_{ij} are estimated by $\hat{e}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_i$

To examine the appropriateness of model [3], the analyst would like to have the true residuals e_{ij} available for study. But, since he only actually gets the estimate $\hat{e}_{ij}$, he can study only whether or not the fitted model $\hat{\beta}_0 + \hat{\beta}_1 x_i$ matches the data.

It is known that the residuals $\hat{e}_{ij}$ have different standard errors; to make them comparable each $\hat{e}_{ij}$ must be divided by its own standard error obtaining the so called "standardized" residuals:

$$\hat{e}'_{ij} = \frac{\hat{e}_{ij}}{\sqrt{RMS(1-h_{ii})}},$$

where h_{ij} is the pertinent term on the diagonal of the Hat matrix (see Sen and Srivastava, (3), p.107). In the case of simple linear regression:

$$h_{ii} = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{I}\sum_{j=1}^{m_i}(x_{ij} - \bar{x})^2} \qquad [5]$$

where $N = \sum_{i=1}^{I} m_i$.

With regard to our example the $\hat{e}'_{ij}$ are reported, against the corresponding predicted values $\hat{y}_{ij}$, in panel (b) of Figure 1. The "fan shape" pattern confirms that the variability within groups increases with mean response, so the constant variance assumption in [3] is inappropriate.

As previously noted, when the sample size is small s_i^2 are poor estimates of σ_i^2. As an alternative one could compute the *average squared error (ase)* for each group (i=1,2,…,I):

$$\sum_{j=1}^{m_i}\frac{(y_{ij} - \hat{y}_{ij})^2}{m_i} = \sum_{j=1}^{m_i}\frac{\hat{e}_{ij}^2}{m_i} .$$

It is expected that the average squared error is preferable to s_i^2 as a measure of within group variability, because the latter is equivalent to using $\hat{y}_{ij} = \bar{y}_i$, not taking advantage of the postulated relationship between Y and X. In the present example the average square errors are respectively $ase_1 = 0.7257$; $ase_2 = 0.6835$; $ase_3 = 6.0345$; $ase_4 = 4.6543$; $ase_5 = 4.9780$; $ase_6 = 18.5321$.

2.2 Weighted Least Squares (WLS)

Now the model is:

$$y_{ij} = \beta_0 + \beta_1 x_i + \sigma_i \varepsilon_{ij} \qquad [6]$$

where σ_i is a scale factor specific to each i.

In equation [3] all the points to be fitted receive the same weight=1/N, as a result of the homoschedasticity assumption. On the contrary in the presence of heteroschedasticity, each point must be weighted so that points for which σ_i^2 is comparatively large should be downweighted: in general the weights (w_i) should be the reciprocal of the variance.

Let's rewrite model [6] multiplying all its terms by $\frac{1}{\sigma_i} = (w_i)^{\frac{1}{2}}$.

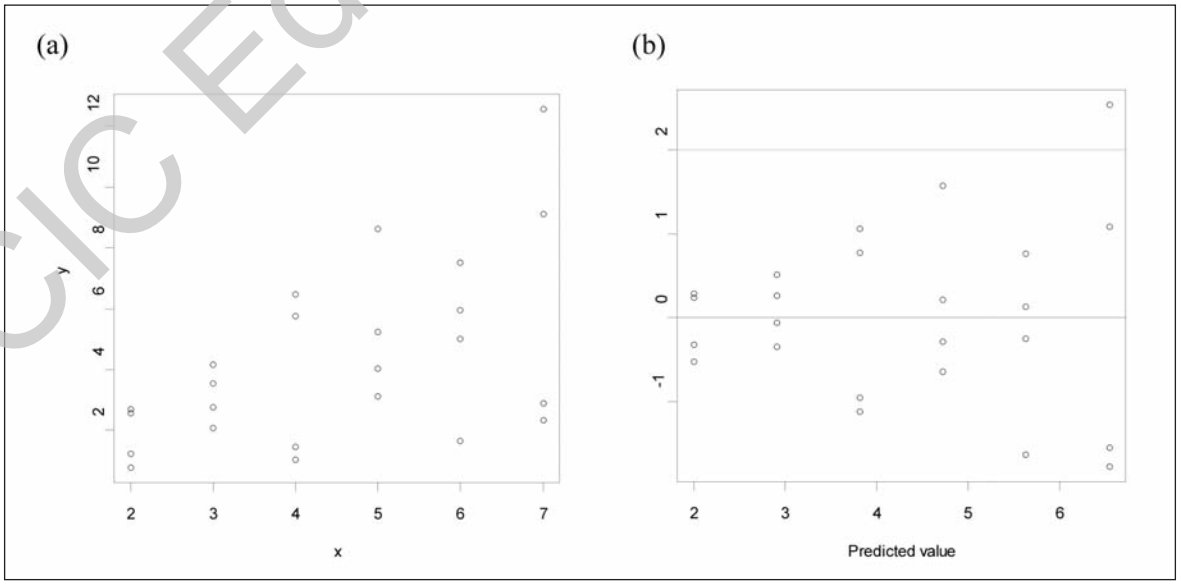

Figure 1. Scatter plot of y_{ij} against x_i (a); Scatter plot of standardized residuals $\hat{e}_{ij}$ against predicted values $\hat{y}_{ij}$ (b).

