Toutenburg, Srivastava, Schaffrin:

Efficiency properties of weighted mixed regression estimation

Sonderforschungsbereich 386, Paper 122 (1998)

Efficiency properties of weighted mixed regression estimation

H. Toutenburg* V. K. Srivastava** B. Schaffrin††

August 14, 1998

Abstract

This paper considers the estimation of the coefficient vector in a linear regression model subject to a set of stochastic linear restrictions binding the regression coefficients, and presents the method of weighted mixed regression estimation which permits to assign possibly unequal weights to the prior information in relation to the sample information. Efficiency properties of this estimation procedure are analyzed when disturbances are not necessarily normally distributed.

1 Introduction

When a set of stochastic linear constraints binding the regression coefficients in a linear regression model is available, Theil and Goldberger (1961) have proposed the method of mixed regression estimation; see Srivastava (1980) for an annotated bibliography. Their method typically assumes that the prior information in the form of stochastic linear constraints and the sample information in the form of observations on the study variable and explanatory variables are equally important and therefore receive equal weights in the estimation procedure. In practice, situations may occur where this assumption may not be tenable. For example, one may conduct a statistical test for the compatibility of sample and prior information; see (Theil, 1963) for instance. If the statistical test reveals that they are compatible, we may combine the two kinds of information assigning equal weights and use the method of mixed estimation accordingly. On the other hand, if the statistical test is indicative of incompatibility, the conventional procedure is to ignore the prior information. This strategy of discarding the prior information outrightly is rather unappealing in comparison to the one which assigns unequal weights to the prior information in comparison to the sample information. Some extraneous considerations may often be suggestive of giving unequal weights. In such circumstances, it may be imperative to assign not necessarily equal weights during the process of combining the prior and sample information. Appreciating this viewpoint, Schaffrin and Toutenburg (1990) have developed the method of weighted mixed regression

*Institut für Statistik, Universität München, Akademiestr. 1, 80799 München, Germany
**Department of Statistics, Lucknow University, Lucknow 226007, India
††Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University, Columbus, OH 43210-1275, U.S.A

estimation. Such a method offers considerable flexibility in the sense that one can assign possibly different weights to sample information and prior information depending upon the degree of belief. Besides this, the method provides a kind of unified treatment to traditional pure and mixed regression methods.

The purpose of this article is to analyze the efficiency properties of the weighted mixed regression method. Section 2 describes the model and the method of weighted mixed regression estimation proposed by Schaffrin and Toutenburg (1990). A feasible version of it is developed when the disturbance variance is not known. In Section 3, we discuss the efficiency properties when disturbances are small but not necessarily normally distributed. The results related to both bias vector and mean squared error matrix are derived in the Appendix. Finally, some remarks are offered in Section 4.

2 Model Specification and Some Estimators

Let us postulate the following linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1}$$

where $\mathbf{y}$ is a $T \times 1$ vector of T observations on the study variable, $\mathbf{X}$ is a $T \times p$ full column rank matrix of T observations on p explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients associated with them and $\mathbf{u}$ is a $T \times 1$ vector of disturbances assumed to be distributed with mean vector $\mathbf{0}$ and variance covariance matrix $\sigma^2\mathbf{I}$ with σ^2 as unknown quantity and $\mathbf{I}$ as identity matrix of order $T \times T$.

Additionally, it is supposed that a set of stochastic linear restrictions binding the regression coefficients is available in the form of independent prior information:

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \mathbf{v} \tag{2}$$

where $\mathbf{r}$ is a $J \times 1$ vector of known elements, $\mathbf{R}$ is a $J \times p$ full row rank matrix with known elements and $\mathbf{v}$ is a $J \times 1$ vector of stochastic elements assumed to be distributed with mean vector $\mathbf{0}$ and variance covariance matrix $\boldsymbol{\Psi}$ with known elements. Further, it is assumed that the elements of $\mathbf{v}$ are stochastically independent of the elements of $\mathbf{u}$.

