AUTHOR'S RESPONSE TO DISCUSSIONS OF PAPER PUBLISHED IN VOLUME LXXXIII

LOSS PREDICTION BY GENERALIZED LEAST SQUARES

LEIGH J. HALLIWELL

1. INTRODUCTION

Having had the pleasure of seeing my paper in the *Proceedings*, I am even more pleased now that Klaus Schmidt and Michael Hamer have deigned to discuss it. But even with their discussions, most of the subject of statistically modeling loss triangles remains *terra incognita*; and I hope that actuaries and academics will continue to explore it.

2. BACKGROUND

Since I wrote the paper late in 1994, I have learned more about statistical modeling. I recommend for interested readers to examine my 1997 *Forum* paper, "Conjoint Prediction of Paid and Incurred Losses," especially its Appendices A and C. Nevertheless, I stand by the conclusions of the earlier paper:

> This paper will argue that the linear modeling and the least squares estimation found in the literature to date have overlooked an important condition of the linear model. In particular, the models for development factors regress random variables against other random variables. Stochastic regressors violate the standard linear model. Moreover, the model assumes that errors are uncorrelated, but stochastic regressors violate this assumption as well. This paper will show that what actuaries are really seeking is found in a general linear model; i.e., a model with nonstochastic regressors but with an error matrix that allows for correlation. [2, p. 436]

> [The use of stochastic regressors] is the fundamental problem with the CL [Chain Ladder] method. Rather than try to rehabilitate it, this paper introduces a different model that honors all the conditions of the Gauss–Markov theorem. [2, p. 441]
>
> A theory becomes very attractive when it unifies partial explanations. Such is the case with loss covariance. CL, prior hypothesis, or BF [Bornhuetter–Ferguson]—which to choose? The answer will lie on a continuum dependent on the variance matrix of the incremental losses. [2, p. 447]
>
> Generalized least squares is a better method of loss prediction than the chain ladder and the other loss development methods. Even when linear models are imposed on loss development methods, they incorporate stochastic regressors, and the estimates are not guaranteed to be either best or unbiased. The confidence intervals derived therefrom are not trustworthy. The fault lies in trying to make the level on one variable affect the level of the next, whereas the statistical idea is that the departure of one variable from its mean affects the departure of the next from its mean. This is the idea of covariance, and it is accommodated in the general linear model and generalized least squares estimation. [2, p. 456]

The problem of stochastic regressors quells my enthusiasm for empirically testing chain-ladder statistical models (as, for example, Gary Venter [6] recommends). The technique of instrumental variables [4, p. 577 and 5, p. 198] solves this problem; but the obvious instrument for a lagged loss is its exposure. And when exposure becomes a regressor, the lagged loss often lacks significance, as Glen Barnett and Ben Zehnwirth have discovered [1, p. 10]. So I am hopeful that actuaries will find their way back to the no-frills "additive model" [2, pp. 442,

449] and thence begin to consider non-trivial covariance structures.[1]

3. AUTHOR'S COMMENTS ON ORIGINAL PAPER

Before responding to the discussions I will point out two flaws of the paper. The first flaw concerns pages 450f. and Exhibit 3. I derived an estimate of β, reweighted the observations, and derived a second estimate of β. I remarked, "The estimate for β changes negligibly (no change within the first ten decimal places)." [2, p. 451] Such a negligible change should have clued me that the estimates of β were identical, the difference owing to computational precision. If one regresses Y against X with error variances σ, the estimate is:

$$\frac{\sum_i \frac{x_i y_i}{\sigma_{ii}}}{\sum_i \frac{x_i x_i}{\sigma_{ii}}}.$$

Therefore, the estimate is invariant to a scale change of the variances. Now the second model applied scale factors according to age. But each element of $\hat{\beta}$ depends on observations of the same age, which have been affected by the same scale factor. Thus the estimate is unchanged.[2]

The second flaw concerns the degrees of freedom in the estimate of σ^2. There were thirty-six observations, eight parameters in β, and two parameters in the variance matrix. I claimed there to be $36 - 8 - 2 = 26$ degrees of freedom [2, p. 453]. But the two parameters that had been estimated in the variance matrix are not like those of β. There is no theoretically right way of accounting for the variance parameters, and twenty-eight degrees of freedom is just as acceptable as twenty-six.

---

[1] My session "Regression Models and Loss Reserving" at the 1999 Casualty Loss Reserve Seminar presents this broad subject with theory and examples.

[2] I am grateful to William A. Niemczyk for pointing this out to me.

