The Gauss-Markov Theorem: Beyond the BLUE

Leigh J. Halliwell, FCAS, MAAA

Abstract: Until now the Gauss-Markov theorem has been the handmaid of least squares; it has served as a proof that the least-squares method produces the Best Linear Unbiased Estimator (BLUE). This theoretical paper shows that it can be, and should be, reformulated as the solution to the problem of the minimization of a quadratic form subject to a linear constraint. The whole theory of linear statistical modeling, from basic to complicated, receives a clean and efficient development on the basis of this reformulation; estimates and predictions based thereon are BLUE from the start, rather than BLUE by subsequent proof. With an intermediate-level background in matrix algebra the reader will understand the frequent interpretations of this development in terms of an n-dimensional projective geometry. Because this paper elevates BLUE to its true role, "Beyond the BLUE" really means "To the True BLUE."

Keywords: Gauss-Markov, BLUE, linear model, projection, distance metric

1. INTRODUCTION

The many treatments of the Gauss-Markov theorem (e.g., Judge [1988, 202-206], Halliwell [2007, Appendix B], and Wikipedia) lead one to believe that the theorem is no more than a proof that $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$ is best linear unbiased estimator (BLUE) of β in the model $\mathbf{y} = X\beta + \mathbf{e}$, where $Var[\mathbf{e}] = \sigma^2 I$. In this capacity the theorem is impressive enough; however, with a little abstraction it becomes much more, as we shall see in the following eleven sections.

2. STATEMENT OF THE THEOREM AND ITS PROOF

The Gauss-Markov theorem is essentially the solution to a constrained-optimization problem, more exactly, to the problem of minimizing a quadratic form subject to a linear constraint. Here is our formulation of the theorem:

> The Gauss-Markov Theorem: If symmetric $\Sigma_{n\times n}$ is positive-definite and $A_{m\times n}$ is of full-row rank, then $\Phi(W) = W'\Sigma^{-1}W$ can be minimized subject to the linear constraint $A_{m\times n}W_{n\times p} = B_{m\times p}$. The value $W^* = \Sigma A'(A\Sigma A')^{-1}B$ uniquely minimizes Φ at $\Phi(W^*) = B'(A\Sigma A')^{-1}B$.

To prove the theorem, we take for granted two theorems about positive-definite matrices.[1] First, positive-definite matrices have inverses; the inverses also are positive-definite. Therefore, symmetric Σ^{-1} exists, and is positive-definite. Second, if $\mathrm{A}_{m\times n}$ is of full-row rank and $\mathrm{T}_{n\times n}$ is positive-definite, then ATA' is positive-definite. From these it follows that $\mathrm{A\Sigma A}'$ is positive-definite and invertible; hence, $\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A\Sigma A}')^{-1}\mathrm{B}$ exists. Moreover, W^* satisfies the constraint, since $\mathrm{AW}^* = \mathrm{A\Sigma A}'(\mathrm{A\Sigma A}')^{-1}\mathrm{B} = \mathrm{I}_m\mathrm{B} = \mathrm{B}$.

Now if W_1 satisfies the constraint, then:

$$\begin{aligned}\mathrm{W}_1'\Sigma^{-1}\mathrm{W}^* &= \mathrm{W}_1'\Sigma^{-1}\Sigma\mathrm{A}'(\mathrm{A\Sigma A}')^{-1}\mathrm{B}\\ &= \mathrm{W}_1'\mathrm{A}'(\mathrm{A\Sigma A}')^{-1}\mathrm{B}\\ &= (\mathrm{AW}_1)'(\mathrm{A\Sigma A}')^{-1}\mathrm{B}\\ &= \mathrm{B}'(\mathrm{A\Sigma A}')^{-1}\mathrm{B}\end{aligned}$$

And since W^* is an allowable instance of W_1, we have the following chain of equalities:

$$\mathrm{W}_1'\Sigma^{-1}\mathrm{W}^* = \mathrm{B}'(\mathrm{A\Sigma A}')^{-1}\mathrm{B} = \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^* = \left(\mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*\right)' = \left(\mathrm{W}_1'\Sigma^{-1}\mathrm{W}^*\right)' = \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}_1$$