When the sample information given by (1) and the prior information depicted by (2) are to be assigned not necessarily equal weights on the basis of some extraneous considerations in the estimation of regression parameters, Schaffrin and Toutenburg (1990) have proposed the method of weighted mixed regression estimation. Their model specification is slightly different as they assume that the variance covariance matrix of $\mathbf{u}$ is fully known while we have assumed that it is not so. However, we follow their technique which essentially comprises the choice of $\boldsymbol{\beta}$ such that the sum of squares

$$\frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + w(\mathbf{r} - \mathbf{R}\boldsymbol{\beta})'\boldsymbol{\Psi}^{-1}(\mathbf{r} - \mathbf{R}\boldsymbol{\beta}) \tag{3}$$

is minimum where w is a nonstochastic and non-negative scalar. This leads to the following solution for $\boldsymbol{\beta}$:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + w\sigma^2\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R})^{-1}(\mathbf{X}'\mathbf{y} + w\sigma^2\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{r})\,. \tag{4}$$

As σ^2 is assumed to be unknown, we propose to replace it by its unbiased estimator based on the residual sum of squares without the constraints (2), namely

$$s^2 = \left(\frac{1}{T-p}\right) \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} \tag{5}$$

so that the weighted mixed regression estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}}_w = (\mathbf{X}'\mathbf{X} + ws^2\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R})^{-1}(\mathbf{X}'\mathbf{y} + ws^2\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{r}) \tag{6}$$

which resembles the f-class of mixed regression estimators proposed by Theil (1963).

If we put $w = 0$, the estimator (6) reduces to

$$\widehat{\boldsymbol{\beta}}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{7}$$

which is the traditional least squares estimator, for it gives no weight to the available stochastic linear restrictions.

If we substitute $w = 1$, we get

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}'\mathbf{X} + s^2\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R})^{-1}(\mathbf{X}'\mathbf{y} + s^2\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{r}) \tag{8}$$

which is the mixed regression estimator proposed by Theil and Goldberger (1961). This estimator gives equal weight to sample and prior information.

It is thus seen that a value of w between 0 and 1 specifies an estimator in which the prior information receives less weight in comparison to the sample information. On the other hand, a value of w greater than 1 implies higher weight to the prior information which, of course, may be of little practical interest. Thus the weighted mixed regression estimator provides a useful framework in which it is possible to incorporate unequal importance to the sample information in relation to the prior information depending upon, for instance, the degree of belief.

3 Efficiency Properties

In order to study the efficiency properties of the weighted mixed regression estimator (6), we simply suppose that the elements of the disturbance vector $\mathbf{u}$ have finite moments up to order four; no specific distribution is assumed as such. Let $\sigma^3\gamma_1$, and $\sigma^4(\gamma_2 + 3)$ be the third and fourth moments of the independently and identically distributed elements of $\mathbf{u}$ with mean 0 and variance σ^2. Thus γ_1 and γ_2 respectively measure the excess of skewness and kurtosis of the distribution. Notice that both γ_1 and γ_2 are zero for the normal distribution.

Under the above fairly general specification, no exact results related to bias vector and mean squared error matrix of the weighted mixed regression estimator can be derived. Although, under the simplified case of normal distribution, the exact expressions can be obtained but they would be sufficiently intricate and would not permit the deduction of meaningful inferences; see, e. g., Swamy and Mehta (1969). We have therefore chosen to employ the small disturbance asymptotic theory in preference to the large sample asymptotic theory as indicated by Srivastava and Upadhyaha (1975). This makes sense since the new sample data are typically collected using the best technology presently available whereas the prior information may stem from obscure sources. We summarize the following results which are derived in the Appendix.

Theorem 1 *When the disturbances are small, the bias vector and the mean squared error matrix of the weighted mixed regression estimator, retaining terms up to order 0(σ^4), are given by*

$$\begin{aligned} \mathrm{bias}(\widehat{\boldsymbol{\beta}}_w) &= \mathrm{E}(\widehat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}) \\ &= -\sigma^3 \left(\frac{w\gamma_1}{T-p}\right) (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}*\mathbf{M})\mathbf{1} \end{aligned} \tag{9}$$

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_w) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^4 w(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{C} + \left(\frac{\gamma_2}{T-p}\right)\mathbf{N}](\mathbf{X}'\mathbf{X})^{-1} \tag{10}$$

where

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{11}$$

$$\mathbf{C} = [2 - w(1 + \frac{2}{T-p})]\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R} \tag{12}$$

$$\begin{aligned} \mathbf{N} &= \mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}*\mathbf{M})\mathbf{X} + \mathbf{X}'(\mathbf{I}*\mathbf{M})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R} \\ &\quad - \frac{w\,\mathrm{tr}\,\mathbf{M}(\mathbf{I}*\mathbf{M})}{T-p}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R} \end{aligned} \tag{13}$$

with $*$ denoting the elementwise Hadamard product operator of matrices and **1** *denoting a column vector having all elements unity.*

