Estimators of the Mean Squared Error of Prediction in Linear Regression

O. Bunke and B. Droge
Department of Statistics
Humboldt-University
Berlin, German Democratic Republic

If a linear regression model is used for prediction, the mean squared error of prediction (MSEP) measures the performance of the model. The MSEP is a function of unknown parameters and good estimates of it are of interest. This article derives a best unbiased estimator and a minimum MSE estimator under the assumption of a normal distribution. It compares the bias and the MSE of these estimators and some others. Similar results are presented for the case in which the model is used to estimate values of the response function.

KEY WORDS: Regression model; Response function; Selection of variables.

1. INTRODUCTION

A frequent situation in analyzing experimental data involves (possibly replicated) observations of an output variable \( y \) obtained for corresponding fixed values of input variables \( x_1, x_2, \ldots, x_k \). The output variable may be the yield in a chemical plant working under technological, chemical, and physical conditions described by input variables such as temperature, \( x_1 \), pressure, \( x_2 \), and the amounts, \( x_3, x_4, x_5 \), of different substances and catalysts in the chemical process. An objective of the analysis may be to predict future values of the output \( y \) for fixed values of the input variables using a function or model depending on the input.

In other cases, the model is used for a rough description of the dependence between output and input variables by an equation, say \( y = f(x_2, x_3, x_5) \). Then the model \( f \) may be interpreted as the approximation to the true response function \( \eta \), which describes the real relation \( y = \eta(x_1, \ldots, x_k) + \varepsilon \) between the output and the input variables. The term \( \varepsilon \) is interpreted here as the random error describing the influence of unobserved variables.

Usually the model will depend on parameters that must be estimated on the basis of data, that is, the model is fitted to the data. The mean squared error of prediction (MSEP) gives a simple description for the performance of the model. The MSEP depends on the unknown response function and the unknown variance of the observations, so good estimates of the MSEP are of interest. Such estimates may be useful as criteria for the comparison of subsets of input variables or of models, although, with the exception of some cases, the selection of models and of corresponding input variables is a more complex process than merely a formal comparison of models by a criterion. But such a criterion is a helpful tool in a model selection strategy (see Bunke 1983), which should also include subject matter considerations and common sense judgment. This has been discussed in the literature, for example, in Kennedy and Bancroft (1971) or Mallows (1973). Surveys on model selection methods, including model comparison criteria, are presented in Gaver and Geisel (1974), Hocking (1976), Thompson (1978), and Montgomery and Peck (1982). Some of these criteria, like \( C_p \) or PRESS, may also be considered as estimates of the MSEP.

Recently Efron (1983), Efron and Gong (1982), and Gong (1982) investigated the behavior of bootstrap and other estimators of the MSEP under somewhat different assumptions, namely of random input variables—an assumption that is also contained in Freedman (1981).

Our article derives a best unbiased estimator and a minimum MSE estimator under the assumption of a normal distribution and compares the bias and the MSEP of these estimators with those of some others. In Section 2 we introduce the problem, assuming replicated observations, and in Section 3 we introduce the different criteria. Their comparison is carried out in Section 4, while in Section 5 numerical results and graphical presentations of MSE for some examples illustrate the quality of different estimators. A discussion of the results is presented in Section 6. In Section 7 we give a short report on analogous results.
for the case of the objective being an estimation of the response function. Estimation of MSEP and MSE in the case of no replications is treated in Section 8. Mathematical proofs are given in the Appendix.

2. THE MEAN SQUARED ERROR OF PREDICTION

Suppose that we have \( n = \sum_j n_j \) observations \( y_{ij} \), assumed to be independent random variables and to have normal distributions

\[
y_{ij} \sim N(\mu_{ij}, \sigma^2)
\]

for \( i = 1, \ldots, m \) and \( j = 1, \ldots, n_j \). The vector of expectations \( \mu = (\mu_1, \ldots, \mu_m)^T \) and the variance \( \sigma^2 \) are unknown. For each \( i \), the values \( y_{1i}, \ldots, y_{ni} \) are replicated observations of the output variable for the same fixed joint values \( x_{1i}, \ldots, x_{ki} \) of the input variables.

We will assume that \( n = \sum n_j > m \); that is, we exclude the case of no replications \( (n_1 = \cdots = n_m = 1) \). Replications are necessary for estimating the variance \( \sigma^2 \) if there is no prior information on the response function or on the expectation \( \mu \).

The problem then is to predict some or all of the random variables \( y_{i1}^*, \ldots, y_{im}^* \) with

\[
E(y_{i*}) = \mu_{i*}, \quad \text{var} (y_{i*}) = \sigma^2,
\]

using the observations \( y_{ij} \). We assume these random variables \( y_{i1}^*, \ldots, y_{im}^* \) to be independent. Often, a pseudolinear regression model \( f(x_1, \ldots, x_k) = \sum_{h=1}^{p} \beta_h g_h(x_1, \ldots, x_k) \), given by

\[
f(x_1, \ldots, x_k) = \sum_{h=1}^{p} \beta_h g_h(x_1, \ldots, x_k),
\]

with fixed real functions \( g_h \) and parameters \( \beta_h \), is used as a basis for prediction.