It becomes

[7]

The sum of squares to be minimized is now:

[8]

From the previous section it appears that three alternatives to compute the weights are available:

in practice this approach can rarely be used because the true value of the variance of each group is usually unknown.

this approach appears to be "naïve" when there are few replicates (<10) for each group, like in the present example.

this approach appears to be "naïve" when there are few replicates (<10) for each group, like in the present example.

All these weights are reported in table 3. Notice that to construct weights 2, s_i^2 are multiplied by ¾, to render the direct comparison of them with weights 3 possible.

The results obtained by weighted least squares regression analysis (function *lm* of software R with option weights) are reported in rows 2, 3 and 4 of table 4; for completeness, in the first row the results of OLS regression are given.

As expected it appears that the estimated standard errors of the coefficients obtained with WLS method are smaller than the corresponding values obtained with OLS method. As the weights are inversely proportional to the variances of each group, the $\bar{\sigma}^2 = 5.79$ is expected to reduce to $\bar{\sigma}^2_{weighted} = 1$; the latter is estimated by WLS_1, WLS_2, WLS_3 as: 0.9704, 1.1403, 1.0902 respectively.

Let's consider WLS_1 model: RMS = 0.9704 is the constant variance for the transformed variables $Y \cdot \sqrt{w_i}$; thus to obtain the estimated variances of each group in the original scale one must divide RMS by the corresponding weights.

Table 3. Different kinds of weights (see text).

(1) i	(2) Weights1	(3) Weights2	(4) Weights3
1	1.0000	1.4447	1.3780
2	0.4444	1.5769	1.4630
3	0.2500	0.1662	0.1657
4	0.1600	0.2285	0.2149
5	0.1111	0.2171	0.2009
6	0.0816	0.0541	0.0540

Table 4. Results of OLS and WLS regression analysis.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
					Residual	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$	Sum of Squares	Mean Square
OLS	0.1842	0.9085	1.4638	0.3041	142.4328	6.4742
WLS1	0.0047	0.9525	0.8258	0.2412	21.3478	0.9704
WLS2	0.1308	0.9317	0.7493	0.2310	25.0857	1.1403
WLS3	0.1232	0.9336	0.7518	0.2318	23.9843	1.0902

2.3 Modelling the variance: Generalized Least Squares (GLS)

The basic idea is to model the variance as a function of the mean and possibly of other factors specific to the process generating the data. Among the several models suggested by Davidian and Giltinan ((4), p.23) the simplest one called "Power Of Mean" POM will be considered here.

The POM model implies:

[9]

where γ is a scale parameter and γ is an unknown parameter to be estimated.

According to Davidian and Giltinan ((4), p.23): "…the scale parameter γ governs the overall level of precision in the response, while the variance parameter θ specifies fully the functional form."

The regression model is now:

[10]

For $\theta = 0$ model [10] reduces to model [3].

For $\theta = \frac{1}{2}$ and $\gamma = 1$, $\sigma_i^2 = Var(Y_{ij}) = \mu_i$ so Y_{ij} is distributed according to a Poisson distribution.

For $\theta = 1$ $\sigma_i = \gamma\mu_i$; this implies that $\gamma = \frac{\sigma_i}{\mu_i}$ and consequently γ corresponds to the coefficient of variation.

In order to investigate the role of θ in determinating γ, the following exercise appears to be useful. Assume that θ is known and takes values: $\theta = 0.5$ and $\theta = 1$ respectively. Furthermore we consider two alternatives regarding μ_i; a priori known $\mu_i = X_i$ (as results from [2]) and the estimate of μ_i : $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The weights become now: $w_i = \frac{1}{\mu_i^{2\theta}}$. The results pertinent to these four scenarios (GLS_1-GLS_4) are reported in table 5.

From table 5 one can observe that:

1) Comparing the results in row 2 with those in 1 and in row 4 with those in 3 it appears that they are very similar; this was expected since the OLS estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ (used to compute μ_i) are known to be unbiased.
2) Comparing row 2 of table 4 with row 3 of table 5 the estimated coefficients as well as their standard errors are exactly the same; the ratio of the two RMS is 4:1, since the weights used in table 4 (row 2) were a quarter of those used in table 5 (row 3).
3) The values of mean square tend to decrease as long as the values of θ tend to increase.

