

On the Coincidence of the Ordinary Least Squares Estimators with the Generalized Least Squares ones

Sulla Coincidenza degli Stimatori dei Minimi Quadrati Ordinari con gli Stimatori dei Minimi Quadrati Generalizzati

Amedeo De Luca

Department of Statistics – Università Cattolica del Sacro Cuore
e_mail: amedeo.deluca@mi.unicatt.it

Riassunto: nel modello di regressione lineare con variabile dipendente dicotomica, caddendo l'ipotesi di omoschedasticità della componente accidentale, per la stima dei parametri si utilizza il metodo dei minimi quadrati generalizzati (GLS) suggerito da Goldberger (1964). Se in detto modello - in un contesto sperimentale - anche i predittori sono di natura dicotomica, utilizzando il metodo GLS si ottengono i medesimi stimatori del metodo dei minimi quadrati ordinari (OLS), come si dimostra nel contributo. Nelle correnti applicazioni il risultato teorico raggiunto ha rilevanza nell'impiego (oltre che nella realizzazione) del software per l'analisi dei dati categorici.

Keywords: heteroskedasticity, generalized least squares.

1. Introduction

The most important part of the experimental design concerns the case in which the theoretical model, linking the dependent variable with levels factors, is a linear function. If this variable is dichotomous (0, 1), considering the heteroskedasticity of the disturbance term, we can have recourse to the GLS method to estimate the parameters of the model. Even if the qualitative experimental factors are represented in terms of dummy variables (0, 1), the GLS estimators coincide with the Ordinary Least Squares (OLS) estimators; this is what we have to show in our contribution.

2. Estimators of the Linear Regression Model on Dummy Variables

Consider a linear regression model without constant term in which the binary dependent variable Y assumes the value 1 if a given category is observed, otherwise it assumes the value 0. Similarly, the independent variables are represented by dummy coding; we define for each variable a set of 0-1 dummy variables Z_i ($i = 1, 2, \dots, I$) so that $Z_i = 1$ if category i th is observed, in all other cases $Z_i = 0$.

The model form in matrix notation, considering - without loss of generality - a single qualitative independent variable, can be written:

$$\mathbf{y} = \mathbf{Z} \boldsymbol{\delta} + \mathbf{e}, \quad (1)$$

where: $\mathbf{y}$ denotes the $n \times 1$ column vector of the observed dichotomous regressand y_j , $j = 1, 2, \dots, n$; $\mathbf{Z}$ denotes the $n \times I$ matrix of observations on the dummy regressors, whose

elements are zeros or ones, with categories z_{ij} , $i = 1, 2, \dots, I$, of I considered levels factor, $j = 1, 2, \dots, n$; δ denotes the $I \times 1$ column vector of unknown parameters δ_i ; $\mathbf{e}$ denotes the $n \times 1$ column vector of stochastic heteroskedastic e_j errors; we assume that the e_j are uncorrelated with each other and with the Z_i . The dichotomous observations y_j are interpreted as values of Bernoulli's random variables, with parameter p_j (fraction of the cases with value 1 in $\mathbf{y}$ vector for which the i th category of the predicted variable is observed). The quantities e_j have zero mean and variance $p_j(1 - p_j)$; the latter can be denoted in the following form: $M(e_j^2) = (\mathbf{z}'_j \delta) (1 - \mathbf{z}'_j \delta)$, where $\mathbf{z}'_j$ indicates the row vector of the independent variable for the j th observation in the matrix $\mathbf{Z}$ of the indicator variables. To estimate the elements of the vector of coefficients δ the GLS method is required; in this case the variance-covariance matrix of disturbances is shown:

$$M(\mathbf{e}\mathbf{e}') = \sigma^2 \begin{bmatrix} \mathbf{z}'_1 \delta (1 - \mathbf{z}'_1 \delta) & 0 & \dots & 0 \\ 0 & \mathbf{z}'_2 \delta (1 - \mathbf{z}'_2 \delta) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathbf{z}'_n \delta (1 - \mathbf{z}'_n \delta) \end{bmatrix} \quad (2)$$