4. AUTHOR'S COMMENTS ON DISCUSSIONS

Drs. Schmidt and Hamer have confined their discussions to the Gauss–Markov theorem and to the best linear unbiased predictor. This is natural, since the Gauss–Markov theorem is the most mathematical topic of the paper and is new material to most actuaries (at least in its matrix form). In several of my papers I have complained that we actuaries know too little about statistical modeling and the matrix algebra that it utilizes. I myself learned what little I know by a time-consuming study of materials outside the actuarial syllabus, particularly [4]. And I believe that even the new actuarial syllabus does not adequately cover this topic. However, I wish that these discussions had gotten beyond the Gauss–Markov theorem and treated the undesirability of stochastic regressors and the distinction between loss covariance and loss development.

Dr. Schmidt's finish, "We thus obtain the predictor proposed by Halliwell by a direct approach which avoids conditioning," provides the basis for my two-fold response. First, as to conditioning, my treatment of the predictor in Appendix C does not depend on Bayes' theorem and a loss distribution. In fact, I wrote that $\mathbf{e}$ is "not necessarily normal" [2, pp. 480, 473]. However, perhaps I invited Dr. Schmidt's criticism when I used conditional-expectation notation [2, pp. 445, 482f] and said that the unknown elements "are affected by the known elements in a Bayesian sense, through the variance matrix." [2, p. 444] My Appendix B demonstrated that if $\mathbf{e}$ is multivariate normal, the predictor can be derived by Bayes' theorem; but I did not say that conditional probability was the rationale of the predictor.

And second, Drs. Schmidt and Hamer have made my argument rigorous, and shown that one can bypass the estimation of β on the way to estimating $\mathbf{Y}_2$ (the "direct approach"). I concur with their assertions that the proof in my Appendix C was not strict, and that it confined itself "to predictors which can be

written as $y_2(\hat{\beta})$ for some admissible estimator $\hat{\beta}$." I had realized these things when I wrote my paper on conjoint prediction [3]. There I formulated the partitioned model (p observations and q predictions):

$$\begin{bmatrix} \mathbf{Y}_{1(p\times 1)} \\ \cdots\cdots \\ \mathbf{Y}_{2(q\times 1)} \end{bmatrix} = \begin{bmatrix} \mathrm{X}_{1(p\times k)} \\ \cdots\cdots \\ \mathrm{X}_{2(q\times k)} \end{bmatrix} \beta_{(k\times 1)} + \begin{bmatrix} \mathbf{e}_1 \\ \cdots \\ \mathbf{e}_2 \end{bmatrix}, \qquad \text{where}$$

$$\operatorname{Var}\begin{bmatrix} \mathbf{e}_1 \\ \cdots \\ \mathbf{e}_2 \end{bmatrix} = \begin{bmatrix} \mathrm{S}_{11(p\times p)} & \vdots & \mathrm{S}_{12(p\times q)} \\ \cdots\cdots\cdots & \vdots & \cdots\cdots\cdots \\ \mathrm{S}_{21(q\times p)} & \vdots & \mathrm{S}_{22(q\times q)} \end{bmatrix}$$

And I showed [3, p. 328] that the best linear unbiased predictor of $\mathbf{Y}_2$ is:

$$\hat{\mathbf{Y}}_2 = (\mathrm{S}_{21}\mathrm{S}_{11}^{-1} + (\mathrm{X}_2 - \mathrm{S}_{21}\mathrm{S}_{11}^{-1}\mathrm{X}_1)(\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{S}_{11}^{-1})\mathbf{Y}_1$$

This agrees with Dr. Schmidt's Theorem 4.3, whose proof Dr. Hamer has provided. This formulation is direct because the estimator $\hat{\mathbf{Y}}_2$ does not involve $\hat{\beta}$. However, if $\mathrm{X}_2 = \mathrm{I}_k$ and $\mathbf{e}_2$ is a zero matrix (and hence S_{21} and S_{22} are zero matrices), then $\mathbf{Y}_2 = \beta$, and:

$$\begin{aligned} \hat{\beta} = \hat{\mathbf{Y}}_2 &= (\mathrm{S}_{21}\mathrm{S}_{11}^{-1} + (\mathrm{X}_2 - \mathrm{S}_{21}\mathrm{S}_{11}^{-1}\mathrm{X}_1)(\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{S}_{11}^{-1})\mathbf{Y}_1 \\ &= (0\mathrm{S}_{11}^{-1} + (\mathrm{I}_k - 0\mathrm{S}_{11}^{-1}\mathrm{X}_1)(\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{S}_{11}^{-1})\mathbf{Y}_1 \\ &= (\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathbf{Y}_1 \end{aligned}$$

