Available online at www.elixirpublishers.com (Elixir International Journal)

Statistics

Elixir Statistics 133 (2019) 53573-53576

ISSN: 2229-712X

Modification of WLS for analyzing Heteroscedastic Models with Missing Data

Gebriel Shamia and Entesar El-Saetti
Department of Statistics, Faculty of Science, University of Benghazi, Benghazi – Libya.

ARTICLE INFO

Article history:
Received: 28 June 2019;
Received in revised form:
20 August 2019;
Accepted: 31 August 2019;

Keywords
Unequal Cell Variances,
Heteroscedasity,
Weighted Least Square (WLS),
Type-iv Sums of Squares,
Proportionality Condition,
Adjustment for Bias.

ABSTRACT

The overriding problem with analyzing unbalanced data which leads to heteroscedasity models is that many methods are available and deciding between them can be a matter of some difficulty. Heteroscedasticity is a problem because ordinary least squares (OLS) in regression assumes that all residuals are drawn from a population that has a constant variance (homoscedasticity). When conditional heteroscedasticity is present, the practice of reweighting the data has long been abandoned in favor of estimating model parameters by ordinary *OLS*, in conjunction with using heteroscedasticity consistent standard errors. However, we argue for reintroducing the practice of reweighting the data, since doing so can lead to large efficiency gains of the resulting weighted least squares *WLS* estimator over *OLS* even when the model for reweighting the data is misspecified. The idea is that, the estimator can also be accompanied by the help of type of proportionality condition on cell sample size using harmonic mean. Special emphasis is given to nested model and on the unbalancedness in the data due to heterogeneity in the environmental conditions of experiments. Estimators of the parameters for this model are found to be independent of the weights under this condition.

© 2019 Elixir All rights reserved.

Introduction

In practice most experimental designs yield unbalanced data. This can include some cells having no data, within the class of unbalance data we make two distinct divisions. One is for data in which all cells contain data; none are empty. We call these all-cells-filled data. Complementary to this are some-cells-empty data; where in there are some cells that have no data, see El-Saeiti and Shamia. (2007). Unbalanced is difficult to define precisely, there it means unequal number of observation in cells of design. The basic problem is that messy data may give rise to heterogeneity of variance a cross cells. In this activity, unequal error variance may be due to the nature of the treatments. Snedcor and Cochran (1967, p. 324) gave some examples of unequal variance due to the treatment errors.

The analysis of variance test is adversely affected when the error variance is heterogeneous. There are other kinds of heteroscidasticity are possible

- The error variance may vary from one cell to another.
- The error variance may vary from one group of observation to another.
- The error variance may vary from one factor to another.

Some authors have dealt with these kinds, see James (1951), Bishop and Dudewicz (1978), Talukder (1978), and Shamia (1991). However, also many researchers have investigated the problem of combined analysis of a group of experiments with the heteroscedasity of error term. Notable among them are Yates and Cochran (1938), Khosla et.al. (1979), Bhuyan and Das (1983), Bhuyan and Miah (1989), and Khiar (1998). A more complete discussion on the sources of such variation can be found in Hartemink et al. (2001).

Problem of Missing Data

Unbalanced factor or empty cell can happen if there are missing data. *ANOVA* is most powerful where replication is equal for the different levels of each factor. However we can still perform; an *ANOVA* if we have unequal replication. For instance, certain treatment combinations may be more expensive or difficult to run than others; thus fewer observations are taken in those cells.

When unbalanced data are not too far from the balanced case, it is sometimes possible to use approximate procedures that convert, the unbalanced problem into a balanced one or partially bvalanced. The method of unweighted means is an approximate procedure when the number of observations in each cell is not dramatically different. This method is inappropriate when empty cells occur or when there are dramatically different. The approach used to develop sums of square for testing main effects and interactions is to represent the analysis of variance model as a regression model. However, there are several ways that this may be done, and their methods may result in different values for the sums of squares. Furthermore, the hypotheses that are not always direct analysis of those from the balanced case. For additional reading, see Searle (1987), Speed and Hocking (1976), Speed et.al. (1978), and Searle et.al. (1981).