As the heart of the Gauss-Markov proof:

$$\begin{aligned}\Phi(\mathrm{W}_1)-\Phi(\mathrm{W}^*) &= \mathrm{W}_1'\Sigma^{-1}\mathrm{W}_1 - \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*\\ &= \mathrm{W}_1'\Sigma^{-1}\mathrm{W}_1 - \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^* - \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^* + \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*\\ &= \mathrm{W}_1'\Sigma^{-1}\mathrm{W}_1 - \mathrm{W}_1'\Sigma^{-1}\mathrm{W}^* - \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}_1 + \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*\\ &= (\mathrm{W}_1-\mathrm{W}^*)'\Sigma^{-1}(\mathrm{W}_1-\mathrm{W}^*)\\ &\geq 0_{p\times p}\end{aligned}$$

[1] For a review of positive-definite and non-negative-definite (or positive-semi-definite) matrices see Judge [1988, Appendix A.14] and Halliwell [1997, Appendix A].

The last line is to be taken in a matrix-definite sense, viz., that the difference $\Phi(W_1)-\Phi(W^*)$ is the non-negative-definite matrix $(W_1-W^*)'\Sigma^{-1}(W_1-W^*)$. And because Σ^{-1} is positive-definite, the difference equals the zero matrix ($0_{p\times p}$) if and only if $W_1=W^*$. Therefore, $W^*=\Sigma A'(A\Sigma A')^{-1}B$ uniquely minimizes $\Phi(W)=W'\Sigma^{-1}W$ subject to $AW=B$. Furthermore, the minimum is $\Phi(W^*)=B'(A\Sigma A')^{-1}B$.

3. GEOMETRICAL INTERPRETATION WITH A DISTANCE METRIC

A geometrical interpretation of the theorem will prove helpful. Again, let W_1 satisfy the constraint, and let $W^*=\Sigma A'(A\Sigma A')^{-1}B$. From the chain of equalities, we derive:

$$(W_1-W^*)'\Sigma^{-1}W^* = W_1'\Sigma^{-1}W^* - W^{*\prime}\Sigma^{-1}W^* = 0_{p\times p} = (0_{p\times p})' = W^{*\prime}\Sigma^{-1}(W_1-W^*)$$

These are unusual quadratic forms. The usual quadratic form is $y'\Sigma x$, where the factors before and after Σ^{-1} are $1\times n$ and $n\times 1$ vectors. Here the form is $Y'\Sigma^{-1}X$, where the factors before and after Σ^{-1} are $p\times n$ and $n\times p$ matrices, and the integer p may exceed one.

But for now, consider the usual quadratic form in the special case that $\Sigma=I_n$. Actuaries know that $x'I_nx=x'x=\sum_{i=1}^{n}x_i^2$ is the square of the distance from the origin of $\Re^n$ to x (or the area of a square the length of whose sides is that distance). Less well known is that $y'x=x'y=\sum_{i=1}^{n}x_iy_i$ represents the area of a rectangle, the length of one of whose sides is the length of the projection of one vector onto the other. Most will recognize, however, an equivalent interpretation, viz., that $\mathrm{y}'\mathrm{x} = \mathrm{x}'\mathrm{y} = 0$ if and only if $\mathrm{x} \perp \mathrm{y}$. The standard (Euclidean) definition of the distance from x to y is $d(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{y}-\mathrm{x})'(\mathrm{y}-\mathrm{x})}$. It has the three properties of a metric on $\Re^n$:

$$\begin{array}{lll} 1. & d(\mathrm{x},\mathrm{y}) \geq 0;\, d(\mathrm{x},\mathrm{y}) = 0 \Leftrightarrow \mathrm{x} = \mathrm{y} & \text{non – negativity; trivially zero} \\ 2. & d(\mathrm{y},\mathrm{x}) = d(\mathrm{x},\mathrm{y}) & \text{symmetry} \\ 3. & d(\mathrm{x},\mathrm{y}) + d(\mathrm{y},\mathrm{z}) \geq d(\mathrm{x},\mathrm{z}) & \text{triangle inequality} \end{array}$$

But for any positive-definite matrix $\Sigma_{n\times n}$, one can define a valid "Σ metric" on $\Re^n$ as $d_\Sigma(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{y}-\mathrm{x})'\Sigma^{-1}(\mathrm{y}-\mathrm{x})}$, which is valid in that it possesses these three properties.[2] The matrix Σ represents a combination of scaling and rotating the axes of $\Re^n$.