So the estimation of β is a special case of the estimation of $\mathbf{Y}_2$ [3, p. 331], which Dr. Hamer calls Case 1 of his Theorem 4.1.[3]

That really is all that I need to say about the Gauss–Markov theorem and best linear unbiased prediction. The task now, as

[3]Dr. Hamer devotes his Appendix to deriving the best linear unbiased estimator (BLUE) of $\mathrm{F}_1\mathbf{Y}_1 + \mathrm{F}_2\mathbf{Y}_2 + \mathrm{A}\beta$. Though correct, the form of this derivation is overly complex. I have shown [3, p. 335f] that the estimator is a linear operator; hence, the BLUE of this expression is $\mathrm{F}_1\hat{\mathbf{Y}}_1 + \mathrm{F}_2\hat{\mathbf{Y}}_2 + \mathrm{A}\hat{\beta} = \mathrm{F}_1\mathbf{Y}_1 + \mathrm{F}_2\hat{\mathbf{Y}}_2 + \mathrm{A}\hat{\beta}$.

I see it, is to get actuaries to understand that this theory is not just a mathematical nicety. Though perhaps not a Copernican revolution, it is revolutionary nonetheless. As it makes inroads, we will see less of development factors and loss adjustments and more of modeling and exposure adjustments.

REFERENCES

[1] Barnett, Glen, and Ben Zehnwirth, "Best Estimates for Reserves," Casualty Actuarial Society *Forum*, Fall 1998, pp. 1–54.

[2] Halliwell, Leigh J., "Loss Prediction by Generalized Least Squares," *PCAS* LXXXIII, 1996, pp. 436–489.

[3] Halliwell, Leigh J., "Conjoint Prediction of Paid and Incurred Losses," Casualty Actuarial Society *Forum*, Summer 1997, pp. 241–380.

[4] Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lütkepohl, and T.-C. Lee, *Introduction to the Theory and Practice of Econometrics*, Second Edition, New York, John Wiley, 1988.

[5] Pindyck, Robert S., and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, Fourth Edition, Boston, Irwin/McGraw-Hill, 1998.

[6] Venter, Gary G., "Testing the Assumptions of Age-to-Age Factors," *PCAS* LXXXV, 1998, pp. 807–847.

APPENDIX A

As an appendix, I wish to comment on the optimization problem of Dr. Schmidt's fifth section, and on Dr. Hamer's generalization of it. Though this problem has occasioned some interesting mathematics, I see the problem as a sidelight, as only loosely related to the Gauss–Markov theorem.

Dr. Schmidt wishes to find the admissible estimator $\hat{\beta}$ that minimizes:

$$E[(\mathbf{Y} - \mathrm{X}\hat{\beta})'\mathrm{S}^{-1}(\mathbf{Y} - \mathrm{X}\hat{\beta})].$$

Estimator $\hat{\beta}$ is admissible if and only if it is a linear function of $\mathbf{Y}_1$ and it is unbiased. In his third section he shows that admissible estimators are of the form $\mathrm{B}_{(k\times p)}\mathbf{Y}_1$ for $\mathrm{BX}_1 = \mathrm{I}_k$, and $\mathrm{Var}[\hat{\beta}] = \mathrm{Var}[\mathrm{B}\mathbf{Y}_1] = \mathrm{B}\,\mathrm{Var}[\mathbf{Y}_1]\mathrm{B}' = \mathrm{BS}_{11}\mathrm{B}'$.

As I had done [2, p. 480f], he factors S^{-1} as $\mathrm{W}'\mathrm{W}$, where:

$$\mathrm{W} = \begin{bmatrix} \mathrm{A}_{(p\times p)} & 0_{(p\times q)} \\ \mathrm{C}_{(q\times p)} & \mathrm{D}_{(q\times q)} \end{bmatrix},$$

such that

$$\mathrm{A}'\mathrm{A} = \mathrm{S}_{11}^{-1},$$

$$\mathrm{D}'\mathrm{D} = (\mathrm{S}_{22} - \mathrm{S}_{21}\mathrm{S}_{11}^{-1}\mathrm{S}_{12})^{-1}, \qquad \text{and}$$

$$\mathrm{C} = -\mathrm{DS}_{21}\mathrm{S}_{11}^{-1}.$$

Now:

$$\mathrm{W}(\mathbf{Y} - \mathrm{X}\hat{\beta}) = \begin{bmatrix} \mathrm{A}_{(p\times p)} & 0_{(p\times q)} \\ \mathrm{C}_{(q\times p)} & \mathrm{D}_{(q\times q)} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1 - \mathrm{X}_1\hat{\beta} \\ \mathbf{Y}_2 - \mathrm{X}_2\hat{\beta} \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{A}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta}) \\ \mathrm{C}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta}) + \mathrm{D}(\mathbf{Y}_2 - \mathrm{X}_2\hat{\beta}) \end{bmatrix}$$

$$= \begin{bmatrix} A(\mathbf{Y}_1 - X_1\hat{\beta}) \\ D\mathbf{Y}_2 - DX_2\hat{\beta} + DD^{-1}C(\mathbf{Y}_1 - X_1\hat{\beta}) \end{bmatrix}$$

$$= \begin{bmatrix} A(\mathbf{Y}_1 - X_1\hat{\beta}) \\ D(\mathbf{Y}_2 - X_2\hat{\beta} + D^{-1}C(\mathbf{Y}_1 - X_1\hat{\beta})) \end{bmatrix}$$

$$= \begin{bmatrix} A(\mathbf{Y}_1 - X_1\hat{\beta}) \\ D(\mathbf{Y}_2 - y_2(\hat{\beta})) \end{bmatrix}$$

Therefore:

$$\begin{aligned}
&(\mathbf{Y} - X\hat{\beta})'S^{-1}(\mathbf{Y} - X\hat{\beta}) \\
&\quad = (\mathbf{Y} - X\hat{\beta})'W'W(\mathbf{Y} - X\hat{\beta}) \\
&\quad = (W(\mathbf{Y} - X\hat{\beta}))'(W(\mathbf{Y} - X\hat{\beta})) \\
&\quad = [(\mathbf{Y}_1 - X_1\hat{\beta})'A' \quad (\mathbf{Y}_2 - y_2(\hat{\beta}))'D'] \begin{bmatrix} A(\mathbf{Y}_1 - X_1\hat{\beta}) \\ D(\mathbf{Y}_2 - y_2(\hat{\beta})) \end{bmatrix} \\
&\quad = (\mathbf{Y}_1 - X_1\hat{\beta})'A'A(\mathbf{Y}_1 - X_1\hat{\beta}) \\
&\qquad + (\mathbf{Y}_2 - y_2(\hat{\beta}))'D'D(\mathbf{Y}_2 - y_2(\hat{\beta})) \\
&\quad = (\mathbf{Y}_1 - X_1\hat{\beta})'S_{11}^{-1}(\mathbf{Y}_1 - X_1\hat{\beta}) \\
&\qquad + (\mathbf{Y}_2 - y_2(\hat{\beta}))'D'D(\mathbf{Y}_2 - y_2(\hat{\beta}))
\end{aligned}$$

And we have Dr. Schmidt's Lemma 5.1:

$$\begin{aligned}
E[(\mathbf{Y} - X\hat{\beta})'S^{-1}(\mathbf{Y} - X\hat{\beta})] = {} & E[(\mathbf{Y}_1 - X_1\hat{\beta})'S_{11}^{-1}(\mathbf{Y}_1 - X_1\hat{\beta})] \\
& + E[(\mathbf{Y}_2 - y_2(\hat{\beta}))'D'D(\mathbf{Y}_2 - y_2(\hat{\beta}))]
\end{aligned}$$

To prove his Theorem 5.2 we have to review the trace function. The trace of a square matrix Q is defined as the sum of its diagonal elements: $\mathrm{tr}(Q_{(n\times n)}) = \sum_{i=1}^{n} q_{ii}$. Some theorems that

should be obvious are:

$$\begin{aligned}
\operatorname{tr}(\alpha \mathrm{Q}) &= \alpha \operatorname{tr}(\mathrm{Q}) \\
\operatorname{tr}(\mathrm{Q}') &= \operatorname{tr}(\mathrm{Q}) \\
\operatorname{tr}(\mathrm{Q}_1 + \mathrm{Q}_2) &= \operatorname{tr}(\mathrm{Q}_1) + \operatorname{tr}(\mathrm{Q}_2) \\
\operatorname{tr}(\mathrm{I}_n) &= n
\end{aligned}$$

If Q is (1×1), then $\operatorname{tr}(\mathrm{Q}) = q_{11} = \mathrm{Q}$. (For our purposes we may ignore the distinction between a scalar and a one-element matrix.) And if $\mathbf{Q}$ is a random matrix:

$$\begin{aligned}
\operatorname{tr}(E[\mathbf{Q}]) &= \sum_{i=1}^{n} E[\mathbf{q}_{ii}] \\
&= E\left[\sum_{i=1}^{n} \mathbf{q}_{ii}\right] \\
&= E[\operatorname{tr}(\mathbf{Q})]
\end{aligned}$$