Models Definition

Unbalanced factors often due to a typographical error, but the empty cell size message can be due to an illegal "nested" design only the random factor can be nested or (hierarchical).

Mixed models are used to describe data from experiments whose treatment structures involve some factors that are fixed and some that are random.

Tele:
E-mail address: gmshamia@yahoo.com

© 2019 Elixir All rights reserved

These models for describing such experiment involve two parts. One part is describing the random effect and the other is describing the fixed effect. Consequently the analysis of mixed consists of two types of analysis a random analysis and a fixed analysis. In some applications of nested classification, the classes have fixed effects that are to be estimated. An instance is an evaluation of breeding value of a set of five sires in cow raising. Each size is mated to a random group of dams. The model is

$$\boldsymbol{y_{ijk}} = \boldsymbol{\mu} + \boldsymbol{\alpha}_{[i]} + \boldsymbol{\beta}_{(i)j} + \in_{(ij)k} \qquad (1)$$

Where α_i are fixed effect associated with the sires but $\beta_{(i)j}$ and $\varepsilon_{(ij)k}$ are random variables corresponding to dams and offspring. Hence the model is called mixed. Here we could obviously expect the variance to differ from size to size. In fact, when the error variance is heterogeneous in this way, the F^{-}test tends to give many significant results.

In matrix notations the general mixed model can be written as following [Searle (1987)].

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad (2)$$

Where

Y is $n \times 1$ vector of observation.

β is vector of unknown fixed effects.

X is $n \times p$ design matrix that relates observations to fixed effects .

u is vector of unknown random effects.

Z is $n \times q$ matrix that relates observation to random effects.

ε is an non-observable random vector of residuals with null mean and

$$\boldsymbol{Var}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}\boldsymbol{\sigma^2}$$

Note that, in computer packages this model known as model III (the methods in SPSS are type-III and IV).

Description of Related Methods

Type-III sum of squares

Type-I and Type-II sums of squares usually are not appropriate for testing hypotheses for factorial *ANOVA* designs with unequal numbers. For *ANOVA* designs with unequal numbers, however, Type-III sums of squares test the same hypothesis that would be tested if the cell numbers were unequal, provided that there is at least one observation in every cell. Specifically, in no-missing-cell designs, Type-III sums of squares test hypotheses about differences in subpopulation (or marginal) means. When there are no missing cells in the homostedastic model, these subpopulation means are least squares means, which are the best linear-unbiased estimates (*BLUE*) of the marginal means for the design. See, Milliken and Johnson, (1992).

The Type-III sums of squares attributable to an effect is computed as the sums of squares for the effect controlling for any effects of equal or lower degree and orthogonal to any higher-order interaction effects (if any) that contain it. The orthogonality to higher-order containing interactions is what gives Type-III sums of squares the desirable properties associated with linear combinations of least squares means in *ANOVA* designs with no missing cells. But for *ANOVA* designs with missing cells, Type-III sums of squares generally do not test hypotheses about least squares means, but instead test hypotheses that are complex functions of the patterns of missing cells in higher-order containing interactions and that are ordinarily not meaningful. This method calculates the sum of squares of an effect in the design as the sum squares of adjusted for any other effects that do not contain it and orthogonal to any effects (if any) that contains it. The Type-III sum of squares has one major advantage in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant. Hence, this type of sums of squares is often considered useful for an unbalanced (homoscedastic) model with no missing cells. In a factorial design with no missing cells. This method is equivalent to the Yates Weighted-squares of means technique for fixed effect models.

Type-IV sum of squares

Type-IV sums of squares are computed by equitably distributing cell contrast coefficients for lower-order effects across the levels of higher-order are containing interactions. Type-IV sums of squares are not recommended for testing hypotheses for lower-order effects in *ANOVA* designs with missing cells, even though this is the purpose for which they were developed.

