University of Oxford

More Notes for Least Squares

Órlaith Burke

Michaelmas Term, 2010

Department of Statistics, 1 South Parks Road,
Oxford OX1 3TG

Within multiple linear regression, the response variable is approximately linearly related to the explanatory (or independent) variables. Consider a situation with n observations of a response variable and p independent variables.

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i \tag{1}$$

for (i = 1,. . .,n; n $\geq$ p), where n is the number of observations in the dataset. There are p independent (or explanatory) variables and hence, p parameters to estimate.

The x_i terms are not random and are measured with negligible error. The model is said to be 'linear in the parameters'. Therefore, transformation of the response and/or independent variables is permitted.

In matrix notation, the model is written as:

$$y = X\beta + \varepsilon \tag{2}$$

where

$$\begin{aligned} y &= (y_1, y_2, \ldots, y_n)^t \\ X &= \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{pmatrix} \\ \beta &= (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^t \\ \varepsilon &= (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^t \end{aligned} \tag{3}$$

There are several approaches to estimating the regression parameters β.

0.1 Ordinary Least Squares (OLS)

The Ordinary Least Squares (OLS) approach to multiple linear regression was introduced by Gauss in 1794. The OLS procedure is the simplest type of estimation procedure used in statistical analyses. However, in order to benefit from the well-behaved properties of an OLS estimate, a number of assumptions must be satisfied.

0.1.1 Assumptions

The estimate of the regression parameters in Equation (2) is denoted as $\hat{\beta}$. There are five assumptions necessary to produce unbiased estimators using OLS. Additional assumptions must be satisfied in order for the estimate to have other favorable qualities.

The assumptions for OLS estimation are:

1. The model must be linear in the parameters.
2. The data are a random sample of the population i.e. residuals are statistically independent/uncorrelated from each other
   i.e. $\text{Cov}(\varepsilon_i,\varepsilon_i') = 0 \; \forall$ i $\neq$ i'.
3. The independent variables are not too strongly collinear.
4. The independent variables are measured precisely such that measurement error is negligible.
5. The expected value of the residuals is always zero
   i.e. $\mathbb{E}(\varepsilon) = 0$.
6. The residuals have constant variance (homogeneous variance)
   i.e. $\mathbb{V}(\varepsilon) = \sigma^2\mathbb{I}$.
   (Tests for this include the Bruesch-Pagan test and Brown-Forsythe test).
7. The residuals are Normally distributed i.e. $\varepsilon \sim N(0, \sigma^2\mathbb{I})$.

0.1.2 Parameter Estimation

Maximum Likelihood Approach

The OLS estimator ($\hat{\beta}$) as defined in Equation (2) above, is identical to the Maximum Likelihood estimate for the regression parameters. The method of Maximum Likelihood essentially chooses the values of the parameters that are most consistent with the data. Assuming that the residuals follow a Normal distribution with mean zero and variance $\sigma^2\mathbb{I}$, the density, f_i, of the response variable, y_i, can be written:

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2\right]. \tag{4}$$

The likelihood function, L, is the product of the densities for the n observations in the dataset. The likelihood function is denoted as a function of the unknown parameters β and σ^2.

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2\right]. \end{aligned} \tag{5}$$

Taking logs, the log-likelihood is of the form:

$$\begin{aligned} l(\beta, \sigma^2) &= lnL(\beta, \sigma^2) \\ &= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \left[\frac{1}{2} \left(\frac{y_i - X_i \beta}{\sigma} \right)^2 \right]. \end{aligned}$$

The score function is calculated by taking the derivative of the log-likelihood:

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= -\frac{1}{2\sigma^2} \left(\frac{\partial [y'y - 2\beta' X'y + \beta' X'X\beta]}{} \right) \\ &= -\frac{1}{2\sigma^2} \left(-2X'y + 2X'X\beta \right) \\ &= -\frac{1}{\sigma^2} \left(-X'y + X'X\beta \right). \end{aligned} \tag{6}$$

To find the maximum likelihood estimator, set the score function equal to zero:

$$\begin{aligned} \frac{1}{\sigma^2} \left(-X'y + X'X\beta \right) &= 0 \\ X'X\beta &= X'y \\ \hat{\beta} &= (X'X)^{-1} X'y. \end{aligned} \tag{7}$$

Least Squares Approach

Estimation of the regression parameters

Least Squares estimation is used to estimate the regression parameters β and σ^2 (the residual variance parameter). The estimator ($\hat{\beta}$) is that which minimizes the residual sum of squares, i.e. satisfies the equation:

$$\frac{\partial}{\partial \beta}\left[(\mathbf{y} - X\beta)'(\mathbf{y} - X\beta)\right] = 0. \tag{8}$$

This partial differentiation gives:

$$-2X'y + 2(X'X)\hat{\beta} = 0 \tag{9}$$

which leads to the set of simultaneous equations known as the 'normal equations':

$$X'X\hat{\beta} = X'y. \tag{10}$$

Solving for the estimator ($\hat{\beta}$):

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{11}$$

Estimate the variance parameter

The variance parameter (σ^2) is estimated using the unbiased estimator, s^2 (the residual mean square):

$$s^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n - p} \tag{12}$$

where p is the number of parameters in the regression model.