If model (3) is fitted to the data by ordinary least squares, we obtain the least squares estimate \( b \) for the parameter vector \( \beta = (\beta_1, \ldots, \beta_p)^T \):

\[
b = (b_1, \ldots, b_p)^T = (A^T L_1^{-1} A)^{-1} A^T L_1^{-1} \tilde{y},
\]

where

\[
\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_m)^T, \quad \tilde{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij},
\]

\[
L = \text{diag} \left[ 1/n_1, \ldots, 1/n_m \right],
\]

and \( A \) is the \( m \times p \) matrix with the elements

\[
a_{hk} = g_h(x_{1k}, \ldots, x_{kk}).
\]

In the following, we assume that \( p = \text{rank} (A) \leq m \). The value \( y_{i*} \) may then be predicted by \( \hat{y}_{i*} = \sum_h \beta_h g_h(x_{1i}, \ldots, x_{ki}) \) if the model (3) is used.

The MSEP is defined as

\[
r = \text{MSEP} (\hat{y}^*) = E \sum_i w_i | \hat{y}_{i*} - y_{i*} |^2,
\]

as it was proposed, for example, by Mallows (1973) or Beaudel and Afifi (1977). The possibly different importance of the prediction errors \( \hat{y}_{i*} - y_{i*} \) is taken into account by the introduction of weights \( w_i \geq 0 \) with \( \sum w_i = 1 \). The weight \( w_i \) could also characterize the future frequency of the occurring of a variable \( y_{i*} \) with (2) to be predicted.

As an example, the observations \( y_{11}, \ldots, y_{14} \) could be the yield of a chemical plant on 14 different days on which the chemical plant was run under the same fixed technological conditions as given by \( x_{11}, \ldots, x_{14} \). The expected yield would be \( \mu_i = \eta(x_{11}, \ldots, x_{14}) \). If we are interested in predicting the yields on several future days under technological conditions that are identical (if the response function \( \eta \) is sufficiently smooth, possibly only close) to those of some of the past days, we may adopt the above mathematical formalization of the problem.

The introduction of weights is necessary to cover cases in which the future output is to be predicted only for some of the \( m \) joint input values already observed in the past, for instance, in our example the yield \( y_{14}, \ldots, y_{14} \) under 14 different technological conditions given by the joint values \( (x_{1i}, \ldots, x_{14}) \) \( (i = 1, \ldots, 14) \). If these technological conditions are considered to be equally important, we will choose the weights \( w_1 = \cdots = w_{14} = 1/14 \); \( w_1 = \cdots = w_m = 0 \). One should choose \( w_i = 0 \) if \( y_{i*} \) is not to be predicted.

Predicting the output for values of the input that are not close to the observed input values may lead to large prediction errors if, as we are assuming in this article, the form of the response function is unknown and if the model (3) used for prediction is possibly not exact. Discussions on the dangers of extrapolating are given in Weisberg (1980) and Montgomery and Peck (1982), for example.

Now we use a vector notation to simplify formulas:

\[
y^* = (y_{1*}, \ldots, y_{m*})^T, \quad \tilde{y}^* = (\tilde{y}_{1*}, \ldots, \tilde{y}_{m*})^T
\]

\[
\| y^* \|^2 = \sum_i w_i y_i^2.
\]

It is useful to write the MSEP as

\[
r = \Delta + \Lambda + \sigma^2,
\]

where

\[
\Delta = \| \mu - E\tilde{y}^* \|^2
\]

is the bias term, and the estimation error term is

\[
\Lambda = E\| \tilde{y}^* - E\tilde{y}^* \|^2.
\]

If the model (3) is not exact, a bias \( \Delta \) cannot be avoided for the "ordinary least squares predictor" \( \tilde{y}^* \) and also for any other predictor of the form \( \tilde{y} = A\hat{\beta} \), where \( \hat{\beta} \) depends on the observations. Therefore, it seems appropriate to use a predictor of \( y^* \) with minimal bias \( \Delta \) and minimal MSEP instead of \( \tilde{y}^* = A\hat{\beta} \), a problem discussed in more detail in Karson, Manson, and Hader (1969), Bunke and Striiby (1975), and
The predictor $\hat{y}$ of $y^*$, which minimizes the MSEP (5) under all predictors $\hat{y} = A\hat{\beta}$ with minimal bias

$$\Delta = \| \mu - E\hat{y} \|^2 = \min_{\hat{\beta}} \| \mu - A\hat{\beta} \|^2,$$

is

$$\hat{y} = A\hat{\beta}, \quad \hat{\beta} = (A^TWA)^+A^TW\hat{\mu},$$

where $C^+$ denotes the Moore-Penrose generalized inverse of the matrix $C$ and $W = \text{diag} \{ w_1, \ldots, w_m \}$, as shown in Bunke and Striiby (1975). This shows us that the predictor $\hat{y}$ and the estimate $\hat{\beta}$ only depend on those means $\bar{y}_i$ that correspond to nonvanishing weights $w_i > 0$. This is intuitively obvious, because without any prior information about the true response function $\eta$, the observations at those joint values of the input at which no prediction is intended have no informative value about the values of $\eta$ at the joint input values of predictive interest. Therefore, the predictor would be the same if we assumed only observations for those $i$ for which $w_i > 0$. Consequently, without loss of generality, we may always assume that all weights are positive. Then the matrix $A^TWA$ is of full rank and thus the best minimum bias predictor is