In practice θ is unknown and must be estimated from the data. First of all, a graphical display should be drawn for evaluating the appropriateness of the variance model. Briefly, by taking logarithms, equation [9] can be rewritten: $\log(\sigma_i) = \log(\gamma) + \theta\log(\mu_i)$.

Since σ_i and μ_i are unknown one needs a substitute for each and can regard $\log(|e_{ij}|)$ as a substitute for $\log(\sigma_i)$ and $\log(\hat{y}_{ij})$ as a substitute for $\log(\mu_i)$. Thus, by regressing $\log(|e_{ij}|)$ against $\log(\hat{y}_{ij})$ one can see if a strong straight line relationship is indicated. As regards our example the plot of $\log(|e_{ij}|)$ against $\log(\hat{y}_{ij})$ is given in figure 2 with the relative regression line.

One could think of using the slope and the intercept of this straight line as estimates of θ and $\log(\gamma)$ respectively. However, Davidian and Haaland (5) "…recommend against using the relationship demonstrated by these plots to estimate the parameters θ and γ of the variance function;…". As an alternative they suggest to resort to an iterative method having better statistical properties. It is based on the following steps:

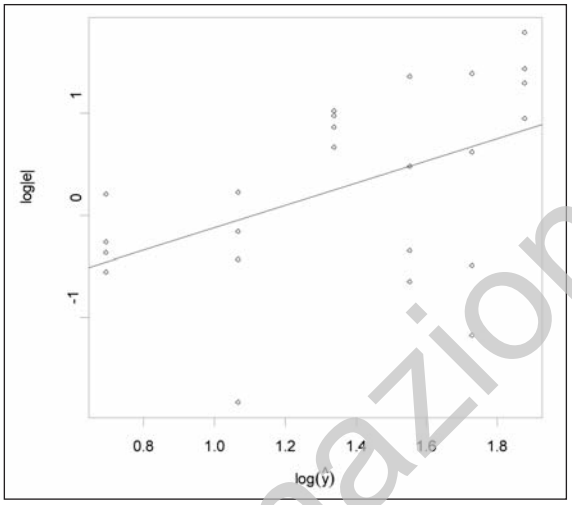

Figure 2. Scatter plot of $\log(|e_{ij}|)$ against $\log(\hat{y}_{ij})$ with the corresponding regression line.

1) Obtain the OLS estimate $\hat{\beta}_{OLS}$. Let $\hat{\beta}^{(0)} = \hat{\beta}_{OLS}$ and set k=0.
2) Obtain the estimate of $\theta^{(k)}$ as shown in section 4.1 reported by Davidian and Haaland (5).
   Form estimated weights $\hat{w}_i = \frac{1}{\hat{\mu}_i^{2\hat{\theta}^{(k)}}}$
   where $\hat{\mu}_i = \beta_0^{(k)} + \beta_0^{(k)} x_1$
3) Use the estimated weights from 2) to obtain $\hat{\beta}_{GLS}$ by minimizing [8].
4) Set k=k+1, let $\hat{\beta}^{(k)} = \hat{\beta}_{GLS}$ and return to 2)

The method is called Generalized Least Squares because the weights are estimated.

The package *calib* of software R takes advantage of this algorithm to estimate parameters for both linear and non

Table 5. Results of GLS regression analysis

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
					Residual					
	$\hat{\beta}_0$	$\hat{\beta}_1$	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$	Sum of Squares	Mean Square	θ	$\hat{\theta}$	μ_i	$\hat{\mu}_i$
GLS_1	0.0980	0.9277	1.0674	0.2593	26.0215	1.1828	0.5		X	
GLS_2	0.1020	0.9268	1.0797	0.2606	27.1400	1.2336	0.5			X
GLS_3	0.0047	0.9525	0.8258	0.2412	5.3370	0.2426	1		X	
GLS_4	0.0163	0.9494	0.8481	0.2425	5.8837	0.2674	1			X
POM	-0.0434	0.9671	0.7533	0.2416	2.9632	0.1347		1.2385		X

linear (logistic) regression models, with heteroscedastic variances modelled in terms of POM.
In *calib* γ is estimated as

[11]

where p is the number of estimated parameters.
In a strict sense $\hat{\gamma}$ can be interpreted like a coefficient of variation only when $\hat{\theta}=1$ in equation [11]; however loosely it is thought like a coefficient of variation even when $\hat{\theta}\neq 1$.
The results of fitting [10] through version 0.1.02 of package calib are reported in the fifth row of table 5.
The estimates $\hat{\theta}=1.2385$ and $\hat{\gamma}=\sqrt{0.1347}=0.367$ enable us to compute $\hat{\sigma}_{i(POM)}$, which are reported in the last row of table 6. These appear to be very good estimates of the pertinent σ_i.