(Goldberger, 1964, pp. 244-245). The best linear unbiased estimators are obtained by means of a two-step procedure: once obtained the OLS estimator $\hat{\delta}_i$ for the parameters δ_i (first step), the same are inserted in the matrix of expression (2). Then the resultant matrix is used to obtain the estimators $\tilde{\delta}_i$ of the GLS (second step), as it is here described. Defining the following diagonal matrix:

$$\Lambda = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sqrt{\lambda_n} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{\mathbf{z}'_1 \delta (1 - \mathbf{z}'_1 \delta)} & 0 & \dots & 0 \\ 0 & 1/\sqrt{\mathbf{z}'_2 \delta (1 - \mathbf{z}'_2 \delta)} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/\sqrt{\mathbf{z}'_n \delta (1 - \mathbf{z}'_n \delta)} \end{bmatrix} \quad (3)$$

premultiplying the elements of equation (1) by the transforming matrix Λ , we have:

$$\Lambda \mathbf{y} = \Lambda \mathbf{Z} \delta + \Lambda \mathbf{e}, \quad (4)$$

consequently the variance-covariance matrix of the transformed disturbance variable is:

$$M(\Lambda \mathbf{e} \Lambda') = \sigma^2 \mathbf{I}. \quad (5)$$

Thus it is correct to apply the OLS method directly to the relationship in equation (4). In the specific case here studied of the dependent variable and the dichotomous predictors, the coincidence of the GLS with OLS estimators has been empirically checked. Extending the result given in De Luca (1999), the formal demonstration is next given.

3. The Coincidence of the OLS Estimators with the GLS ones

Lemma

In the linear regression model (1) without intercept of a dichotomous regressand on dummy variables, the elements δ_i of the vector of parameters δ correspondent to a

qualitative independent variable estimated by the OLS method, are equivalent to the conditional expectations of Y given z_i , and are denoted by $M(Y_{|z_i})$, as it is in De Luca (1999):

$$\delta_i = \sum_{j=1}^n y_j z_{ij} / \sum_{j=1}^n z_{ij} = M(Y_{|z_i}), \quad i = 1, 2, \dots, I. \quad (6)$$

Theorem

In a linear stochastic regression model if the regressand is dichotomous and also the predictors are dichotomous, the GLS estimators $\tilde{\delta}_i$ of coefficients δ_i are equal to OLS estimators $\hat{\delta}_i$:

$$\tilde{\delta}_i = \hat{\delta}_i, \quad i = 1, 2, \dots, I. \quad (7)$$

Proof.

With reference to the transformed model in equation (4), in order to estimate the parameters, the expression to be minimized is:

$$S(\Lambda y, \tilde{\delta}) = \sum_{j=1}^n (\sqrt{\lambda_j} y_j - \sum_{i=1}^I \sqrt{\lambda_j} z_{ij} \tilde{\delta}_i)^2 \quad (8)$$

so the least squares estimators are the solution of: $\frac{\partial S}{\partial \tilde{\delta}_h} = 0$, $h = 1, 2, \dots, I$. Differentiating

$S(\Lambda y, \tilde{\delta})$ with respect to each element $\tilde{\delta}_h$ of $\tilde{\delta}$ and setting expression (8) equal to 0, we have:

$$-\sum_{j=1}^n \sqrt{\lambda_j} z_{hj} (\sqrt{\lambda_j} y_j - \sum_{i=1}^I \sqrt{\lambda_j} z_{ij} \tilde{\delta}_i) = 0;$$

it follows:

$$\sum_{i=1}^I \tilde{\delta}_i (\sum_{j=1}^n \lambda_j z_{hj} z_{ij}) = \sum_{j=1}^n \lambda_j y_j z_{hj} \quad h = 1, 2, \dots, I. \quad (9)$$