But a theorem that is not obvious is that if A is $(m \times n)$ and B is $(n \times m)$, then the traces of AB and BA are equal. The proof is:

$$\begin{aligned}
\operatorname{tr}(\mathrm{AB}) &= \sum_{i=1}^{m} [\mathrm{AB}]_{ii} \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} b_{ji} \right) \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{m} b_{ji} a_{ij} \right) \\
&= \sum_{j=1}^{n} [\mathrm{BA}]_{jj} = \operatorname{tr}(\mathrm{BA})
\end{aligned}$$

With this knowledge of the trace we can prove Theorem 5.2.

We reduce the first term on the right side of Lemma 5.1, mindful of the fact that the expressions within the expectation operators are (1×1) matrices:

$$\begin{aligned}
&E[(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})'\mathrm{S}_{11}^{-1}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})] \\
&\quad = E[\mathrm{tr}((\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})'\mathrm{S}_{11}^{-1}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta}))] \\
&\quad = E[\mathrm{tr}(\mathrm{S}_{11}^{-1}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})')] \\
&\quad = \mathrm{tr}(E[\mathrm{S}_{11}^{-1}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})']) \\
&\quad = \mathrm{tr}(\mathrm{S}_{11}^{-1}E[(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})']) \\
&\quad = \mathrm{tr}(\mathrm{S}_{11}^{-1}E[(\mathbf{Y}_1 - \mathrm{X}_1\mathrm{B}\mathbf{Y}_1)(\mathbf{Y}_1 - \mathrm{X}_1\mathrm{B}\mathbf{Y}_1)']) \\
&\quad = \mathrm{tr}(\mathrm{S}_{11}^{-1}E[((\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathbf{Y}_1)((\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathbf{Y}_1)'])
\end{aligned}$$

But because $\hat{\beta}$ is admissible, $\mathrm{BX}_1 = \mathrm{I}_k$ and:

$$\begin{aligned}
E[(\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathbf{Y}_1] &= (\mathrm{I}_p - \mathrm{X}_1\mathrm{B})E[\mathbf{Y}_1] \\
&= (\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathrm{X}_1\beta \\
&= \mathrm{X}_1\beta - \mathrm{X}_1\mathrm{B}\mathrm{X}_1\beta \\
&= \mathrm{X}_1\beta - \mathrm{X}_1\mathrm{I}_k\beta \\
&= 0
\end{aligned}$$

So:

$$\begin{aligned}
&E[((\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathbf{Y}_1)((\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathbf{Y}_1)'] \\
&\quad = \mathrm{Var}[(\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathbf{Y}_1] \\
&\quad = (\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathrm{Var}[\mathbf{Y}_1](\mathrm{I}_p - \mathrm{X}_1\mathrm{B})' \\
&\quad = (\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathrm{S}_{11}(\mathrm{I}_p - \mathrm{X}_1\mathrm{B})'
\end{aligned}$$

Therefore:

$$
\begin{aligned}
&E[(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})'\mathrm{S}_{11}^{-1}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})] \\
&\quad = \mathrm{tr}(\mathrm{S}_{11}^{-1}E[((\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathbf{Y}_1)((\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathbf{Y}_1)']) \\
&\quad = \mathrm{tr}(\mathrm{S}_{11}^{-1}(\mathrm{I}_p - \mathrm{X}_1\mathrm{B})\mathrm{S}_{11}(\mathrm{I}_p - \mathrm{X}_1\mathrm{B})') \\
&\quad = \mathrm{tr}(\mathrm{S}_{11}^{-1}(\mathrm{S}_{11} - \mathrm{S}_{11}\mathrm{B}'\mathrm{X}_1' - \mathrm{X}_1\mathrm{B}\mathrm{S}_{11} + \mathrm{X}_1\mathrm{B}\mathrm{S}_{11}\mathrm{B}'\mathrm{X}_1')) \\
&\quad = \mathrm{tr}(\mathrm{I}_p - \mathrm{B}'\mathrm{X}_1' - \mathrm{S}_{11}^{-1}\mathrm{X}_1\mathrm{B}\mathrm{S}_{11} + \mathrm{S}_{11}^{-1}\mathrm{X}_1\mathrm{Var}[\hat{\beta}]\mathrm{X}') \\
&\quad = \mathrm{tr}(\mathrm{I}_p) - \mathrm{tr}(\mathrm{B}'\mathrm{X}_1') - \mathrm{tr}(\mathrm{S}_{11}^{-1}\mathrm{X}_1\mathrm{B}\mathrm{S}_{11}) + \mathrm{tr}(\mathrm{S}_{11}^{-1}\mathrm{X}_1\mathrm{Var}[\hat{\beta}]\mathrm{X}')
\end{aligned}
$$