$$\hat{y} = A(A^TWA)^{-1}A^TWh.$$  

Its mean squared error of prediction is

$$r = \text{MSEP} (\hat{y}) = \Delta + \sigma^2(1 + t),$$

where

$$t = \text{tr} \, D, \quad D = WA(A^TWA)^{-1}A^TWL.$$  

In the following we will always use the best minimum bias predictor (6) for prediction. In the case of weights

$$w_i = n_i/n \quad (i = 1, \ldots, m),$$

which are proportional to the number $n_i$ of replications, it is identical to the ordinary least squares predictor $Ab$.

3. ESTIMATES OF THE MSEP

Some criteria for model selection use the RSS, which will be defined here under consideration of the weights and the number of replications at each value of the input as

$$\hat{r} = \text{RSS} = \sum_{i,j} w_i n_i^{-1}(y_{ij} - \bar{y}_i)^2.$$  

This “weighted” RSS is a negatively biased estimator of MSEP since

$$E\hat{r} - r = -2\sigma^2 t \leq 0,$$

as shown in the Appendix. With the unbiased estimate

$$\hat{\sigma}^2 = (n - m)^{-1} \sum_{i,j} (y_{ij} - \bar{y}_i)^2$$

of $\sigma^2$, we may adjust $\hat{r}$ for bias and obtain the adjusted RSS,

$$\hat{r}' = \hat{r} + 2\hat{\sigma}^2 t.$$  

In the case (9) of weights, $w_i$, which are proportional to the number, $n_i$, of replications, the adjusted estimate $\hat{r}' = \sigma^2(C_i/n + 1)$ is just the $C_i$-criterion (see Mallows 1973 or Montgomery and Peck 1982) up to a factor and an additive constant that do not depend on the model.

A “plug in” estimator of the MSEP may be obtained by replacing the unknown expectation $\mu$ and variance $\sigma^2$ in the formula (7) by their estimates $\hat{\mu}$ and $\hat{\sigma}^2$:

$$\hat{r}_B = \| \hat{\mu} - \bar{y} \|^2 + \hat{\sigma}^2(1 + t).$$

This estimator has a nonnegative bias,

$$E\hat{r}_B - r = \sigma^2(\text{tr} \, WL - t) \geq 0,$$

as calculated in Appendix A.1. A bias adjustment provides the “plug in” estimator

$$\tilde{r}_B = \| \hat{\mu} - \bar{y} \|^2 + \hat{\sigma}^2(1 - \text{tr} \, WL + 2t).$$

This estimator $\tilde{r}_B$ is a best unbiased estimator of $r$ because it depends on $y = (y_{11}, \ldots, y_{1m}, \ldots, y_{nm})^T$ through the statistic $(\hat{\mu}, \hat{\sigma}^2)$, which is sufficient and complete under (1) (see Bunke and Bunke 1984). In the special case,

$$w_1 = \cdots = w_m = m^{-1}, \quad n_1 = \cdots = n_m = h,$$

of equal weights and an equal number of replications, the estimator (16) is identical to the adjusted RSS $\hat{r}'$.

Mallows (1973) has already incidentally suggested the introduction of weights as in (5), and it may be seen that under our assumptions the criterion proposed in that paper is identical to (16). Besides the suggestion, this criterion has not been further investigated. The estimators (14) and (16) may also be obtained by application of the bootstrap approach, as shown in Bunke and Droge (1982). There the performance of Allen’s PRESS (a special case of cross-validation, see Stone 1974) as an estimator of MSEP is investigated, and its inferiority in comparison with the estimator (16) is shown under some general assumptions.

In this article, we want to compare “classical” criteria like the RSS and the adjusted RSS with the “plug in” estimator (14), the best unbiased estimator (16) and a minimum MSE estimator of MSEP, which we will now derive.

Looking at the form of the previous estimators, we may ask whether a linear combination of RSS, $\| \hat{\mu} - \bar{y} \|^2$ and $\hat{\sigma}^2$ different from $\hat{r}, \hat{r}', \tilde{r}_B$, and $\tilde{r}_B$, could provide a better MSE for estimating $r$. In Appendix A.4 we prove that the criterion

$$r = \| \hat{\mu} - \bar{y} \|^2 + q\hat{\sigma}^2,$$

TECHNOMETRICS © VOL. 26, NO. 2, MAY 1984
Table 1. MSE for Different Estimators of the MSEP