So what is special in the Gauss-Markov theorem about $\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$? Adapting the concept of perpendicularity to a metric, we have:

$$(\mathrm{W}_1 - \mathrm{W}^*)'\Sigma^{-1}\mathrm{W}^* = 0_{p\times p}$$

[2] Some confusion results from using the inverse of Σ in the quadratic form; one must think twice to determine whether something is a Σ metric or a Σ^{-1} metric. However, consider the usual formula for the ellipse whose major semi-axis is two units and minor semi-axis is one: $(x_1/2)^2 + (x_2/1)^2 = 1^2$. As a quadratic form this would be:

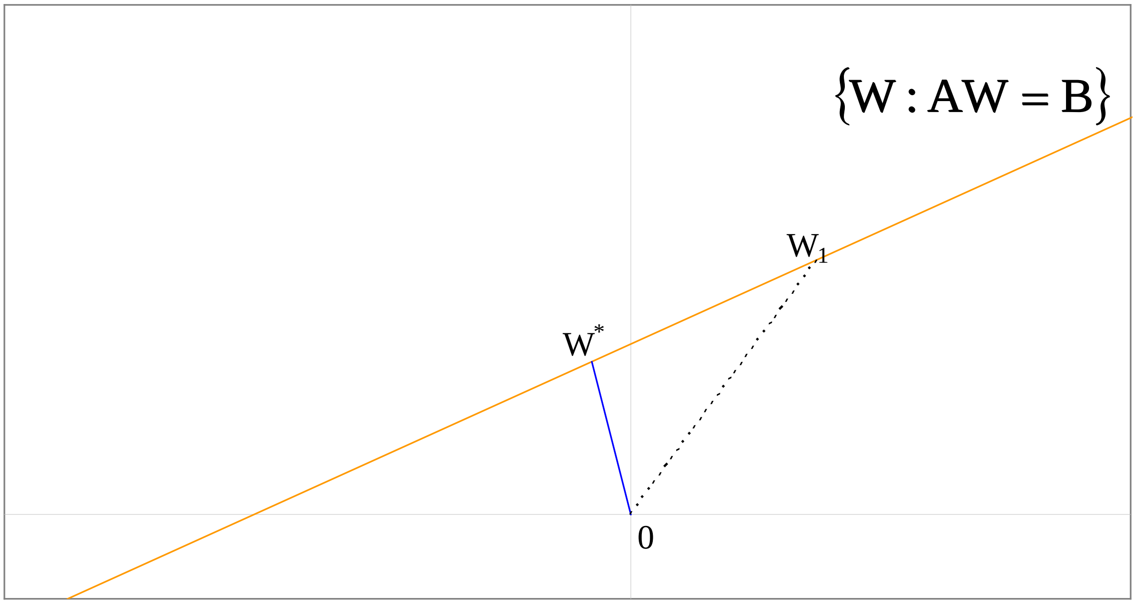
$$[x_1 \quad x_2]\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}^{-1}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1^2$$

It seemed more natural to call this a [2 0; 0 1] metric (as if to say, "Two units on the first axis count as one unit on the second."), rather than to call it a [½ 0; 0 1] metric. This ellipse is the set of points in $\Re^2$ whose distance from the origin is one unit according to the [2 0; 0 1] metric. It may help some readers to know that $d_\Sigma(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{y}-\mathrm{x})'\Sigma^{-1}(\mathrm{y}-\mathrm{x})}$ is called the "Mahalanobis distance" (cf. Wikipedia), in whose definition the Σ matrix is inverted. Appendix A provides a proof of the triangle inequality, as well as a justification of the geometric interpretation of $\mathrm{x}'\mathrm{y}$ as the product of the length of x and the length of the projection of y onto x.

This means that according to the Σ metric W^* is perpendicular to $W_1 - W^*$ (and vice versa). In mathematical notation, $W^* \underset{\Sigma}{\perp} \left(W_1 - W^*\right)$. The heart of the Gauss-Markov theorem, expressed above as $\Phi(W_1) - \Phi(W^*) = \left(W_1 - W^*\right)' \Sigma^{-1}\left(W_1 - W^*\right)$, is really just the Pythagorean theorem adapted to the Σ metric:

$$W^{*'}\Sigma^{-1}W^* + \left(W_1 - W^*\right)'\Sigma^{-1}\left(W_1 - W^*\right) = W_1'\Sigma^{-1}W_1$$

W^* is the element of the constraint set closest to the origin according to the Σ metric. The following diagram clarifies this:



The orange line represents the constraint set.[3] The origin, W^*, and W_1 form the Σ-right triangle, of which W_1 is the hypotenuse, and W^* and W_1 - W^* are the legs. The salient point is that the Σ area of the square with side W^* is less than or equal to that of the square with side W_1, or $W^{*'}\Sigma W^* \le W_1'\Sigma W_1$, and equal if and only if $W_1 = W^*$. This is valid even

[3] Since the constraint on W is linear, the constraint set is a hyperplane (technically, an affine apace). The Gauss-Markov theorem requires a *linear* constraint; constraints involving curvature are inadmissible.

when the area concept is abstracted from a non-negative scalar to a non-negative-definite matrix.