It is observed from (9) that the skewness, and not the kurtosis, of the distribution of the disturbances influences the bias at least to the order of our approximation. This bias vanishes when the distribution is symmetric. The bias also vanishes when $\mathbf{X}'(\mathbf{I}*\mathbf{M})\mathbf{1}$ is equal to a null vector, i.e., when

$$\sum_{t=1}^{T} x_{tj} m_{tt} = 0 \quad (j = 1, \ldots, p) \tag{14}$$

where m_{tt} is the tth diagonal element of the matrix $\mathbf{M}$.

Looking at the expression (10), it is seen that the variability of the estimator as measured by the mean squared error matrix to the order of our approximation is influenced only by the kurtosis of the distribution of the disturbances; the skewness has no role at least to the order of our approximation.

Comparing the weighted mixed regression estimator with the pure regression estimator which ignores the prior information completely, we observe from (10) that

$$V(\widehat{\boldsymbol{\beta}}_0) - \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_w) = \sigma^4 w(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{C} + \left(\frac{\gamma_2}{T-p}\right)\mathbf{N}](\mathbf{X}'\mathbf{X})^{-1}\,. \tag{15}$$

If $\underline{m}$ and $\overline{m}$ denote the smallest and largest diagonal elements of the idempotent matrix $\mathbf{M}$, we have in Löwnerïs partial ordering of matrices

$$\underline{m}\mathbf{I} < (\mathbf{I}*\mathbf{M}) < \overline{m}\mathbf{I}\,. \tag{16}$$

Premultiplying by $\mathbf{M}$ and then taking the trace, we get

$$\underline{m} < \frac{\operatorname{tr} \mathbf{M}(\mathbf{I} * \mathbf{M})}{T-p} < \overline{m}\,. \tag{17}$$

It therefore follows from (13), (16) and (17) that

$$(2\underline{m} - w\overline{m})\mathbf{R}'\mathbf{\Psi}^{-1}\mathbf{R} < \mathbf{N} < (2\overline{m} - w\underline{m})\mathbf{R}'\mathbf{\Psi}^{-1}\mathbf{R}\,. \tag{18}$$

Utilizing this result, we find that the matrix difference (15) is at least a positive semidefinite matrix when

$$0 < w < 2\left[\frac{T-p+\gamma_2\underline{m}}{T-p+2+\gamma_2\overline{m}}\right] \quad \text{for} \quad \gamma_2 > 0 \tag{19}$$

$$0 < w < 2\left[\frac{T-p+\gamma_2\overline{m}}{T-p+2+\gamma_2\underline{m}}\right] \quad \text{for} \quad \gamma_2 < 0 \tag{20}$$

$$0 < w < 2\left[\frac{T-p}{T-p+2}\right] \quad \text{for} \quad \gamma_2 = 0\,. \tag{21}$$

As $\mathbf{M}$ is an idempotent matrix, its smallest and largest characteristic roots are 0 and 1 repectively. Thus we have

$$\mathbf{0} \leq \mathbf{M} \leq \mathbf{I}$$

and since $\mathbf{M}$ is also symmetric it follows that

$$0 \leq \underline{m} < \overline{m} \leq 1\,. \tag{22}$$

Using this, we see that the condition (19) will be satisfied as long as

$$0 < w < 2\left[\frac{T-p}{T-p+2+\gamma_2}\right] \quad \text{for} \quad \gamma_2 > 0\,. \tag{23}$$

Similarly, the condition (20) will hold true as long as

$$0 < w < 2\left[\frac{T-p+\gamma_2}{T-p+2}\right] \quad \text{for} \quad \gamma_2 < 0\,. \tag{24}$$

It is thus seen that the weighted mixed regression estimator dominates the pure regression estimator according to the criterion of mean squared error matrix to the order $O(\sigma^4)$ under the constraint (19) or (23) for platykurtic distributions of the disturbances, the constraint (20) or (24) for leptokurtic distributions and the constraint (21) for mesokurtic distributions such as the normal distribution, in particular.

Next, let us compare the weighted mixed regression estimator with the conventional or unweighted mixed regression estimator which gives equal weight to sample and prior information ($w = 1$).