<table>
<thead>
<tr>
<th>Estimator</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{r} )</td>
<td>( R + 2\sigma^4{\text{tr}(WL - D)^2 + (n - m)} - (1 - \text{tr}WL + 2t) } )</td>
</tr>
<tr>
<td>( \hat{r}' )</td>
<td>( R + 2\sigma^4{\text{tr}(WL - D)^2 + 4(n - m)^2(1 + t) } )</td>
</tr>
<tr>
<td>( \hat{r}_B )</td>
<td>( R + 2\sigma^4{\text{tr}(WL - D)^2 + (n - m)(n - m + 2)(1 - \text{tr}WL + 2t) } )</td>
</tr>
<tr>
<td>( \hat{r}'_B )</td>
<td>( R + 2\sigma^4{\text{tr}(WL - D)^2 + (n - m - 1)(1 - \text{tr}WL + 2t) } )</td>
</tr>
<tr>
<td>( \hat{r} )</td>
<td>( R + 2\sigma^4{\text{tr}(WL - D)^2 + (n - m + 2)(1 - \text{tr}WL + 2t) } )</td>
</tr>
</tbody>
</table>

- \( q = (n - m)(n - m + 2)^{-1}(1 - \text{tr}WL + 2t) \), (19)
- \( \hat{r} = (1 - k)(\hat{\mu} - \hat{\mu})^2 + k \text{RSS} + k \sigma^2 \). (20)

The bias of this optimal estimator is

\[ E\hat{r} - r = -2(n - m + 2)^{-1}(1 - \text{tr}WL + 2t)\sigma^2 \] (21)

(see Appendix A.2).

This result shows that for estimating the MSEP, a term like \( \| \hat{\mu} - \hat{\mu} \| ^2 \) is more favorable than the RSS. The estimator (18) is very close to the best unbiased estimator (16) and its bias (21) will be relatively small. Only if the number of replications is small, that is, \( n - m \) is small, will there be a noticeable difference.

### 4. COMPARISONS OF THE MSEP ESTIMATORS

Table 1 contains the MSE formulas (proved in Appendix A.3) for the different estimators of MSEP con-
ESTIMATORS OF PREDICTION MSE

considered in Section 3. We use the notation
\[ R = 4\sigma^2 \sum_i w_i n_i^{-1} (\mu_i - \hat{E}_i)^2. \]  
(22)

From Table 1 we obtain the following:

1. The adjusted RSS \( \hat{r} \) has smaller MSE than the RSS \( r \) iff
   \[ n - m > 2 \quad \text{and} \quad t > 2(n - m - 2)^{-1}(1 - \text{tr} WL). \]

2. The “plug-in” estimator \( \hat{r}_B \) has smaller MSE than the RSS \( r \) if
   \[ t > (3n - 3m - 2)^{-1}((n - m - 2) \text{tr} WL + 4) \]
   (as shown in Appendix A.5).

3. As shown in Appendix A.6 the “plug-in” estimator \( \hat{r}_B \) never has a smaller MSE than its adjusted version \( \hat{r}_B^\prime \).

4. With \( n - m \geq 18 \), the best unbiased estimator \( \hat{r}_B \) is nearly as good as the optimal estimator \( \hat{r} \):
   \[ \text{MSE} (\hat{r}_B) - \text{MSE} (\hat{r}) \leq (0.1) \text{MSE} (\hat{r}_B) \]
   (see Appendix A.7).

In the case (9) of weights, which are proportional to the number of replications, there are simplifications in the formulas, since
\[ R = 4\sigma^2 \Delta/n, \]
\[ WL = n^{-1} I, \]
\[ D = WA(ATA)^{-1}AT/n, \]
\[ D^2 = D/n, \]
\[ t = n^{-1} \text{tr} [WA(ATA)^{-1}AT] \]
\[ = n^{-1} \text{rank} [ATA] = p/n, \]
\[ = n^{-1} \text{rank} [ATA] = p/n, \]
\[ \text{tr} (WL - D)^2 = (m - p)/n, \]
\[ \text{tr} (WL - D)^2 + \text{tr} W^2L - \text{tr} W^2L? = (1 - p/n)/n. \]

We see that the MSE of the different estimators only depends on the unknown response function in the term \( R \), namely by the “local biases,” \( \mu_i - \hat{E}_i \), in (22) or by the bias \( \Delta \) in the special case (23). The other term only depends on \( \sigma^2 \) and on the matrix \( D \) (see (8)), that is, on the model (3).

5. SOME NUMERICAL RESULTS

In order to illustrate the differences between the estimators of the MSEP, we calculated for each estimator \( \hat{r} \) the values
\[ M = \sigma^{-4}[\text{MSE} (\hat{r}) - R] \]

---

**Figure 3.** The case \( m = 10, h = 2 \).

**Figure 4.** The case \( m = 10, h = 5 \).

Figures 3 and 4. \( M = \sigma^{-4}[\text{MSE} (\hat{r}) - R] \) vs. \( p \) for different estimators \( \hat{r} \) of the MSEP (\( m \): number of input values; \( h \): number of replications).

TECHNOMETRICS ©, VOL. 26, NO. 2, MAY 1984
in the case (17) of equal weights and equal number of replications. This was done for different \( m = 2, 5, 10, 20 \), different numbers of replications \( h = 2, 5, 10, 20 \), and different model dimensions \( p = 1, \ldots, m \). \( M \) is just the term in the formulas for the MSE (4) (see Table 1), which is different for the different estimators \( \hat{r} \) of \( r \), while \( R \) is identical for all estimators.

\( M \) does not depend on any unknown parameter. It depends on the model (3) only by the dimension \( p \), and as it is proportional to the MSE (3 up to an additive constant, it is especially adequate for a numerical or graphical comparison of estimators.