This ends the geometric interpretation. Gauss-Markov reasoning happens whenever a quadratic form is to be minimized subject to a linear constraint. Gauss-Markov/BLUE proofs are abstractions of what we all learned in plane Geometry, viz., that the shortest distance from a point to a straight line is along a line segment perpendicular to the line. Lines are abstracted into linear constraints and distance is abstracted into a Σ metric.

It is hardly necessary to memorize the formula for W^*. With the following heuristic reasoning one can derive it on the fly. Since $A_{m\times n}$ is of full-row rank (or of rank *m*), the $m \times m$ matrix AA' is invertible. In fact, as stated above, for any positive-definite $T_{n\times n}$, ATA' is invertible. Thus, there is a family of "right inverses" of A that have the form $TA'(ATA')^{-1}$. W^* will be the matrix product of one of these right inverses and B, i.e., $W^* = TA'(ATA')^{-1}B$. Since we seek to minimize $W'\Sigma^{-1}W$, distance is measured according to a $T = \Sigma$ metric. According to this metric $W^* = \Sigma A'(A\Sigma A')^{-1}B$ is the element of the constraint set closest to the origin.

4. PROJECTION INTO THE CONSTRAINT SPACE

In the interest of economy and precision, let us introduce some more formalism. Our 'W' variables denote elements of $\Re^{n\times p}$, the real space of $n \times p$ dimensions. Let us use 'Ω' to denote the constraint set: $\Omega = \{W \in \Re^{n\times p} : AW = B\}$. Obviously, $\Omega \subseteq \Re^{n\times p}$; but it is not empty under the assumption that $A_{m\times n}$ is of full-row rank. In fact, we have just seen that

$\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$ is the element of Ω closest to the origin of $\Re^{n\times p}$ according to the Σ metric. We may say that $\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$ is the Σ projection of the origin into Ω. In general, what is the Σ projection of *any* element of $\Re^{n\times p}$ into Ω?

Using '*P*' for projection, we define $P(\mathrm{W}_0; \Omega, \Sigma)$ as the function which projects $\mathrm{W}_0 \in \Re^{n\times p}$ into Ω according to the Σ metric. As before, Ω is the non-empty solution set of the linear constraint $\mathrm{AW} = \mathrm{B}$, and Σ is positive-definite. When these parameters are understood, we will use the abbreviation $P(\mathrm{W}_0)$. So $P(\mathrm{W}_0; \Omega, \Sigma)$ is *an* element of Ω that minimizes the Σ-metric distance from W_0 to Ω. Equivalently, it minimizes the quadratic form $\Phi(\mathrm{W}) = (\mathrm{W} - \mathrm{W}_0)'\Sigma^{-1}(\mathrm{W} - \mathrm{W}_0)$ subject to $\mathrm{AW} = \mathrm{B}$.

We could argue from scratch as in Section 2, but the following analysis is more insightful. The constraint $\mathrm{AW} = \mathrm{B}$ is equivalent to $\mathrm{A}(\mathrm{W} - \mathrm{W}_0) = \mathrm{B} - \mathrm{AW}_0$. So the projection problem is to minimize $(\mathrm{W} - \mathrm{W}_0)'\Sigma^{-1}(\mathrm{W} - \mathrm{W}_0)$ subject to $\mathrm{A}(\mathrm{W} - \mathrm{W}_0) = \mathrm{B} - \mathrm{AW}_0$. This is the Gauss-Markov problem with two changes in variables:

$$\begin{aligned} \mathrm{W} &\rightarrow \mathrm{W} - \mathrm{W}_0 \\ \mathrm{B} &\rightarrow \mathrm{B} - \mathrm{AW}_0 \end{aligned}$$