From (10), we observe that

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_1) - \mathrm{MSE}(\widehat{\boldsymbol{\beta}}_w) = \sigma^4(1-w)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}_w(\mathbf{X}'\mathbf{X})^{-1} \tag{25}$$

where

$$\mathbf{D}_w = -\left(\frac{\gamma_2}{T-p}\right)[\mathbf{R}'\mathbf{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}*\mathbf{M})\mathbf{X} + \mathbf{X}'(\mathbf{I}*\mathbf{M})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\mathbf{\Psi}^{-1}\mathbf{R}] + [(1+w)\left(1+\frac{2}{T-p}+\frac{\gamma_2 \operatorname{tr}\mathbf{M}(\mathbf{I}*\mathbf{M})}{(T-p)^2}\right) - 2]\mathbf{R}'\mathbf{\Psi}^{-1}\mathbf{R}\,. \quad (26)$$

Assuming that $0 < w < 1$ and using (16) and (17), it is easy to see that the matrix difference (25) is at least positive semidefinite as long as

$$\left.\begin{array}{l} [1-2\left(\frac{2-\gamma_2(\overline{m}-\underline{m})}{T-p+2+\gamma_2\underline{m}}\right)] < w < 1 \\ (T-p) > 2-\gamma_2(2\overline{m}-\underline{m}) \end{array}\right\} \quad \text{for} \quad \gamma_2 > 0 \quad (27)$$

$$\left.\begin{array}{l} [1-2\left(\frac{2+\gamma_2(\overline{m}-\underline{m})}{T-p+2+\gamma_2\overline{m}}\right)] < w < 1 \\ (T-p) > 2+\gamma_2(\overline{m}-2\underline{m}) \end{array}\right\} \quad \text{for} \quad \gamma_2 < 0 \quad (28)$$

$$\left.\begin{array}{l} \left(\frac{T-p-2}{T-p+2}\right) < w < 1 \\ (T-p) > 2 \end{array}\right\} \quad \text{for} \quad \gamma_2 = 0\,. \quad (29)$$

Further, utilizing (22), the condition will hold true as long as

$$\left.\begin{array}{l} \left(\frac{T-p-2(1-\gamma_2)}{T-p+2}\right) < w < 1 \\ (T-p) > 2(1-\gamma_2) \end{array}\right\} \quad \text{for} \quad \gamma_2 > 0\,. \quad (30)$$

Similarly, (28) will be satisfied when

$$\left.\begin{array}{l} \left(\frac{T-p-2-\gamma_2}{T-p+2+\gamma_2}\right) < w < 1 \\ (T-p) > (2+\gamma_2) \end{array}\right\} \quad \text{for} \quad \gamma_2 < 0\,. \quad (31)$$

Thus the conditions (27) or (30) for platykurtic distributions, (28) or (31) for leptokurtic distributions, and (29) for mesokurtic including the normal distributions specify the cases where providing less weight to the prior information in comparison to the sample information for the estimation of regression coefficients would be a better strategy than assigning equal weight.

Next, let us consider the relatively less interesting case in which w exceeds 1, i.e., the prior information is given higher weight than the sample information. Such a situation may arise, for instance, when the prior information is known to have high credibility.

Assuming that $w > 1$, it is observed from (25) that the weighted mixed regression estimator is no more efficient than the mixed regression estimator with respect to the criterion of mean squared error matrix to the order $O(\sigma^4)$ unless $\mathbf{D}_w$ turns out to be nonpositive-definite. However, if we compare the weighted mixed regression estimator with the pure regression estimator, it follows from (15) that incorporating the dominant prior information via weighted mixed regression is a better strategy than ignoring it alltogether when the condition (19) or (23) holds for platykurtic distributions, condition (20) or (24) holds for leptokurtic distributions, and condition (21) holds for mesokurtic including the normal distributions provided that the upper bound of w as specified by the respective condition is larger than 1.

4 Some Remarks

Our investigations have brought out some interesting properties of the method of weighted mixed regression estimation for the coefficients in a linear regression model when a set of stochastic linear constraints is available to represent the prior information. For example, it is observed that the weighted mixed regression estimator is nearly unbiased when the distribution of the disturbances is symmetric irrespective of the nature of kurtosis. If the distribution is skewed, the estimator is generally biased. And the magnitude of the bias, to the order of our approximation, will always be smaller than that of the mixed regression estimator which gives equal weights to prior and sample information provided that w is less than one. Of course, the pure regression estimator is exactly unbiased but it ignores the prior information completely.