On the other hand, it is clear that the \( M \) values essentially describe the differences in the MSE of the estimators if \( R \) (or \( \Delta \)) is small; that is, if the model is nearly exact, while the comparative use of differences in \( M \) values decreases with increasing \( R \). Already a few selected values of \( M \) seem to be sufficient to give a good impression of these numerical results, say for \( m = 5, 10 \) and \( h = 2, 5 \) (see Figures 1-4).

For comparison, we have also included the values \( M \) for a leaving-m-out variant, \( \hat{r}^{*,*} \), of a cross-validation criterion (in the sense of Stone 1974) defined in Bunke and Droge (1982), which is based on calculating for each \( j \) an estimate \( \hat{\theta} \) without the observations \( y_{ij}, \ldots, y_{mj} \).

We see that the optimal estimator \( \hat{r} \) is essentially better than the other estimators if small MSE for all \( p \) is wanted. The best unbiased estimator \( \hat{r}_0 \) should be first chosen among the other estimators. For sufficiently large \( h \), it seems to be nearly as good as the optimal estimator \( \hat{r} \) (for \( h = 5 \) at most 10% difference in MSE).

6. DISCUSSION OF THE RESULTS

The theoretical and numerical comparisons show that in general it is better to use estimators based on an unbiased estimate of the variance and on the “empirical model bias,” \( \| \hat{\theta} - \hat{\theta} \|^2 \), than estimators based on the RSS. The criterion \( \hat{r} \) with smallest MSE for estimating the MSEP of a model is given by (18), and in general its MSE is essentially smaller than that of the other estimators, with the exception of the best unbiased estimator \( \hat{r}_0 \), given by (16), which behaves nearly like \( \hat{r} \).

In the case (9) of weights that are proportional to the number of replications, our results show that the \( C_p \)-criterion of Mallows, which then is equivalent to the adjusted RSS, leads to a best unbiased estimator and, moreover, to a nearly optimal MSE in estimating the MSEP.

7. ESTIMATORS FOR THE MSE

If the regression model is used for approximating the unknown response function \( \eta \) or estimating its values at some values of the input instead of predicting the future value of the output (that is the value of the response function plus random error, as in the previous sections), the MSE for estimating the response function

\[
\rho = E \sum_{i} w_{ii} [\hat{y}_{i} - \eta(x_{i1}, \ldots, x_{iL})]^2 = \Delta + \sigma^2 t \tag{24}
\]

should be used instead of the MSEP (5) for an assessment of the model performance. Different estimators could be constructed as in Section 3. Their comparison can be made analogously as for the estimators of the MSEP and, therefore, we state the results without proofs.

Of special interest is the RSS \( \hat{r} \), its adjusted unbiased variant,

\[
\hat{\rho} = \hat{\rho} + \hat{\theta}^2(t - 1),
\]

the “plug-in” estimator,

\[
\hat{\rho}_0 = \| \hat{\theta} - \hat{\theta} \|^2 + \hat{\theta}^2 t,
\]

and its adjusted unbiased variant,

\[
\hat{\rho}_0 = \| \hat{\theta} - \hat{\theta} \|^2 + \hat{\theta}^2 (2t - \text{tr} WL) = \hat{\rho}_0 - \hat{\theta}^2, \tag{25}
\]

which is a best unbiased estimator for \( \rho \). In the case of estimators of the form

\[
\hat{\rho} = (1 - k) \| \hat{\theta} - \hat{\theta} \|^2 + k\hat{r} + h\hat{\theta}^2,
\]

we obtain a minimal MSE (\( \hat{\rho} \)) for

\[
\hat{\rho} = \| \hat{\theta} - \hat{\theta} \|^2 + q\hat{\theta}^2, \tag{26}
\]

where

\[
q = 4(n - m)(n - m - 2)^{-1}(2t - \text{tr} WL). \tag{27}
\]

The MSE and the bias for the different estimators are presented in Table 2, where \( R \) is defined as in (22). An examination of Table 2 yields the following relations:

1. MSE (\( \hat{\rho}_0 \)) \( \leq \) MSE (\( \hat{\rho} \)).
2. MSE (\( \hat{\rho}_0 \)) \( \leq \) MSE (\( \hat{\rho}_0 \)) if \( n \geq m + 2 \).
3. With \( n \geq m + 2 \), it holds that MSE (\( \hat{r} \)) \( \leq \) MSE (\( \hat{\rho} \)) if

\[
1 \leq 2t \leq 4(n - m - 2)^{-1}(1 - \text{tr} WL) + 1.
\]

Therefore,

(a) \( \lambda := \min_i (n_i) \geq 2 \) and

(b) \( n_i = nw_i, \quad 2p \leq n \)

are sufficient conditions for

MSE (\( \hat{\rho} \)) \( \leq \) MSE (\( \hat{r} \)).