But

$$
\begin{aligned}
\mathrm{tr}(\mathrm{I}_p) &= p, \\
\mathrm{tr}(\mathrm{B}'\mathrm{X}_1') &= \mathrm{tr}(\mathrm{X}_1\mathrm{B}), \\
\mathrm{tr}(\mathrm{S}_{11}^{-1}\mathrm{X}_1\mathrm{B}\mathrm{S}_{11}) &= \mathrm{tr}(\mathrm{X}_1\mathrm{B}\mathrm{S}_{11}\mathrm{S}_{11}^{-1}) \\
&= \mathrm{tr}(\mathrm{X}_1\mathrm{B}) = \mathrm{tr}(\mathrm{B}\mathrm{X}_1) = \mathrm{tr}(\mathrm{I}_k) = k, \qquad \text{and} \\
\mathrm{tr}(\mathrm{S}_{11}^{-1}\mathrm{X}_1\,\mathrm{Var}[\hat{\beta}]\mathrm{X}_1') &= \mathrm{tr}(\mathrm{A}'\mathrm{A}\mathrm{X}_1\,\mathrm{Var}[\hat{\beta}]\mathrm{X}_1') \\
&= \mathrm{tr}(\mathrm{A}\mathrm{X}_1\,\mathrm{Var}[\hat{\beta}]\mathrm{X}_1'\mathrm{A}') \\
&= \mathrm{tr}((\mathrm{A}\mathrm{X}_1)\mathrm{Var}[\hat{\beta}](\mathrm{A}\mathrm{X}_1)').
\end{aligned}
$$

So we arrive at the second equation of Theorem 5.2:

$$
\begin{aligned}
&E[(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})'\mathrm{S}_{11}^{-1}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})] \\
&\quad = \mathrm{tr}(\mathrm{I}_p) - \mathrm{tr}(\mathrm{X}_1\mathrm{B}) - \mathrm{tr}(\mathrm{S}_{11}^{-1}\mathrm{X}_1\mathrm{B}\mathrm{S}_{11}) + \mathrm{tr}(\mathrm{S}_{11}^{-1}\mathrm{X}_1\,\mathrm{Var}[\hat{\beta}]\mathrm{X}') \\
&\quad = p - 2k + \mathrm{tr}((\mathrm{A}\mathrm{X}_1)\mathrm{Var}[\hat{\beta}](\mathrm{A}\mathrm{X}_1)')
\end{aligned}
$$

Then we reduce the second term:

$$\begin{aligned}
&E[(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))'\mathrm{D}'\mathrm{D}(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))] \\
&\quad = E[\mathrm{tr}((\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))'\mathrm{D}'\mathrm{D}(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta})))] \\
&\quad = E[\mathrm{tr}(\mathrm{D}'\mathrm{D}(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))')] \\
&\quad = \mathrm{tr}(E[\mathrm{D}'\mathrm{D}(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))']) \\
&\quad = \mathrm{tr}(\mathrm{D}'\mathrm{D}E[(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))']) \\
&\quad = \mathrm{tr}(\mathrm{D}'\mathrm{D}\,\mathrm{Var}[\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta})]) \\
&\quad = \mathrm{tr}(\mathrm{D}\,\mathrm{Var}[\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta})]\mathrm{D}')
\end{aligned}$$

The next-to-last step follows from the fact that $\mathrm{y}_2(\hat{\beta})$ is an admissible predictor of $\mathbf{Y}_2$ (as Dr. Schmidt states in his fourth section); hence, $E[\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta})] = 0$. But according to Lemma 4.1, $\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}) = \mathrm{D}^{-1}h(\hat{\beta})$ and:

$$\mathrm{Var}[h(\hat{\beta})] = (\mathrm{CX}_1 + \mathrm{DX}_2)\mathrm{Var}[\hat{\beta}](\mathrm{CX}_1 + \mathrm{DX}_2)' + \mathrm{I}_q$$