This parameter quantifies the variation of the residuals about the estimated regression line ($\hat{y} = X\hat{\beta}$). The residual mean square is an unbiased estimator (assuming that the proposed model is correct).

0.1.3 Properties of the estimator

Unbiased estimator

Under the assumption that $\mathbb{E}(\varepsilon) = 0$, it can be shown that the OLS estimator ($\hat{\beta}_{OLS}$) is an unbiased estimator:

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \mathbb{E}((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X'\mathbb{E}y \\ &= (X'X)^{-1}X'X\beta \\ &= \beta. \end{aligned} \tag{13}$$

Variance of the estimator

Under the assumptions of uncorrelated residuals with homogeneous variance, the variance-covariance matrix for $\hat{\beta}$ is given by:

$$\mathbb{V}(\hat{\beta}) = \sigma^2(X'X)^{-1}. \tag{14}$$

0.1.4 Advantages & Deficiencies

The first four assumptions (Section 0.1.1) are needed for any estimation of the regression parameters. Assumptions 5 to 7 give the OLS estimator other favorable qualities.

If the expected value of the residuals is always zero (Assumption 5), then the OLS estimator is unbiased (as shown above).
If the residuals have homogeneous variance (Assumption 6), then the OLS estimator has the minimum variance of all linear unbiased estimators (BLUE) by the Gauss-Markoff Theorem.
Finally, if the residuals are also Normally distributed (Assumption 7), then t and F tests can be used in inference. It should be noted that the bias and dispersion properties of the estimator do not depend on the Normality of the residuals.

However, if all seven of the assumptions are satisfied, the OLS estimator has the uniformly minimum variance of all unbiased estimators (UMVU).

Another characteristic of OLS is that in a model $Y = AX + BZ$, we can estimate $\hat{A}$ and $\hat{B}$ by first regressing Y on X, calculating the residuals and regressing those residuals on Z.

However, there are also some disadvantages to OLS. The assumptions required for OLS are stringent. If any of these assumptions are not met, the OLS estimation procedure breaks down and the estimator no longer enjoys all of the properties discussed above.

Of the seven OLS assumptions, the two which most often cause issues are

the assumption of homogeneous variance in the residuals (Assumption 6) and Normally distributed residuals (Assumption 7). If these conditions do not hold, the OLS estimator will be unbiased and consistent[1]. However, the estimates will be inefficient[2] i.e. OLS will give incorrect estimates of the parameter standard errors.

The third assumption listed above (Section 0.1.1) can also raise issues in practical applications. Multicollinearity in the data can caused serious problems (Section 0.1.5). OLS estimation does not allow for correlation within the residuals. Other approaches (discussed later) do not require the assumption of homogeneous independent Normal residuals.

0.1.5 Multicollinearity

Assumption 3 requires that the independent variables are not too strongly collinear. Multicollinearity is an issue often raised in multiple regression since it prohibits accurate statistical inference. This condition occurs when there are near-linear relationships between the independent variables which make up the columns of X. In mathematical terms, it implies that there exists a set of constant d_j (which are not all zero) such that:

$$\sum_{j=1}^{k} d_j x_j \cong 0 \tag{16}$$

where x_j is the j^{th} column of X.

[1] An estimator is said to be a consistent estimator if, for all positive c,

$$n \overset{lim}{\to} \infty \; P\left(|T - \theta| \geq c\right) = 0, \tag{15}$$

where n is the sample size.

[2] The efficiency of an unbiased estimator is the ratio of its variance to the Cramér-Rao lower bound. For an efficient estimator, this ratio is 1.

The normal equations are then said to be 'ill-conditioned'. If this term is exactly zero, the linear dependencies are exact and the (X'X) matrix is singular.

Procedures for reducing multicollinearity in the independent variables include variable selection, transformations (which may reduce the dimensionality of the system) and (the somewhat controversial) ridge regression.

Ridge regression replaces the usual parameter estimate ($\hat{\beta}$) with:

$$\hat{\beta}^* = (X'X + q\mathbb{I})X'y \tag{17}$$

for some constant q. The value of q (often refered to as the shrinkage parameter) is determined through a study of the ridge trace (a plot of the elements of β^* against q).

0.2 Generalised Least Squares (GLS)

Generalised Least Squares (GLS) was introduced by Aitken (1935). The model equation is of the same form as that used in OLS (Equation 2) with one main difference. The residuals need not follow the same assumptions as required by OLS.

GLS is designed to produce an optimal unbiased estimator of β for situations with heterogeneous variance. In such cases, OLS estimates are unbiased and consistent but inefficient. OLS tends to underestimate the parameter standard errors which, in turn, affects the hypothesis testing

procedures.