4. Under \( \lambda := \min_i (n_i) \geq 2 \), it holds that

MSE (\( \hat{\rho}_0 \)) \( \leq \) MSE (\( \hat{r} \))

and

\[
| \text{bias (\( \hat{\rho}_0 \))} | \leq | \text{bias (\( \hat{r} \))} |.
\]

5. MSE (\( \hat{\rho}_0 \)) - MSE (\( \hat{\rho} \)) \( \leq (1) \) MSE (\( \hat{\rho}_0 \)) if \( n \geq m + 18 \).
### Table 2. MSE and Bias for Different Estimators of the MSE \( \rho \)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>MSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{f} )</td>
<td>( R + 2\sigma^2 \left{ \text{tr}(WL - D)^2 + \text{tr}(W^2L - W^2L^2) + (1 - 2t)^2/2 \right} )</td>
<td>( (1 - 2t)\sigma^2 )</td>
</tr>
</tbody>
</table>
| \( \hat{\rho} \) | \( R + 2\sigma^2 \left\{ \text{tr}(WL - D)^2 + \text{tr}(W^2L - W^2L^2) \right\} 
+ (n - m)^{-1} \left\{ (2t - 1)(2t - 2tW + 1) \right\} \) | 0 |
| \( \hat{\rho}_0 \) | \( R + 2\sigma^2 \left\{ \text{tr}(WL - D)^2 + (n - m)^{-1}t^2 + (\text{tr}WL - t)^2/2 \right\} \) | \( (\text{tr}WL - t)\sigma^2 \) |
| \( \hat{\rho}_1 \) | \( R + 2\sigma^2 \left\{ \text{tr}(WL - D)^2 + (n - m)^{-1}(2t - \text{tr}WL)^2 \right\} \) | 0 |
| \( \hat{\rho} \) | \( R + 2\sigma^2 \left\{ \text{tr}(WL - D)^2 + (n - m + 2)^{-1}(2t - \text{tr}WL)^2 \right\} \) | \( -2(2t - \text{tr}WL) \sigma^2 / (n - m + 2) \) |

We see that the optimal estimators \( \hat{r} \) for the MSEP and \( \hat{\rho} \) for the MSE only differ by a term that does not depend on the model \( (3) \). Consequently, any of these, say \( \hat{\rho} \), could be used as a criterion for the comparison of models with the objective of predicting the output or, alternatively, of estimating the response function. As discussed in Section 1, comparison of models by a criterion is only a tool in model selection. After finally arriving at some potentially good models by some more complex strategy, a description of their performance by a good estimate of their corresponding MSEP (or MSE), like \( \hat{r} \) (or \( \hat{\rho} \)), is obviously of interest.

### 8. ESTIMATING THE MSEP AND THE MSE WITHOUT REPLICATED OBSERVATIONS

If there are no replicated observations—that is, for each input \( x_{i1}, \ldots, x_{ik} \) \( (i = 1, \ldots, m) \), there is one observation \( y_i \) \( (n_i = 1) \)—then an estimate of the unknown variance \( \sigma^2 \), on the basis of the observations, may only be constructed if there is some information on the response function, for example, leading to an (at least approximately) exact linear model for the observation vector \( y = (y_1, \ldots, y_m)^T \).

We assume
\[
y \sim N(H\lambda, \sigma^2I), \quad \lambda \in R^q,
\]
where \( H \) is a known \( n \times q \) matrix of rank \( q < n = m \) and \( \lambda \) an unknown parameter. Then
\[
\hat{\mu} = H(H^TH)^{-1}H^Ty \quad (29)
\]

and
\[
\hat{\sigma}^2 = (n - q)^{-1} \| y - \hat{\mu} \|^2 \quad (30)
\]
are best unbiased estimators of the unknown parameters \( \mu = H\lambda \) and \( \sigma^2 \).

If the model \( (28) \) would not be exact, then the estimator \( (30) \) would overestimate the variance and an estimator of MSEP based on \( (30) \) would be misleading in comparing models, because models with few parameters (i.e., smaller constant \( t \)) would be unduly favored.

Now we may derive, with almost the same calculations as in the previous sections, best unbiased estimators for the MSEP \( (7) \) and the MSE \( (24) \) and, moreover, minimum MSE estimators in the class of estimators of the form \( (20) \). Here and in the following, the estimators \( (29) \) and \( (30) \) always replace the estimators \( \hat{\mu} \) (see \( (4) \)) and \( y \) (see \( (6) \)) by a best minimum bias predictor. The formulas \( (16), (18), (19), (25), (26), \) and \( (27) \) for the optimal estimators of MSEP and MSE remain the same; only the matrix \( L \) (see \( (4) \)) is to be replaced by the matrix
\[
L = H(H^TH)^{-1}H^T.
\]

**Remark 1.** The assumption \( (28) \) covers the cases of replicated observations—for example, case \( (1) \), as a special case, where the corresponding structure of the matrix \( H \) is obvious.

**Remark 2.** If there is knowledge on the structure of the expectation \( \mu = H\lambda \) in the sense of assumption \( (28) \), there is still interest in predictors \( \hat{y} = A\hat{\mu} \) (see \( (6) \)) given by another matrix \( A \) of smaller rank (i.e., using an inexact model) because the MSEP \( (7) \) could be essentially diminished (smaller constant \( t \)) by such a predictor, even if then there is some bias \( A \).

### ACKNOWLEDGMENT

The authors are very grateful to the referees for their remarks, which improved the presentation of the article considerably.