Hence, the Gauss-Markov theorem states that $(\mathrm{W} - \mathrm{W}_0)^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B} - \mathrm{AW}_0)$ uniquely minimizes $\Phi(\mathrm{W} - \mathrm{W}_0) = (\mathrm{W} - \mathrm{W}_0)'\Sigma^{-1}(\mathrm{W} - \mathrm{W}_0)$. But since W_0 is a constant, $(\mathrm{W} - \mathrm{W}_0)^* = \mathrm{W}^* - \mathrm{W}_0$, or $\mathrm{W}^* = \mathrm{W}_0 + (\mathrm{W} - \mathrm{W}_0)^* = \mathrm{W}_0 + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B} - \mathrm{AW}_0)$. So there is not just *an* element of projection, but a unique element:

$$\begin{aligned} P(\mathrm{W};\Omega,\Sigma) &= \mathrm{W}+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B}-\mathrm{AW}) \\ &= \{I_n-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathrm{W}+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B} \end{aligned}$$

As a check:

$$\begin{aligned} \mathrm{A}P(\mathrm{W};\Omega,\Sigma) &= \mathrm{A}\{\mathrm{I}_n-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathrm{W}+\mathrm{A}\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B} \\ &= \{\mathrm{A}-\mathrm{A}\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathrm{W}+\mathrm{A}\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B} \\ &= \{\mathrm{A}-\mathrm{I}_m\mathrm{A}\}\mathrm{W}+\mathrm{I}_m\mathrm{B} \\ &= \mathrm{B} \end{aligned}$$

Hence, for all $\mathrm{W}\in\Re^{n\times p}$, $\mathrm{A}P(\mathrm{W})\in\Omega$. So P is a mapping from $\Re^{n\times p}$ into Ω, i.e., $P:\Re^{n\times p}\to\Omega$. In particular, the mapping of the origin is:

$$P(0_{n\times p})=\{\mathrm{I}_n-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}0+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}=\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B},$$

which is the $'\mathrm{W}^*'$ of the theorem itself. Accordingly, we may employ the formulation $P(\mathrm{W})=\{\mathrm{I}_n-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathrm{W}+P(0)$.

P maps element $\mathrm{W}\in\Re^{n\times p}$ to the closest element of constraint set Ω according to the Σ metric. Geometrically, *P* sends a Σ perpendicular from W into Ω; in symbols, $P(\mathrm{W})-\mathrm{W}\underset{\Sigma}{\perp}\mathrm{W}_1-P(\mathrm{W})$, for every $\mathrm{W}_1\in\Omega$, as the following algebra shows:

$$\begin{aligned}
&(P(\mathrm{W})-\mathrm{W})'\Sigma^{-1}(\mathrm{W}_1-P(\mathrm{W}))\\
&\quad=\left(\left\{\mathrm{I}_n-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right\}\mathrm{W}+P(0)-\mathrm{W}\right)'\Sigma^{-1}(\mathrm{W}_1-P(\mathrm{W}))\\
&\quad=\left(P(0)-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\mathrm{W}\right)'\Sigma^{-1}(\mathrm{W}_1-P(\mathrm{W}))\\
&\quad=\left(\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\mathrm{W}\right)'\Sigma^{-1}(\mathrm{W}_1-P(\mathrm{W}))\\
&\quad=\left(\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B}-\mathrm{A}\mathrm{W})\right)'\Sigma^{-1}(\mathrm{W}_1-P(\mathrm{W}))\\
&\quad=(\mathrm{B}-\mathrm{A}\mathrm{W})'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\Sigma^{-1}(\mathrm{W}_1-P(\mathrm{W}))\\
&\quad=(\mathrm{B}-\mathrm{A}\mathrm{W})'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}(\mathrm{W}_1-P(\mathrm{W}))\\
&\quad=(\mathrm{B}-\mathrm{A}\mathrm{W})'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{A}\mathrm{W}_1-\mathrm{A}P(\mathrm{W}))\\
&\quad=(\mathrm{B}-\mathrm{A}\mathrm{W})'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B}-\mathrm{B})\\
&\quad=0
\end{aligned}$$

Because of the first property of a metric (zero-triviality), within the restricted domain Ω, *P* is the identity mapping. Hence, not only is *P* a mapping *into* the constraint set Ω; it is also a mapping *onto* Ω. Nonetheless, we will prove it algebraically. If $\mathrm{W}\in\Omega$:

$$\begin{aligned}
P(\mathrm{W})&=\left\{\mathrm{I}_n-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right\}\mathrm{W}+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}\\
&=\mathrm{W}-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{A}\mathrm{W})+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}\\
&=\mathrm{W}-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B})+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}\\
&=\mathrm{W}
\end{aligned}$$

Conversely, if $P(\mathrm{W})=\mathrm{W}$, then $\mathrm{A}\mathrm{W}=\mathrm{A}P(\mathrm{W})=\mathrm{B}$ and $\mathrm{W}\in\Omega$. Therefore, *P* is a many-to-one mapping from $\Re^{n\times p}$ onto constraint set Ω, and an element of $\Re^{n\times p}$ belongs to Ω if and only if *P* acts upon it as an identity mapping.