So by substitution we arrive at the third equation of Theorem 5.2:

$$\begin{aligned}
&E[(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))'\mathrm{D}'\mathrm{D}(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))] \\
&\quad = \mathrm{tr}(\mathrm{D}\,\mathrm{Var}[\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta})]\mathrm{D}') \\
&\quad = \mathrm{tr}(\mathrm{D}\,\mathrm{Var}[\mathrm{D}^{-1}h(\hat{\beta})]\mathrm{D}') \\
&\quad = \mathrm{tr}(\mathrm{Var}[\mathrm{D}\mathrm{D}^{-1}h(\hat{\beta})]) \\
&\quad = \mathrm{tr}(\mathrm{Var}[h(\hat{\beta})]) \\
&\quad = q + \mathrm{tr}((\mathrm{CX}_1 + \mathrm{DX}_2)\mathrm{Var}[\hat{\beta}](\mathrm{CX}_1 + \mathrm{DX}_2)')
\end{aligned}$$

Dr. Schmidt denotes the Gauss–Markov estimator

$$(\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathbf{Y}_1$$

as β^*. Adapting my notation to his, I can restate the last formula of my Appendix A [2, p. 474] as:

$$\begin{aligned}\mathrm{Var}[\hat{\beta}] - \mathrm{Var}[\beta^*] = \{&\mathrm{BA}^{-1} - (\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{A}'\} \\ &\times \{\mathrm{BA}^{-1} - (\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{A}'\}' \geq 0,\end{aligned}$$

where, as above, $\mathrm{A}'\mathrm{A} = \mathrm{S}_{11}^{-1}$ and $\mathrm{BX}_1 = \mathrm{I}_k$. And equality obtains if and only if:

$$\begin{aligned}\mathrm{BA}^{-1} - (\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{A}' &= 0 \\ \mathrm{BA}^{-1} &= (\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{A}' \\ \mathrm{B} &= (\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{A}'\mathrm{A} \\ &= (\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{S}_{11}^{-1}\end{aligned}$$

Therefore, $\mathrm{Var}[\hat{\beta}] - \mathrm{Var}[\beta^*]$ is non-negative definite (or, as Dr. Schmidt calls it, positive semidefinite).[4]

Winding up the optimization problem, we have:

$$\begin{aligned}&E[(\mathbf{Y} - \mathrm{X}\hat{\beta})'\mathrm{S}^{-1}(\mathbf{Y} - \mathrm{X}\hat{\beta})] - E[(\mathbf{Y} - \mathrm{X}\beta^*)'\mathrm{S}^{-1}(\mathbf{Y} - \mathrm{X}\beta^*)] \\ &\quad = E[(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})'\mathrm{S}_{11}^{-1}(\mathbf{Y}_1 - \mathrm{X}_1\hat{\beta})] \\ &\qquad - E[(\mathbf{Y}_1 - \mathrm{X}_1\beta^*)'\mathrm{S}_{11}^{-1}(\mathbf{Y}_1 - \mathrm{X}_1\beta^*)] \\ &\qquad + E[(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))'\mathrm{D}'\mathrm{D}(\mathbf{Y}_2 - \mathrm{y}_2(\hat{\beta}))] \\ &\qquad - E[(\mathbf{Y}_2 - \mathrm{y}_2(\beta^*))'\mathrm{D}'\mathrm{D}(\mathbf{Y}_2 - \mathrm{y}_2(\beta^*))] \\ &\quad = \mathrm{tr}((\mathrm{AX}_1)(\mathrm{Var}[\hat{\beta}] - \mathrm{Var}[\beta^*])(\mathrm{AX}_1)') \\ &\qquad + \mathrm{tr}((\mathrm{CX}_1 - \mathrm{DX}_2)(\mathrm{Var}[\hat{\beta}] - \mathrm{Var}[\beta^*])(\mathrm{CX}_1 - \mathrm{DX}_2)')\end{aligned}$$

[4]See [3, pp. 306–309] for an overview of non-negative definite matrices.

The arguments of the trace functions are non-negative definite matrices, whose diagonal elements must be non-negative. Therefore, the traces are non-negative, and:

$$E[(\mathbf{Y}-\mathrm{X}\hat{\beta})'\mathrm{S}^{-1}(\mathbf{Y}-\mathrm{X}\hat{\beta})]-E[(\mathbf{Y}-\mathrm{X}\beta^*)'\mathrm{S}^{-1}(\mathbf{Y}-\mathrm{X}\beta^*)]\geq 0$$

$$E[(\mathbf{Y}-\mathrm{X}\beta^*)'\mathrm{S}^{-1}(\mathbf{Y}-\mathrm{X}\beta^*)]\leq E[(\mathbf{Y}-\mathrm{X}\hat{\beta})'\mathrm{S}^{-1}(\mathbf{Y}-\mathrm{X}\hat{\beta})]$$

β^* minimizes the expected quadratic loss, though it may not be unique among all admissible estimators of β.