Statisticians who have examined the usefulness of Type-IV sums of squares have concluded that Type-IV sums of squares are not up to the task for which they were developed:

- Milliken and Johnson (1992, p. 204) has written: "It seems likely that few, if any, of the hypotheses tested by the Type-IV analysis of [some programs] will be of particular interest to the experimenter."
- Searle (1987, p.463-464) has written: "In general, [Type-IV] hypotheses determined in this nature are not necessarily of any interest." and (p. 465) "This characteristic of Type-IV sums of squares for rows depending on the sequence of rows establishes their non-uniqueness, and this in turn emphasizes that the hypotheses they are testing are by no means necessarily of any general interest."
- Hocking (1985, p. 152), in an otherwise comprehensive introduction to general linear models, writes: "for the missing cell problem, [some programs] offers a fourth analysis, Type-IV, which we shall not discuss."

So, we recommend that you use the Type-IV sums of square solution with caution, and that you understand fully the nature of the (often non-unique) hypotheses that are being testing, before attempting interpretations of the results. Furthermore, in *ANOVA* designs with no missing cells, Type-IV sums of squares are always equal to Type-III sums of squares, so the use of Type-IV sums of squares is either (potentially) inappropriate, or unnecessary, depending on the presence of missing cells in the design. The hypotheses selected depend upon the pattern of filled cells; and then the F-statistics are in accord with the good statistical practice of setting up a hypothesis and testing it. This does mimic reality, but only as an algorithmic automaton looking at which cells contain data, and not as a knowledgeable scientist thinking a bout data. Type-IV sum of squares is not unique, for a given set of some-cells-empty data, and so the corresponding hypothesis is not unique either.

A related technique is the weighted linear regression method, proposed by Yates (1934). This technique is also based on sum squares of the cell means, but the term in the sums of squares are weighted in inverse proportion to their variances. For further details of the procedure, see Searle (1987) and Speed, *et.al.* (1978).

The method of reweighted squares of means

In some cases it is desirable to apply differential weights to the observations, and to compute so-called weighted least squares estimates. This method is commonly applied when the variances of the residuals are not constant over the range of the independent variable values. In that case, one can apply the inverse values of the variances for the residuals as weights and compute weighted least square estimates. (In practice, these variances are usually not known, however, they are often proportional to the values of the independent variable(s), and this proportionality can be exploited to compute appropriate case weights.) Neter, *et.al.* (1985) describe an example of such an analysis.

The method of weighted least squares of means is used when the errors of variance are not equal then the error variance is heterogeneous. So to estimate the fixed factor we use the method of (WLS).

Note that, the expression for mixed model can be formed as in equation (2).

So that part of fixed effect estimators of the parameters are obtained by *WLS* method.

Talukder (1991) suggests general principle of *WLS* analysis for fixed the heteroscedastic model with

$$\mathrm{E}(\varepsilon) = 0 \quad \text{and} \quad \mathrm{Var}(\varepsilon) = dig(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) = V.$$

A diagonal matrix with error variance $\sigma_i^2{}'s$, as the diagonal elements and these variances may not be all distinct.

For known error variances, the *WLS* estimator β of is obtained by minimizing the quadratic form

$$\acute{\varepsilon}\, v^{-1}\varepsilon = (Y\text{-}X\beta)'\, V^{1}(Y\text{-}X\beta)$$

And the resultant normal (GLS) equations are given by

$$X'\, V^{1}\, X\beta = X'\, V^{1}Y, \tag{3}$$

where $V^{1} = W^{\delta}$ is a diagonal matrix with $w_{(i)j}$ as the diagonal elements.

Here in our model (equation (1)), the weights, due to Hartmink, *et.al.* (2001), are given by

$$w_{(i)j} = \frac{(1/\sigma^2_{(ii)j})}{\sum_i^a \sum_j^b (1/\sigma^2_{(i)j})} = \frac{\hbar}{ab}((\sigma^2_{(i)j})^{-1} = \frac{\hbar}{ab}\left(\frac{\lambda_i}{n_{(i)j}}\right),$$

where $\hbar$ is the harmonic mean of the cell variances, and we assume that $n_{(i)j} \cong \lambda_i \sigma^2_{(i)j}$ is the proportional condition.

Here, λ_i is the level-specific proportionality constant which varies with the levels of the factor-*A*.

