To Estimate or Not to Estimate?

Benjamin Kedem and Shihua Wen [1]

In linear regression there are examples where some of the coefficients are known but are estimated anyway for various reasons not least of which is failure to recognize any problem. Over-fitting is a special case. We show that this practice may lead to inefficient estimates. A simulation study confirms closely the theory.

1 Introduction

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ be a full rank design matrix where $\mathbf{X}_1$ and $\mathbf{X}_2$ are matrices of dimensions $n \times p$ and $n \times q$, respectively. Consider the linear model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is a column vector of observations of length n and where $\boldsymbol{\beta}_1$ is a *known* column vector of length p, but $\boldsymbol{\beta}_2$ is an *unknown* column vector of coefficients of length q which must be estimated. Assume that the error term $\boldsymbol{\epsilon}$ has mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}_n$, $\mathbf{I}_n$ is the identity matrix of dimension $n \times n$. Since $\boldsymbol{\beta}_1$ is known it need not be estimated, but the question is what happens if it is estimated along with $\boldsymbol{\beta}_2$ anyway?

A special case of this is encountered in *over-fitting* when $\boldsymbol{\beta}_1 = \mathbf{0}$ is a "known" vector whose inadvertent estimation results in inflated variability of the components of $\hat{\boldsymbol{\beta}}_2$. See the discussion in [2, Sec. 9.2.2].

[1]Benjamin Kedem is Professor and Shihua Wen is a graduate student in the Department of Mathematics, University of Maryland, College Park, MD 20742, phone: 301-405-5061/5112, fax: 301-314-0827, email: bnk@math.umd.edu, wen@math.umd.edu.

To answer the question we first estimate only the unknown $\boldsymbol{\beta}_2$ using the shifted observations $\mathbf{y}^* = \mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1$ in the modified model referred to as Model I,

$$\text{Model I}: \qquad \mathbf{y}^* = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{2}$$

and then we estimate both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ using (1) rewritten as Model II,

$$\text{Model II}: \quad \mathbf{y} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

with $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ a full rank $n \times (p+q)$ matrix, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$.

2 Estimation of $\boldsymbol{\beta}_2$

It is helpful to define the matrices

$$\mathbf{A}_{11} = \mathbf{X}_1'\mathbf{X}_1, \quad \mathbf{A}_{12} = \mathbf{X}_1'\mathbf{X}_2, \quad \mathbf{A}_{21} = \mathbf{X}_2'\mathbf{X}_1, \quad \mathbf{A}_{22} = \mathbf{X}_2'\mathbf{X}_2$$

and the partitioned matrix $\mathbf{A} = \mathbf{X}'\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)'(\mathbf{X}_1, \mathbf{X}_2)$,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

An important role is played by the matrix,

$$\mathbf{A}_{11.2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

The assumption that $\mathbf{X}$ is of full rank implies that $\mathbf{A}$ is nonsingular and therefore ([1, p. 594])

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11.2}^{-1} & -\mathbf{A}_{11.2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11.2}^{-1} & \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11.2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1} \end{pmatrix}$$

2.1 Estimation: Model I

Since $\mathbf{X}_2$ has full rank, Model I gives a linear unbiased least square estimator of $\boldsymbol{\beta}_2$ in terms of $\mathbf{y}^*$,

$$\hat{\boldsymbol{\beta}}_2 = \mathbf{A}_{22}^{-1}\mathbf{X}_2'\mathbf{y}^* \tag{4}$$

with

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_2) = \sigma^2 \mathbf{A}_{22}^{-1} \tag{5}$$

and an unbiased estimate for σ^2 with $n - q$ degrees of freedom

$$\hat{\sigma^2} = \frac{\|\mathbf{y}^* - \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2\|^2}{n - q} \tag{6}$$

2.2 Estimation: Model II

Since $\mathbf{X}_1$ and $\mathbf{X}_2$ are full rank matrices, the least square estimates of both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \mathbf{A}^{-1} \begin{pmatrix} \mathbf{X}_1' \mathbf{y} \\ \mathbf{X}_2' \mathbf{y} \end{pmatrix} \tag{7}$$

Therefore,

$$\hat{\boldsymbol{\beta}}_1 = \mathbf{A}_{11.2}^{-1} \mathbf{X}_1' \mathbf{y} - \mathbf{A}_{11.2}^{-1} \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{X}_2' \mathbf{y} \tag{8}$$

$$\hat{\boldsymbol{\beta}}_2 = -\mathbf{A}_{22}^{-1} \mathbf{A}_{21} \mathbf{A}_{11.2}^{-1} \mathbf{X}_1' \mathbf{y} + \left[\mathbf{A}_{22}^{-1} \mathbf{A}_{21} \mathbf{A}_{11.2}^{-1} \mathbf{A}_{12} \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1} \right] \mathbf{X}_2' \mathbf{y} \tag{9}$$