Comparing the estimators with respect to the criterion of mean squared error matrix to the order of our approximation, we have spelled out the conditions for the dominance of the weighted mixed regression estimator over both the pure regression estimator and the mixed regression estimator. An interesting aspect of these conditions is that they are easy to check in any given application provided that the kurtosis of the distribution of the disturbances is available.

Looking at the expressions for the bias vector and the mean squared error matrix to the order of our approximation, it is interesting to note that the bias is influenced by the skewness of the distribution, and not the kurtosis, while the mean squared error is affected by the kurtosis alone and the skewness of the distribution has no role. An interesting implication of it is that any conclusions drawn under the conventional specification of normality of disturbances could be quite different from those where the distribution is skewed, platykurtic and leptokurtic.

At this point we are reluctant to draw more specific conclusions beyond the above as far as the the "best" choice of w (given γ_2) is concerned. The main reason is that we here compare the mean squared error matrices themselves, not just scalar-valued functions of them (like the "trace", "determinant", or "largest eigenvalue"). Moreover, we refer to the geodetic literature, e.g., Schaffrin and Bock (1994), for questions of practical relevance. It is this context in which we would like to present a simulation study in the future which would tell us a bit more about the appropriateness of our small error assumption and the validity of our approximations based upon it. This is, however, beyond the scope of the present paper.

Acknowledgement

The third author would like to thank Helge Toutenburg for his hospitality during his stay at the Ludwig-Maximilians University in Munich as visiting research fellow in the framework of the Sonderforschungsbereich 386 : "Statistische Analyse diskreter Strukturen, Modellierung und Anwendung in Biometrie und Ökonometrie". This and the financial support through the German Science Foundation (DFG) is most gratefully acknowledged.

References

Schaffrin, B. (1983). Variance-covariance component estimation for the adjustment of heterogeneous repeated measurements (in german), *Publication of the german geodetic commission*, C-282, Munich, Germany.

Schaffrin, B. and Bock, Y. (1994). Geodetic deformation analysis based on robust inverse theory, *Manus. Geodaet.* **19**: 31–44.

Schaffrin, B. and Toutenburg, H. (1990). Weighted mixed regression, *Zeitschrift für Angewandte Mathematik und Mechanik* **70**: 735–738.

Srivastava, V. K. (1980). Estimation of linear single-equation and simultaneous-equation models under stochastic linear constraints: An annotated bibliography, *International Statistical Review* **48**: 79–82.

Srivastava, V. K. and Upadhyaha, S. (1975). Small-disturbance and large sample approximations in mixed regression estimation, *Eastern Economic Journal* **2**: 261–265.

Swamy, P. A. V. B. and Mehta, J. S. (1969). On Theil's mixed regression estimator, *Journal of the American Statistical Association* **64**: 273–276.

Theil, H. (1963). On the use of incomplete prior information in regression analysis, *Journal of the American Statistical Association* **58**: 401–414.

Theil, H. and Goldberger, A. S. (1961). On pure and mixed estimation in econometrics, *International Economic Review* **2**: 65–78.

Ullah, A., Srivastava, V. K. and Chandra, R. (1983). Properties of shrinkage estimators in linear regression when disturbances are not normal, *Journal of Econometrics* **21**: 389–402.

Appendix

In order to derive the small disturbance asymptotic approximations for the bias vector and the mean squared error matrix of the weighted mixed regression estimator, we replace $\mathbf{u}$ in (1) by $\sigma\mathbf{Z}$ so that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{Z}\,. \tag{32}$$

Now the elements of $\mathbf{Z}$ are independently and identically distributed with the first four moments as $0, 1, \gamma_1$ and $(\gamma_2 + 3)$. Thus we have

$$\begin{aligned}
\mathrm{E}(\mathbf{Z}) &= \mathbf{0} && (33)\\
\mathrm{E}(\mathbf{Z}\mathbf{Z}') &= \mathbf{I}_T && (34)\\
\mathrm{E}(\mathbf{Z}\mathbf{Z}'\mathbf{M}\mathbf{Z}) &= \gamma_1(\mathbf{I} * \mathbf{M})\mathbf{1} && (35)\\
\mathrm{E}(\mathbf{Z}\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{Z}') &= \gamma_2(\mathbf{I} * \mathbf{M}) + (T-p)\mathbf{I} + 2\mathbf{M} && (36)
\end{aligned}$$

where $*$ denotes the elementwise Hadamard product operator of matrices and $\mathbf{1}$ is a $T \times 1$ vector with all elements equal to 1; see Ullah, Srivastava and Chandra (1983) for the derivation, or alternatively Schaffrin (1983).