Just as the Σ-metric right inverse $\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}$ is conspicuous in the formula $P(0_{n\times p}) = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$, so too is it conspicuous in the formula for what we will call the "Σ-projection matrix" $\mathrm{I}_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}$.[4] Since Σ is positive-definite, it can be Cholesky-decomposed as $\Sigma = \mathrm{QQ}'$ for some non-singular $\mathrm{Q}_{n\times n}$. Then the matrix can be factored as $\mathrm{I}_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A} = \mathrm{Q}\{\mathrm{I}_n - \mathrm{Q}'\mathrm{A}'(\mathrm{AQQ}'\mathrm{A}')^{-1}\mathrm{AQ}\}\mathrm{Q}^{-1} = \mathrm{QMQ}^{-1}$. So the rank of the matrix is the rank of $\mathrm{M} = \mathrm{I}_n - \mathrm{Q}'\mathrm{A}'(\mathrm{AQQ}'\mathrm{A}')^{-1}\mathrm{AQ}$. But M is a (symmetric) idempotent matrix (i.e., $\mathrm{M} = \mathrm{MM}'$), and the rank of an idempotent matrix equals its trace (Judge [1988, Appendix A.4 and A.12] and Halliwell [1997, Appendix B, 317; also Note 3]). Employing basic theorems about the trace operator, we derive:

$$\begin{aligned}
rank\left(\mathrm{I}_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right) &= rank\left(\mathrm{Q}\{\mathrm{I}_n - \mathrm{Q}'\mathrm{A}'(\mathrm{AQQ}'\mathrm{A})^{-1}\mathrm{AQ}\}\mathrm{Q}^{-1}\right) \\
&= rank\left(\mathrm{I}_n - \mathrm{Q}'\mathrm{A}'(\mathrm{AQQ}'\mathrm{A})^{-1}\mathrm{AQ}\right) \\
&= Tr\left(\mathrm{I}_n - \mathrm{Q}'\mathrm{A}'(\mathrm{AQQ}'\mathrm{A})^{-1}\mathrm{AQ}\right) \\
&= Tr(\mathrm{I}_n) - Tr\left(\mathrm{Q}'\mathrm{A}'(\mathrm{AQQ}'\mathrm{A})^{-1}\mathrm{AQ}\right) \\
&= Tr(\mathrm{I}_n) - Tr\left((\mathrm{AQQ}'\mathrm{A})^{-1}\mathrm{AQQ}'\mathrm{A}'\right) \\
&= Tr(\mathrm{I}_n) - Tr(\mathrm{I}_m) \\
&= n - m
\end{aligned}$$

So every column of the elements of constraint set Ω has m fewer degrees of freedom than the columns of $\Re^{n\times p}$; in a sense, the dimensionality of Ω is $\Re^{(n-m)\times p}$. This can be surmised from the full-row rank of $\mathrm{A}_{m\times n}$, which imposes m independent restrictions on the elements of $\Re^{n\times p}$ that belong to Ξ. This will prove useful in Section 9, in which we will treat linear statistical models with parameter constraints.

[4] The projection matrix shows to its greatest effect in the homogeneous form, i.e., in the differential form, $P(\mathrm{W}_2) - P(\mathrm{W}_1) = \{\mathrm{I}_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}(\mathrm{W}_2 - \mathrm{W}_1)$.

5. VARIANCE AS A METRIC

In this section we will show how the variance of a random vector serves as its natural metric. Let **x** be an $n \times 1$ random vector with mean μ and non-degenerate variance Σ. Since the variance is non-degenerate, the variance of every non-zero linear combination of its elements is positive, i.e., Σ is positive-definite and Σ^{-1} exists. The variance of a random vector is a measure of its ability to differ from its mean. So the distances of random vectors from their means should somehow be invariant, when their variances serve as their distance metrics.