### APPENDIX

In this appendix we give the proofs of the statements in Sections 3 and 4, using the following notation.

1. \( \otimes \): Kronecker product.
2. \( 1_n \): \( n \)-vector of ones.
3. \( I_k \): \( k \times k \) unity matrix.
4. \( y = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{m1}, \ldots, y_{mn_n})^T \).
Furthermore, it follows from (31) and (36) that
\[ E \| y - \hat{\mu} \|_2^2 = E \| y \|_2^2 = \| \mu \|_2^2 + \sigma^2 \text{ tr } Q \]
\[ = \sigma^2 (1 - \text{ tr } WL), \]
which, together with (32) and (39), provides
\[ E \hat{r} = \Delta + \sigma^2 (1 - \text{ tr } D) \]
and, therefore, the formula for \( E \hat{r} - r \).

A.1 Proof of (15)

Equation (15) follows directly from (7), (14), (39), and \( E \hat{\sigma}^2 = \sigma^2 \). We remark that \( \text{ tr } WL - \text{ tr } D = \sum_i n_i^{-1} w_i (1 - p_{it}) \) can never be positive because \( p_{it} \leq 1 \) for \( i = 1, \ldots, m \).

A.2 Proof of (21)

Using (7), (18), and (39), we obtain
\[ E \hat{r} - r = \sigma^2 (q - 1 + \text{ tr } WL - 2t) \]
\[ = \sigma^2 (1 - \text{ tr } WL + 2t) \]
\[ \times \{ (n - m) (n - m + 2) \} \]
and, therefore, (21).

A.3 Proof of the MSE Formulas in Table 1

1. For the calculation of
\[ \text{MSE}(\hat{r}) = (E \hat{r} - r)^2 + \text{ var}(\hat{r}), \]
we will use the formulas for the bias of the estimates and (37). First we consider an estimate of \( r \) of the form
\[ \hat{r} = \hat{\mu} + \hat{y}, \]
where \( \hat{\mu} \) and \( \hat{y} \) are independent under (1) (see Rao 1972). Therefore, we obtain
\[ \text{ var } \hat{r} = \text{ var } \hat{\mu} + \text{ var } \hat{y} \]
and
\[ U \hat{\mu} = WL - D \]
which, together with (33), (37), and (39), yields
\[ \text{ var } \hat{\mu} = \text{ var } \| \hat{\mu} \|_2^2 = 4 \sigma^2 \mu^T U L \mu = 4 \sigma^2 \mu^T A \Sigma A \mu = 4 \mu^T A \Sigma A \mu + 2 \text{ tr } (A \Sigma)^2, \]
and thus
\[ A \sigma^2 \mu^T U L U \mu = 4 \sigma^2 \| \mu \|_2^2 \]
From
\[ U L = W L - D \]
we derive
\[ \text{ tr } (U L)^2 = \text{ tr } (W L - D)^2, \]
which, together with (31), (33), (37), and (43), yields
\[ \text{ var } \| \hat{\mu} - \hat{y} \|_2^2 = \text{ var } \| \hat{\mu} \|_2^2 \]
since
\[ \hat{\mu} \sim N(\mu, \sigma^2 L). \]
Using (34), we may write
\[ \| \tilde{y} - \hat{\mu} \|^2 = \| y \|^2. \] (45)

The identities in (35) lead to
\[ \| \hat{\mu} \|^2 = 0, \] (46)

and, moreover, with (34) and (35) to
\[ \text{tr } T^2 = \text{tr } ([I - K]S)^2 \]
\[ = \text{tr } \{ a[I - K] + b[V - VK] \}^2 \]
\[ - a^2 \text{ tr } (I - K) + 2ab \text{ tr } (V - VK) \]
\[ + b^2 \text{ tr } (V - VK)^2 \]
\[ = a^2 (n - m) + 2ab(1 - \text{tr } WL) \]
\[ + b^2 \text{ tr } (V - VK)^2. \] (47)

From (37), (45), (46), and (47), we obtain
\[ \text{var } \| y - \hat{\mu} \|^2 = 2\sigma^4 \{ a^2(n - m) + \]
\[ + 2ab(1 - \text{tr } WL) + b^2 \text{ tr } (V - VK)^2 \}, \]

which, together with (39) and (40), provides
\[ \var \left( \hat{\phi}^* (a, b) \right) = R + \]
\[ 2a^4 \{ \text{tr } (WL - D)^2 + h^2(n - m)^{-1} \]
\[ + 2hk(n - m)^{-1}(1 - \text{tr } WL) \]
\[ + (h + k + 1 + \text{tr } WL) \]
\[ - k \text{ tr } WL - 2t^2 / 2 \}. \] (48)

2. To find the optimum pair \((h, k)\), we minimize \(\phi(h, k)\) with respect to \((h, k)\). From the necessary condition for a minimum \((h^*, k^*)\),
\[ \frac{\partial \phi(h, k)}{\partial h} \Big|_{(h^*, k^*)} = 0, \]

it results that
\[ k^*(1 - \text{tr } WL) = \]
\[ (n - m)^{-1}(n - m + 2)(n - m)^{-1} \]
\[ + (n - m)(n - m + 2)^{-1}(1 - \text{tr } WL + 2t) \] (51)