Substituting (32) and (2) in (6) and expanding in terms of increasing powers of σ, we find

$$\begin{aligned}
(\widehat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}) &= [\mathbf{X}'\mathbf{X} + \sigma^2 w \left(\frac{\mathbf{Z}'\mathbf{M}\mathbf{Z}}{T-p}\right) \mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}]^{-1}\\
&\quad \times[\sigma\mathbf{X}'\mathbf{Z} + \sigma^2 w \left(\frac{\mathbf{Z}'\mathbf{M}\mathbf{Z}}{T-p}\right) \mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{v}]\\
&= \sigma f_1 + \sigma^2 f_2 + \sigma^3 f_3 + +\sigma^4 f_4 + O_p(\sigma^5) && (37)
\end{aligned}$$

where

$$\begin{aligned}
f_1 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} && (38)\\
f_2 &= w\left(\frac{\mathbf{Z}'\mathbf{M}\mathbf{Z}}{T-p}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{v} && (39)\\
f_3 &= -w\left(\frac{\mathbf{Z}'\mathbf{M}\mathbf{Z}}{T-p}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} && (40)\\
f_4 &= -w^2\left(\frac{\mathbf{Z}'\mathbf{M}\mathbf{Z}}{T-p}\right)^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{v}\,. && (41)
\end{aligned}$$

Now utilizing the results (33)–(36) it is easy to see that

$$\begin{aligned}
\mathrm{E}(f_1) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\mathbf{Z})\\
&= \mathbf{0}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(f_2) &= \left(\frac{w}{T-p}\right)\mathrm{E}(\mathbf{Z}'\mathbf{M}\mathbf{Z})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\,\mathrm{E}(\mathbf{v})\\
&= \mathbf{0}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(f_3) &= -\left(\frac{w}{T-p}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\mathbf{Z}\mathbf{Z}'\mathbf{M}\mathbf{Z})\\
&= -\left(\frac{w\gamma_1}{T-p}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}*\mathbf{M})\mathbf{1}\\
\mathrm{E}(f_4) &= -\left(\frac{w}{T-p}\right)^2 \mathrm{E}((\mathbf{Z}'\mathbf{M}\mathbf{Z})^2)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\boldsymbol{\Psi}^{-1}\,\mathrm{E}(\mathbf{v})\\
&= \mathbf{0}\,.
\end{aligned}$$

Employing these results in

$$\mathrm{bias}(\widehat{\boldsymbol{\beta}}_w) = \sigma\,\mathrm{E}(f_1) + \sigma^2\,\mathrm{E}(f_2) + \sigma^3\,\mathrm{E}(f_3) + \sigma^4\,\mathrm{E}(f_4)\,, \tag{42}$$

we obtain the expression (9) stated in Theorem 1.

Similarly, by virtue of the distributional properties of $\mathbf{Z}$ and $\mathbf{v}$, it is easy to see that

$$\begin{aligned}
\mathrm{E}(f_1 f_1') &= (\mathbf{X}'\mathbf{X})^{-1}\\
\mathrm{E}(f_2 f_1') &= \mathbf{0}\\
\mathrm{E}(f_3 f_1') &= -\left(\frac{w\gamma_2}{T-p}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}*\mathbf{M})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\\
&\quad -w(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\\
\mathrm{E}(f_2 f_2') &= \left(\frac{w^2}{T-p}\right)\left[T-p+2+\left(\frac{\gamma_2}{T-p}\right)\mathrm{tr}\,\mathbf{M}(\mathbf{I}*\mathbf{M})\right]\times\\
&\quad \times(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\Psi}^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\,.
\end{aligned}$$

Using these expressions and observing that the mean squared error matrix to the order $O(\sigma^4)$ is given by

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_w) = \sigma^2\,\mathrm{E}(f_1 f_1') + \sigma^3\,\mathrm{E}(f_2 f_1' + f_1 f_2') + \sigma^4\,\mathrm{E}(f_3 f_1' + f_1 f_3' + f_2 f_2')\,, \tag{43}$$

we obtain the result (10) of Theorem 1.