Now, let $Z = W^{\delta/2}$ Y and $A = W^{\delta/2}\, X$; where $W^{\delta/2}$ is a diagonal matrix with $W_i^{1/2}$ as the diagonal elements. Then Cov(*Z*)=I, an identity matrix; and the equation (3) can be written such as

$$A'\, A\,\beta = A'\, Z \tag{4}$$

With rank (*A*) = rank (*X*). These are the normal equation of ordinary least square (OLS) method in transformed data (*Z*) so that the *WLS* estimators posses the optimal properties of OLS method. The usual least square analysis can therefore be performed here also under assumption of normality of errors. The estimator of $\hat{\beta}$ satisfies the optimality properties of *OLS* estimation.

For the mixed model (equation (2)), the difficulty with this method is that *V* matrix is very large one. Henderson (1963) such as suggested an alternative method

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{U} \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}$$

in which $\hat{\beta}$ is a solution to equation (3). Here

$$R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1} = V^{-1}\cdot$$

Conclusion

A large statistical literature is devoted to missing values. If a data set contains a large share of missing entries, the way they are imputed can affect the analysis substantially, for example inducing spurious features. However, some statistical techniques allow imputation of missing values as part of the estimation process.

This study attempts to investigate the theoretical side of design analysis in the case of unbalanced models with missing data in the situation where approximate methods such as unweighted means are inappropriate. It was examined by using proportionality condition on cell sample size using harmonic mean for variances the cell to get the weighted value to translate the model from unbalanced into the partial balanced form as exact method. By applying such modification of *WLS,* we adjusting (reduction) of sum of squares and provide an improvement of the analysis of variance method.

This procedure is used for mixed-nested design with unequal cell variances to exemplify the method, but it can be used for crossed-nested designs and split-plot designs hving random, mixed, and fixed-effect models as well.

Reference

[1] Bhuyan, K.C. (1984). A group of split-plot designs with a heteroscedastic model. *Austral. J. Statistic.* 26(2), 133-141.

[2] Bhuyan, K.C. and Das, A.K. (1983). On the combined analysis of group of factorial experiments with a heteroscedastic model. *Jahagirnagar Review, Part A*, Vol. 7, Bangladesh.

[3] Bhuyan, K.C. and Miah, A.B.M.A.S. (1989). A group of randomized block designs. *Sankhya*, 51(3), 429-433.

[4] Bishop, T.A. and Dudewicz, E.J. (1978*).*Exact analysis of variance with unequal cell variances: Test procedures and Tables. *Technometrics*, 20, 419-430.

[5] Bishop, T.A. and Dudewicz, E.J. (1981). Heteroscedastic ANOVA. *Sankhya*, 33B, 40-57.

[6] Bliss, C.I. (1952). *The Statistics of Bioassay*. Academic Press, Inc. New York.

[7] Casella, G., McCullock, C. E. and Searle, S. R., (1992). *Variance components.* New York: Wiley.

[8] Cochran, W.G. (1937). Problem arising in the analysis of a series of similar experiments. *Suppl. Jour. Roy. Stat. Sco.* 4, 102-118.

[9] Cochran, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Jour. of the American Statistical Association*. 34, 492-510.

[10] Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.

[11] Cochran, W.G. and Yates, F. (1938). The analysis of groups of experiments. *Jour. Agr. Sci*, 28, 556-580.

[12] Corbeil, R.R. and Searle, S.R. (1976). A comparison of Variance Component Estimators. *Biometrics*. 32. 779-791.

[13] David, B.D., Roger, A.H. and Susan, D.H. (1975). Estimating heteroscedastic variances. *Journal of the American Statistical Association*. 70, 380-385.

[14] David, L.W. and Urquhart, N.S. (1978). Linear models in messy data: some problems and alternatives. *Biometrics*. 34, 696-705.

[15] El-Saeiti, I.N. and Shamia, G.M. (2007). Analysis of four-stage nested design with missing data. Journal of Science and its Applications. 1(1), 24-32.

[16] Fisher, R.A. (1925): *statistical methods for research workers, First edition*, Oliver and Boyd, London.