The second necessary condition,
\[ \frac{\partial \phi(h, k)}{\partial k} \Big|_{(h^*, k^*)} = 0, \]

leads to
\[ 0 = 2h^*(n - m)^{-1}(1 - \text{tr } WL) \]
\[ + 2k^* \text{ tr } (W^2L - W^2L^2) \]
\[ + (h^* + k^* - 1 + \text{tr } WL - k^* \text{ tr } WL \]
\[ - 2t(1 - \text{tr } WL). \] (52)

Now let
\[ c := (n - m)^{-1}(1 - \text{tr } WL)^2 \]
\[ - \text{tr } (W^2L - W^2L^2). \] (53)

Then, from (51) and (52), we obtain
\[ 2ck^* = 0. \] (54)

Because of \(\sum_i (n_i - 1) = n - m\) and the strict convexity of the function \(f(x) = x^2\), Jensen’s inequality yields
\[ \sum_i (n_i - 1)(n - m)^{-1}w_i n_i^{-2} \]
\[ \geq (n - m)^{-2} \left( \sum_i (n_i - 1)w_i n_i^{-1} \right)^2, \] (55)

TECHNOMETRICS ©, VOL. 26, NO. 2, MAY 1984
where equality holds iff
\[ w_i n_i^{-1} = w_j n_j^{-1} \text{ for } i, j = 1, \ldots, m. \] (56)

We remark that (56) is equivalent to \( nV = I \). Since
\[ \text{tr} (W^2 L - W^2 L^2) = \sum_i (n_i - 1) w_i^2 n_i^{-2} \]
and
\[ (1 - \text{tr} WL)^2 = \left(1 - \sum_i w_i n_i^{-1}\right)^2 = \left(\sum_i (n_i - 1) w_i n_i^{-1}\right)^2, \]
we obtain from (53), (55), and (56)
\[ c \leq 0, \] (57)
where
\[ c = 0 \iff nV = I. \] (58)

In the case of \( c \neq 0 \), it follows from (54) that
\[ k^* = 0 \] (59)
and, therefore, together with (51),
\[ h^* = \hat{q}. \] (60)

For \( c = 0 \), that is, \( nV = I \) (see (58)), any solution \((h^*, k^*)\) of (51) provides
\[ \bar{r} = \| \hat{y} - \hat{\mu} \|^2 + k^* \| y - \hat{\mu} \|^2 + h^* \sigma^2 \]
\[ - \| \hat{\mu} - \hat{y} \|^2 + q \sigma^2 - \bar{r}, \]
since then
\[ \| y - \hat{\mu} \|^2 = (n - m) \sigma^2 \text{ and } \text{tr} WL = m/n. \]

It is easily verified that \( h^* \) and \( k^* \), defined as in (59) and (60), always minimize \( \phi(h, k) = \text{MSE} (\bar{r}) \).

### A.5 Proof of Statement 2 in Section 4

Note that
\[ \sigma^{-4} \text{MSE} (\hat{r}) - \text{MSE} (\bar{r}) = 2 \text{tr} (W^2 L - W^2 L^2) \]
\[ - \text{tr} WL - t^2 + 4r^2 - 2(n - m)^{-1}(1 + t)^2. \] (61)

Using (57) we have
\[ \text{tr} (W^2 L - W^2 L^2) \geq (n - m)^{-1}(1 - \text{tr} WL)^2, \]
which, together with (61), leads to
\[ \sigma^{-4} \text{MSE} (\hat{r}) - \text{MSE} (\bar{r}) \geq (n - m)^{-1}(t + \text{tr} WL) \]
\[ \times \{ (3n - 3m - 2)t - (n - m - 2) \text{tr} WL - 4 \}. \] (62)

Thus the statement directly follows from (62), since
\[ t + \text{tr} WL > 0. \]

### A.6 Proof of Statement 3 in Section 4

From Table 1 we derive
\[ \sigma^{-4} \text{MSE} (\bar{r}) - \text{MSE} (\bar{r}) = (n - m)^{-1}(t - \text{tr} WL) \]
\[ \times \{ (n - m - 6)t - (n - m - 2) \text{tr} WL - 4 \}. \] (63)

Now it can be easily verified that
\[ (n - m - 2) \text{tr} WL + 4 > (n - m - 6)t, \]
which, together with (63) and \( 0 \leq t \leq \text{tr} WL \), leads to
\[ \text{MSE} (\bar{r}) - \text{MSE} (\bar{r}) \geq 0. \]

### A.7 Proof of Statement 4 in Section 4

Table 1 yields
\[ \text{MSE} (\bar{r}) - \text{MSE} (\bar{r}) = 4 \{(n - m)(n - m + 2)\}^{-1} \]
\[ \times (1 - \text{tr} WL + 2t)^2 \sigma^2 (n - m + 2)^{-1} \text{MSE} (\bar{r}), \]
from which statement 4 in Section 4 directly follows because
\[ 2(n - m + 2)^{-1} \leq 1 \iff n - m \geq 18. \]

[Received July 1981. Revised December 1983.]

### REFERENCES