This estimator is unbiased with covariance matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_2) = \sigma^2 \mathbf{A}_{22}^{-1} \mathbf{A}_{21} \mathbf{A}_{11.2}^{-1} \mathbf{A}_{12} \mathbf{A}_{22}^{-1} + \sigma^2 \mathbf{A}_{22}^{-1} \tag{10}$$

We observe that since $\mathbf{A}$ is positive definite, there is a multivariate normal distribution with covariance matrix $\mathbf{A}$ and therefore $\mathbf{A}_{11.2}$ is the covariance matrix of the corresponding conditional normal distribution. Therefore $\mathbf{A}_{22}^{-1} \mathbf{A}_{21} \mathbf{A}_{11.2}^{-1} \mathbf{A}_{12} \mathbf{A}_{22}^{-1}$ is a covariance matrix as well. It follows that the variances of the components of $\hat{\boldsymbol{\beta}}_2$ under Model II are larger than the corresponding variances under Model I, unless the columns of $\mathbf{X}_1$ are orthogonal to the columns of $\mathbf{X}_2$ (i.e. $\mathbf{A}_{12} = \mathbf{0}$) in which case (5) and (10) and also (4) and (9) are equal.

Thus, in general, estimating known regression coefficients in a model such as (1) results in a loss of efficiency of the least square estimates.

Under Model II, an unbiased estimate for σ^2 with $n-p-q$ degrees of freedom is

$$\hat{\sigma^2} = \frac{\|\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2\|^2}{n-p-q} \tag{11}$$

The loss of degrees of freedom may be the cause of efficiency loss when estimating σ^2 as we illustrate in the following example.

3 An Illustrations

3.1 Estimation of the Angle of a Parallelogram

The preceding discussion can be illustrated by a problem from [3, p. 66] regarding aerial observations of a parallelogram.

Consider a parallelogram with angles $\theta, \pi-\theta, \theta, \pi-\theta$, and suppose that we are given one noisy observation on each angle as follows: $y_1 = \theta + \epsilon_1$, $y_2 = \pi - \theta + \epsilon_2$, $y_3 = \theta + \epsilon_3$, and $y_4 = \pi - \theta + \epsilon_4$, where the ϵ_i have mean 0 and variance σ^2. This can be recorded more conveniently in vector notation akin to model (1) as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \pi + \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \theta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix} \tag{12}$$

Then Model I gives from (4)

$$\hat{\theta} = \mathbf{A}_{22}^{-1}\mathbf{X}_2'\mathbf{y}^* = \frac{1}{4}(1,-1,1,-1)\begin{pmatrix} y_1 \\ y_2 - \pi \\ y_3 \\ y_4 - \pi \end{pmatrix} = \frac{y_1 - y_2 + y_3 - y_4}{4} + \frac{\pi}{2} \tag{13}$$

and

$$\mathrm{Var}(\hat{\theta}) = \sigma^2\mathbf{A}_{22}^{-1} = \sigma^2/4 \tag{14}$$

On the other hand, Model II gives

$$\mathbf{A} = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}, \qquad \mathbf{A}^{-1} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}, \qquad \mathbf{A}_{11.2} = 1$$

so that from (9)

$$\hat{\theta} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11.2}^{-1}\mathbf{X}_1'\mathbf{y} + \left[\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11.2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\right]\mathbf{X}_2'\mathbf{y} = \frac{y_1 + y_3}{2} \quad (15)$$

with variance

$$\mathrm{Var}(\hat{\theta}) = \sigma^2\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11.2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} + \sigma^2\mathbf{A}_{22}^{-1} = \sigma^2/2 \quad (16)$$

Substituting $y_1 \approx y_3 \approx \theta$, $y_2 \approx y_4 \approx \pi - \theta$, we see that both estimators (13) and (15) are sensible, $(y_1 - y_2 + y_3 - y_4)/4 + \pi/2 \approx \theta$ and $(y_1 + y_3)/2 \approx \theta$. We refer to (13) as the *standard* estimator and to (15) as the *redundant* estimator. Since both estimators are unbiased, we conclude from (14) and (16) that the standard estimator (13) is twice as efficient as the redundant one. Interestingly, from (8) $\hat{\pi} = (y_1 + y_2 + y_3 + y_4)/2 \approx \pi$.

Figure 1 is a bar plot of mean square errors obtained by simulating Models I and II one million times using uniform, logistic, and normal errors with mean 0 and standard deviation of 10 degrees. The experimental results in the top part of Figure 1 verify closely the fact that the standard estimator is twice as efficient as the redundant estimator. The bottom part points to the loss of efficiency resulting from the estimation of σ^2 under Model II.

3.2 Estimation of the Angles of a Triangle

We close with a similar example concerning the estimation by least squares of the angles of a triangle. Viewed as a least squares problem with a linear restriction, this example is also discussed in [2, p. 59], but here, as in the previous example, we are concerned with the redundant least squares estimation of π and its consequence when estimating the angles of a triangle.