[17] Giesbrecht, F.G. (1983): An efficient procedure for computing MINQUE of variance components and generalized least squares estimates of fixed effects. *Communication in statistics.* A 12, 2169-2177.

[18] Goodnight, J.H. and Hemmerle, W.J. (1978): *A simplified algorithm for the W-transformation in variance component estimation.* SAS Technical report R-104, SAS Institute, and Cary, North Carolina.

[19] Gosslee, D.G. and Lucas, H.L. (1965): Analysis of variance of disproportionate data when interaction is present. *Biometrics*. 21, 115-133.

[20] Hartley, H.O. and Rao, J.N.K. (1967): Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*. 54, 93-108.

[21] Hartmink, A.J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. (2001). *Maximum likelihood estimation of optimal scaling factors for expression array normalization. Proc. SPIE 4266, Microarrays: Optical Technologies and Informatics*. http://groups.csail.mit.edu/cgs/pubs/spie.pdf

[22] Hocking, R. R. and Speed, F. M. (1975): A full rank analysis of some linear model problems. *Journal of the American Statistical Association*. 70, 707-712.

[23] Hocking, R.R. (1985): *The Analysis of Linear Models*. Books/ Coles, Monterey, CA.

[24] James, G.S. (1951): The comparison of several groups of observation when the ratios of the population variance are unknown. *Biometrika*. 38, 324.

[25] Khiars, A.S. (1998): *On Combined Analysis of a Group of Nested Designs.* M.Sc. Thesis, Garyounis Univ.

[26] Khosla, R.K., Rao, P.P and Das, M.N. (1979): A note on the study of the experimental error in-groups of agricultural field experiments conducted in different years. *J. Indian Soc. Agri. Statist*. 31, 65-68.

[27] LaMotte, L.R. (1973): Quadratic estimation of variance components. Biometrics. 29, 311-330.

[28] Marssaglia, George, and Styan, George (1974): Equalities and Inequalities for rank of matrices. *Linear and Multilinear Algebra*. 2, 269-292.

[29] Milliken, G. A., and Johnson, D. E. (1992*): Analysis of messy data: Vol. I. Designed experiments.* New York: Chapman & Hall.

[30] Montgomery, D.C. (1990): *Probability and statistic in Engineering and Management science.* Wily, New York.

[31] Neter, J., Wasserman, W., and Kutner, M. H. (1985): *Applied linear statistical models: Regression, analysis of component of variance, and experimental designs*. Homewood, IL: Irwin.

[32] Qaas, R.L. and Bolgiano, D.C. (1979): *Sampling variance of the MIVQUE and method 3 estimator of sire component of variance*. Cornell Univ. Ithaca New York, 99-106.

[33] Rao, C.R. (1972*):* Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*. 67, 112-115.

[34] Rao, C.R. (1973): On the estimation of heteroscedastic variance. *Biometrics*. 29, 11-24.

[35] Rao, S.R.S. (1997): *Variance components estimation mixed models, methodologies and applications.* Chapman & Hall.

[36] Satterthwaite, F.E. (1946): An approximate distribution of estimates of variance components. *Biometrics*. 2, 110-112.

[37] Searle, S.R. (1979): Annotated computer output for analysis of variance of unequal-subclass-numbers data. *The American statistician*. 33, 222-223.

[38] Searle, S.R. (1987): *Linear models for unbalanced data*. John Wily and sons, Inc. New York.

[39] Shamia, G.M. (1991): *Analysis of some designs of experiments with heteroscedastic models.* M.Sc. Thesis, Garyounis Univ.

[40] Snedecor, G.W and Cochran, W.G (1967): *Statistical methods.* Ames, Iowa, USA.

[41] Swallow, W.H. (1974): *Minimum norm and minimum variance quadratic unbiased estimation of variance components.* Ph.D. Thesis. Cornell University, Ithaca, NY.

[42] Talukder, M.A.H. (1978): General block designs with a heteroscedastic model. *Sankhya.* 40B, 217-226.

[43] Yates, F. (1934): The analysis of multiple classifications with unequal numbers in the different classes. *Jour. of the American Statistical Association.* 29, 51-66.