3. Final comments

1) The standardized residuals (not given here) of WLS_1, WLS_2 and WLS_3 models show that the heteroscedasticity condition in the raw data has been satisfactory removed. Therefore it appears that weighting enables fulfilling the basic assumptions to compute confidence intervals of β_0 and β_1 and to test null hypothesis on β_0 and β_1.
2) Let's consider the column "mean square" in tables 4 and 5.
   For the data in this example, $\sigma_i = \frac{1}{2}x_i = \frac{1}{2}\mu_i$; for POM model $\sigma_i = \gamma\mu_i^{\theta}$. In computing the weights for WLS_1 it was assumed that both γ and θ were known, namely: $\gamma = \frac{1}{2}$ and $\theta = 1$. So the weights were $w_i = \frac{4}{\mu_i^2}$. Therefore the RMS obtained with this WLS_1 model was an estimate of $Var(\varepsilon_{ij}) = 1$. On the other hand if one assumes that only θ is known and is equal to 1, the weights are: $w_i = \frac{1}{\mu_i^2}$ and the RMS of GLS_3 is estimate of $\gamma^2 \cdot Var(\varepsilon_{ij})$, but being the latter equal 1, this RMS reduces to be an estimate of $\gamma^2 = \frac{1}{4} = 0{,}25$. In the present example $RMS_{WLS1} = 0.9704$, whereas $RMS_{GLS3} = 0.2426$; as expected $RMS_{WLS1} = 4 \cdot RMS_{GLS3}$. However, both the estimates of the parameters and their pertinent standard errors furnished by WLS_1 and GLS_3 are identical.
3) By observing "mean square" column of table 5 it appears that the values of θ influence inversely the estimation of γ^2. Furthermore being $\hat{\theta}$ given by POM model >1, the corresponding γ^2 is the smallest one, implying the maximum level of estimated precision.
4) Consider now the problem of determining the value of a future observation $\tilde{y}_0$ at the point x_0. For whatever value x_0 included in the range of X, $\tilde{y}_0 = \beta_0+\beta_1 x_0+e_0$. Such a $\tilde{y}_0$ is estimated by $\hat{\tilde{y}}_0 = \hat{\beta}_0+\hat{\beta}_1 x_0$. Under the assumption of homoschedasticity it can be shown (see Sen and Srivastava, (3), p.71) that the standard error of $\tilde{y}_0$ is given by:

[12]

where, coherently with [5],

Note that the first term of equation [12] accounts for the random error e_0 and the second one reflects the uncertainty in estimating $\hat{\tilde{y}}_0$ by means of $\hat{\beta}_0$ and $\hat{\beta}_1$.
In the case of POM model, according to Davidian and Giltinan ((4), p.291), the standard error of $\tilde{y}_0$ is estimated by:

where

Table 6. True and POM estimated values of the standard deviations for each group.

	i	1	2	3	4	5	6
(1)	σ_i	1	1.5	2	2.5	3	3.5
(2)	$\hat{\sigma}_{i(GLS)}$	0.8666	1.3777	1.9289	2.5126	3.1238	3.7590

Note that, in the case of heteroschedasticity, the first term of $S.E.(\tilde{y}_0)$: $\sqrt{\hat{\gamma}^2\hat{\mu}_0^{2\hat{\theta}}}$ properly accounts for the variance as modelled by the variance function adopted.í

Unfortunately the WLS approaches developed in section 2.2 do not enable modelling the variance. Thus in the presence of heteroscedasticity they can be used to obtain unbiased estimates of β_0 and β_1 together with their standard errors, but they fail in estimating the variance of a future observation and, consequently, in solving the problem of "calibration" ((4), p.276) in heteroschedastic biological settings.

4. References

1. Carroll RJ, Ruppert D. Transformation and weighting in regression. New York: Chapman and Hall, 1988.
2. Carroll RJ, Cline DBH. An asymptotic theory for weighted least squares with weights estimated by replication. Biometrika 1988; 75: 35-43.
3. Sen A, Srivastava M. Regression analysis: theory, methods and applications. New-York: Springer-Verlag, 1990.
4. Davidian M, Giltinan DM. Nonlinear models for repeated measurement data. New York: Chapman and Hall, 1995.
5. Davidian M, Haaland P. Regression and calibration with nonconstant error variance. Chemometrics and Intelligent Laboratory Systems 1990; 9: 231–248.