0.2.1 Assumptions

The GLS equation is identical to the OLS equation:

$$y = X\beta + \varepsilon \tag{18}$$

with the exception that:

$$\varepsilon \sim N(0, \Omega). \tag{19}$$

OLS is clearly a special case of GLS ($\Omega = \sigma^2 \mathbb{I}$).

0.2.2 Parameter Estimation

Maximum Likelihood Approach

The log-likelihood of a single observation (i where i = 1, . . . , n) is written in terms of the unknown regression parameter β and the variance matrix Ω:

$$l_i(\beta, \Omega) = -\frac{1}{2}\left(nln(2\pi) + ln(det\Omega) + (y_i - X_i\beta)\Omega^{-1}(y_i - X_i\beta)\right). \tag{20}$$

The log-likelihood for the entire dataset is then:

$$L(\beta, \Omega) = \sum_{i=1}^{n} l_i(\beta, \Omega). \tag{21}$$

The score vector is calculated by taking the derivatives with respect to the parameters.

$$\begin{aligned}
\frac{\partial}{\partial\beta}L(\beta,\Omega) &= \sum_{i=1}^{n}\frac{\partial}{\partial\beta}l_j(\beta,\Omega) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}\left((y_i - X_i\beta)\Omega^{-1}(y_i - X_i\beta)\right) \\
&= \sum_{i=1}^{n}X_i'\Omega^{-1}(y_i - X_i\beta) \qquad (22)
\end{aligned}$$

Setting the score function to zero gives:

$$\begin{aligned}
\beta_{MLE} &= \left(\sum_{i=1}^{n}X_i'\Omega^{-1}X_i\right)^{-1}\sum_{i=1}^{n}X_i'\Omega^{-1}y_i \\
&= \hat{\beta} \qquad (23)
\end{aligned}$$

where β_{MLE} represents the Maximum Likelihood estimate and $\hat{\beta}$ the GLS estimate of the parameters.

Estimation by Regression

The heterogeneous variance of the residuals is taken into account in the GLS estimation of the regression parameters ($\hat{\beta}$) and the standard errors of $\hat{\beta}$.

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \qquad (24)$$

and

$$\hat{\mathbb{V}}(\hat{\beta}_{GLS}) = (X'\Omega^{-1}X)^{-1}. \qquad (25)$$

0.2.3 Advantages & Deficiencies

GLS allows for heterogeneous variance in the residuals. The GLS estimator, $\hat{\beta}$, is an unbiased estimator. It is also the Maximum Likelihood estimator (under the assumption that $\varepsilon \sim N(0, \Omega)$). If the assumption of Normality is relaxed, the general Gauss-Markov theorem applies and $\hat{\beta}$ is the 'best' (i.e. minimum variance) linear unbiased estimator.

The assumptions of GLS allow for heterogeneous variance within the residuals. This can also be expanded to allow for non-zero covariances between the residual terms. This in turn can be used to allow for different forms of correlation (such as cross-correlation or autocorrelation) in the data. However, the variance-covariance matrix is constant and cannot vary through time.

As in OLS, multicollinearity (Section 0.1.5) can also be a serious issue in GLS estimation.
One of the main issues associated with GLS is that the variance-covariance matrix for GLS (Ω) is unknown. Therefore, an estimation must be used of this matrix ($\hat{\Omega}$) must be used. This brings us to Feasible Generalised Least Squares (FGLS).

0.3 Feasible Generalised Least Squares (FGLS)

Feasible Generalised Least Squares (FGLS) follows the same method as GLS with the exception that an estimated variance-covariance matrix for the residuals ($\hat{\Omega}$) is used in place of the unknown Ω.

0.3.1 Parameter Estimation

In order to generate $\hat{\Omega}$, OLS is first applied to the model. This gives consistent estimates of β. The residuals are then estimated:

$$\hat{\varepsilon} = y - X\hat{\beta}. \tag{26}$$

These residual values are also consistent and are used to estimate the variance-covariance matrix Ω:

$$\hat{\Omega} = \hat{\varepsilon}\hat{\varepsilon}'. \tag{27}$$

The estimated variance-covariance matrix ($\hat{\Omega}$) is then substituted into the GLS equations to give the FGLS estimate:

$$\hat{\beta} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y \tag{28}$$

and

$$\hat{\mathbb{V}}(\hat{\beta}) = (X'\hat{\Omega}^{-1}X)^{-1}. \tag{29}$$

0.3.2 Advantages & Deficiencies

FGLS allows for the practical application of GLS. It therefore, enjoys the same advantages and suffers the similar disadvantages to GLS (Section 0.2.3). FGLS has been shown to be equivalent to the Maximum Likelihood estimator in its limit. It therefore also possesses the asymptotic properties of the Maximum Likelihood.

One down-side to FGLS is that while the procedure works well for large samples, little is known about its behaviour in small finite sample.