Taking into consideration that z_{hj} and z_{ij} are dummy variables which can take on only two values we have designated 0 and 1, and that for $h \neq i$ the result is: $z_{hj} z_{ij} = 0$, while for $h = i$ the result is: $z_{hj} z_{ij} = z_{ij}^2 = z_{ij}$, developing the summation denoted with $\sum_{i=1}^I$ in the first member of equation (9) for each equation, the system (9) is reduced to a linear system of I equations in one unknown:

$$\tilde{\delta}_i (\sum_{j=1}^n \lambda_j z_{ij}) = \sum_{j=1}^n \lambda_j y_j z_{ij}, \quad i = 1, 2, \dots, I, \quad (10)$$

and the generic estimator of the equation (10) is represented in the following form:

$$\tilde{\delta}_i = \sum_{j=1}^n \lambda_j y_j z_{ij} / \sum_{j=1}^n \lambda_j z_{ij}, \quad i = 1, 2, \dots, I. \quad (11)$$

In the numerator of equation (11) it is: $y_j z_{ij} = 0$ if $y_j \vee z_{ij} = 0$; $y_j z_{ij} = 1$ if $y_j \wedge z_{ij} = 1$. In addition, according to equation (3) it is:

$$\lambda_j = 1 / \mathbf{z}'_j \hat{\delta} (1 - \mathbf{z}'_j \hat{\delta}), \quad (12)$$

then, taking into account that the elements z_{ij} of the row vector $\mathbf{z}'_j$ are mutually exclusive and exhaustive, for the generic estimator $\hat{\delta}_i$ we have, for example, for $i = 3$: $\mathbf{z}'_j \hat{\delta} = [0 \ 0 \ 1 \ 0 \ 0] \hat{\delta} = z_{ij} \hat{\delta}_i = \hat{\delta}_i$, it follows that in the vector $\mathbf{z}'_j$ there is only one element (i th) which is not null. Since the first member of (11) does not depend on j , considering equation (12) the result is: $\lambda_j = \lambda$; it follows that the numerator of (11) can be written:

$$\sum_{j=1}^n \lambda_j y_j z_{ij} = \lambda \sum_{j=1}^n y_j z_{ij}. \quad (13)$$

Taking into account the relationship: $\lambda_j = \lambda$ also for the denominator of (11), this is reduced to, by virtue of equation (13):

$$\hat{\delta}_i = \lambda \sum_{j=1}^n y_j z_{ij} / \lambda \sum_{j=1}^n z_{ij} = \sum_{j=1}^n y_j z_{ij} / \sum_{j=1}^n z_{ij} = M(Y_{|z_i}) = \hat{\delta}_i, \quad (14)$$

finally equation (7), taking (6) into consideration, is thus shown $\square$

4. Conclusions

In the field of data analysis, the theoretical result which has been obtained is most significant for the algorithms of iterative calculations, both to estimate the parameters (typically in the class of generalized linear models, whose development has been increased by a noticeable improvement in the power of electronic calculus), and to utilize, as well as to produce, software for the analysis of categorical data. For example, if in some practical applications the response variable of the linear model does not naturally assume values between 0 and 1, it is necessary to impose the $0 \leq p \leq 1$ inequality constraints on the expected values (p) and have recourse to Quadratic Programming (QP) (restricted least squares); by virtue of the coincidence of OLS estimators with the GLS estimators the objective function of QP is considerably simplified (De Luca, 2001).

References

- De Luca A. (1999) Un modello di misurazione della customer satisfaction con codifica binaria additiva dei predittori ordinali, in: *Valutazione della Qualità e Customer Satisfaction*, Giornata di Studio SIS, Bologna, 24 settembre, 219-230.
- De Luca A. (2001) Un modello di valutazione della qualità percepita con variabile risposta politomica, in: *Processi e Metodi Statistici di Valutazione*, Atti del Convegno Intermedio SIS, Roma, 4-6 giugno, 87-90.
- Goldberger A. S. (1964) *Econometric Theory*, Wiley, New York.