This problem has led Dr. Hamer to define the "generalized Schmidt best (GSB)" estimator as the admissible (i.e., linear-in-$\mathbf{Y}_1$ and unbiased) estimator $\mathbf{P}^*$ that minimizes $E[(\mathbf{Y}_2-\mathbf{P})'\mathrm{W}'\cdot\mathrm{W}(\mathbf{Y}_2-\mathbf{P})]$ over all admissible $\mathbf{P}$, regardless of W.[5] He proves in his Theorem 5.1 that $\mathbf{P}^*$ is GSB if and only if it is the best linear unbiased predictor $\hat{Y}_2$. Therefore, GSB and "uniformly best (UB)" are equivalent. Now the set of admissible estimators in Dr. Schmidt's problem is a subset of the set of those in Dr. Hamer's definition; hence, $\hat{Y}$ will dominate $X\beta^*$ in the optimization of $E[(\mathbf{Y}-\mathbf{P})'\mathrm{W}'\mathrm{W}(\mathbf{Y}-\mathbf{P})]$.

In his Section 6 Dr. Hamer proves that $\mathrm{X}\beta^*$ is best if and only if X_1 is square. I wish to present here another proof. The relevant formulas are:

$$\begin{aligned}\hat{Y}&=\begin{bmatrix}\hat{\mathbf{Y}}_1\\ \hat{\mathbf{Y}}_2\end{bmatrix}\\ &=\begin{bmatrix}\mathbf{Y}_1\\ (\mathrm{S}_{21}\mathrm{S}_{11}^{-1}+(\mathrm{X}_2-\mathrm{S}_{21}\mathrm{S}_{11}^{-1}\mathrm{X}_1)(\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{S}_{11}^{-1})\mathbf{Y}_1\end{bmatrix}\\ &=\begin{bmatrix}\mathrm{I}_p\\ \mathrm{S}_{21}\mathrm{S}_{11}^{-1}+(\mathrm{X}_2-\mathrm{S}_{21}\mathrm{S}_{11}^{-1}\mathrm{X}_1)(\mathrm{X}_1'\mathrm{S}_{11}^{-1}\mathrm{X}_1)^{-1}\mathrm{X}_1'\mathrm{S}_{11}^{-1}\end{bmatrix}\mathbf{Y}_1\end{aligned}$$

[5] I've changed his notation, but not his meaning.

$$X\beta^* = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} (X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} \mathbf{Y}_1$$
$$= \begin{bmatrix} X_1 (X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} \\ X_2 (X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} \end{bmatrix} \mathbf{Y}_1$$

The two estimators are identical (i.e., equal, regardless of the value of $\mathbf{Y}_1$) if and only if $X_1(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} = I_p$ and

$$X_2(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1}$$
$$= S_{21} S_{11}^{-1} + (X_2 - S_{21} S_{11}^{-1} X_1)(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1}.$$

However, if $X_1(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} = I_p$:

$$\begin{aligned} &X_2(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} \\ &\quad = S_{21} S_{11}^{-1} + X_2(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} - S_{21} S_{11}^{-1} I_p \\ &\quad = S_{21} S_{11}^{-1} + X_2(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} \\ &\qquad - S_{21} S_{11}^{-1} X_1 (X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} \\ &\quad = S_{21} S_{11}^{-1} + (X_2 - S_{21} S_{11}^{-1} X_1)(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} \end{aligned}$$

Therefore, the two estimators are identical if and only if $X_1(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} = I_p$.

Now if $X_1(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} = I_p$:

$$\begin{aligned} p &= \operatorname{tr}(I_p) \\ &= \operatorname{tr}(X_1(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1}) \\ &= \operatorname{tr}((X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} X_1) \\ &= \operatorname{tr}(I_k) \\ &= k \end{aligned}$$

And if $p = k$, then since the rank of X_1 is k (guaranteeing that $X_1' S_{11}^{-1} X_1$ has an inverse), X_1 has an inverse. And:

$$\begin{aligned} X_1(X_1' S_{11}^{-1} X_1)^{-1} X_1' S_{11}^{-1} &= X_1 (X_1)^{-1} (S_{11}^{-1})^{-1} (X_1')^{-1} X_1' S_{11}^{-1} \\ &= I_p (S_{11}^{-1})^{-1} I_p S_{11}^{-1} \\ &= (S_{11}^{-1})^{-1} S_{11}^{-1} \\ &= I_p \end{aligned}$$

So $X\beta^*$ is best if and only if X_1 is square, in which case the observations constitute a system of simultaneous equations that has the unique solution $\beta^* = X_1^{-1} \mathbf{Y}_1$.