A surveyor measures once each of the angles α, β, γ of an area that has the shape of a triangle, and obtains unbiased measurements Y_1, Y_2, Y_3 (in radians). It is known that $\mathrm{Var}(\mathrm{Y_i}) = \sigma^2$, $i = 1, 2, 3$. Then Model I reduces to

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 - \pi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

and the least squares estimates of the unknown angles are

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 2Y_1 - Y_2 - Y_3 + \pi \\ 2Y_2 - Y_1 - Y_3 + \pi \end{pmatrix}$$

with covariance matrix

$$\mathrm{Var} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{\sigma^2}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

On the other hand, Model II reduces to

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \pi \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

This time the least squares estimates are different yet sensible,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\pi} \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_1 + Y_2 + Y_3 \end{pmatrix}$$

and

$$\mathrm{Var} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\pi} \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 3 \end{pmatrix}$$

It follows that if π is estimated then $\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}(\hat{\beta}) = \sigma^2$, whereas if π is not estimated the estimates are more precise since $\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}(\hat{\beta}) = 2\sigma^2/3$.

4 summary

We have discussed the loss of efficiency resulting from the estimation of known coefficients in linear regression. Treating the known coefficients as unknown parameters alters the design matrix and produces inefficient, albeit unbiased, least square estimates. This of course is unavoidable when over-fitting is not recognized in time. In the examples, the loss of efficiency is coupled with a loss of information in the sense that Model I gives estimators that use all the observations in the estimation of each unknown angle, which is not the case in Model II where only partial information is used.

Acknowledgment: We wish to thank Victor De Oliveira and Paul Smith for very useful comments.

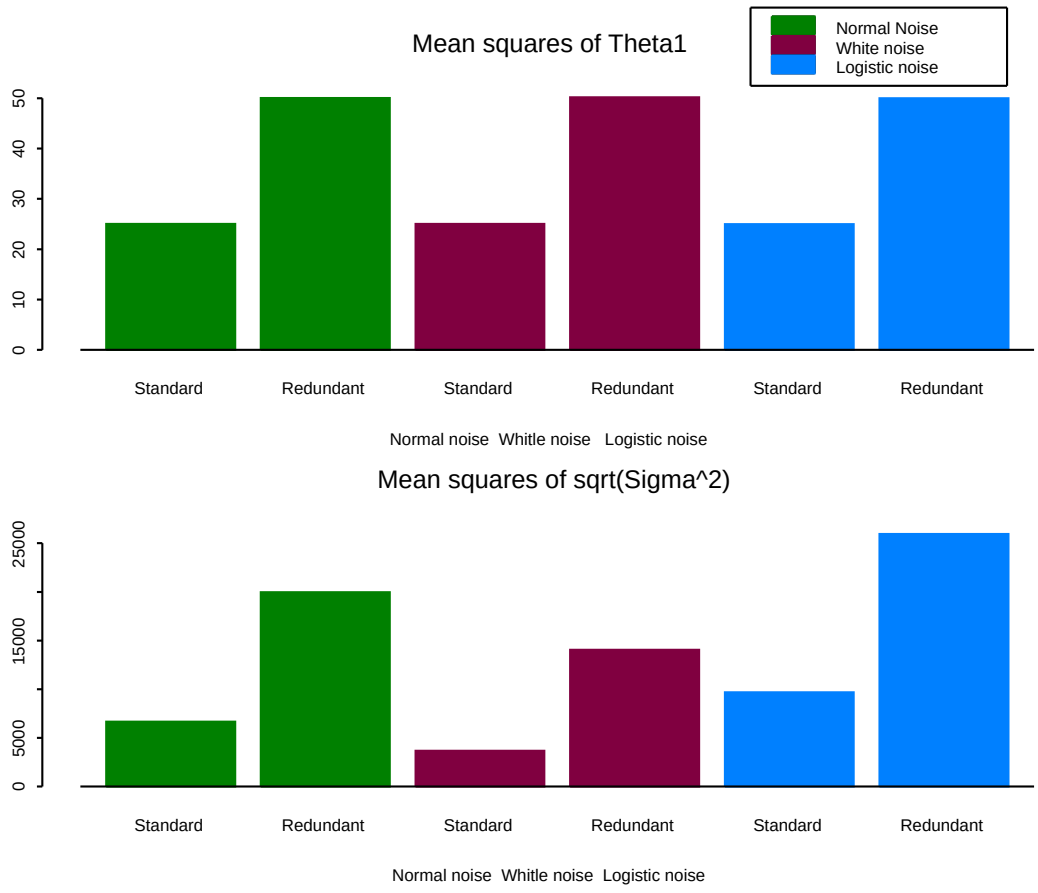

Figure 1: Mean square error results obtained by a simulation.

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis* 2nd ed., Wiley, New York.

[2] Seber, G.A.F., and Lee, A.J. (2003). *Linear regression analysis*, 2nd ed., Wiley, Hoboken, N.J.

[3] Silvey, S.D. (1975). *Statistical Inference*, Chapman-Hall, London.