The square of the Σ-metric distance of **x** from its mean is $d^2_\Sigma(\mu, \mathbf{x}) = (\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)$. And by definition, the variance of **x** is $Var[\mathbf{x}] = \Sigma = E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right]$. Using again the trace-operator theorems of the previous section, we find:

$$\begin{aligned}
E\left[d^2_\Sigma(\mu,\mathbf{x})\right] &= E\left[(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right] \\
&= Tr\left(E\left[(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right]\right) \\
&= E\left[Tr\left((\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right)\right] \\
&= E\left[Tr\left(\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right)\right] \\
&= Tr\left(E\left[\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right]\right) \\
&= Tr\left(\Sigma^{-1}E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right]\right) \\
&= Tr\left(\Sigma^{-1}\Sigma\right) \\
&= Tr(I_n) = n
\end{aligned}$$

Hence, a random vector's variance is its natural metric, according to which its expected squared distance from its mean equals its dimensionality *n*, the degrees of its freedom.[5]

6. PROJECTIONS OF RANDOM VECTORS

As in the previous section, let **x** be an *n*×1 random vector with mean μ and variance Σ, i.e., $\mathbf{x} \sim (\mu, \Sigma)$. Since the Gauss-Markov theorem has to do with abstract projective geometry, we may inquire about the moments of the Σ projection of **x** into the constraint space $\Omega = \left\{ \mathrm{x} \in \Re^n : \mathrm{A}_{m\times n}\mathrm{x}_{n\times 1} = \mathrm{b}_{m\times 1} \right\}$.

The Σ projection is $P(\mathbf{x};\Omega,\Sigma) = \left\{ I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A} \right\}\mathbf{x} + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b}$. Therefore:

$$\begin{aligned} E[P(\mathbf{x};\Omega,\Sigma)] &= E\left[\left\{ I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A} \right\}\mathbf{x} + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b}\right] \\ &= \left\{ I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A} \right\}E[\mathbf{x}] + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b} \\ &= \left\{ I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A} \right\}\mu + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b} \\ &= \mu + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{b} - \mu) \end{aligned}$$

The variance follows from the standard formula $Var[\mathrm{Q}\mathbf{x}] = \mathrm{Q}Var[\mathbf{x}]\mathrm{Q}'$:

[5] Moreover, if **x** is multivariate normal, or if **x** ~ N(μ, Σ), then $(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu) \sim \chi^2_n$ (Judge [1988], §2.5.9). The multivariate normal distribution is unique in that its probability distribution is a function of its variance metric: $f_{\mathbf{x}}(\mathrm{x}) \propto e^{-\frac{1}{2}(\mathrm{x}-\mu)'\Sigma^{-1}(\mathrm{x}-\mu)}$. Most likely, this is the ultimate reason why normality is preserved under any linear transformation.

$$
\begin{aligned}
Var[P(\mathbf{x};\Omega,\Sigma)] &= Var\left[\left\{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right\}\mathbf{x} + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b}\right] \\
&= Var\left[\left\{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right\}\mathbf{x}\right] \\
&= \left\{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right\}Var[\mathbf{x}]\left\{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right\}' \\
&= \left\{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right\}\Sigma\left\{I_n - \mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\right\} \\
&= \Sigma - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma \\
&= \Sigma - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma \\
&= \Sigma - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma
\end{aligned}
$$

Therefore, if $\mathbf{x} \sim (\mu, \Sigma)$, then $P(\mathbf{x};\Omega,\Sigma) \sim \left(P(\mu;\Omega,\Sigma), \Sigma - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\right)$. As a check, $E[\mathrm{A}\mathbf{x}] = \mathrm{A}P(\mu) = \mathrm{b}$ and $Var[\mathrm{A}\mathbf{x}] = \mathrm{A}\left(\Sigma - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\right)\mathrm{A}' = 0_{m\times m}$. This will prove useful in Section 11.

7. PARAMETER ESTIMATION IN THE LINEAR STATISTICAL MODEL

Now let us apply our Gauss-Markov theorem to the linear statistical model. First, and as an easy start, we will apply it to derive the best linear unbiased estimator (BLUE) of the parameter β in the model $\mathbf{y} = \mathrm{X}_{t\times k}\beta_{k\times 1} + \mathbf{e}$, where $Var[\mathbf{e}] = \Sigma_{t\times t}$. X is of full-column rank, and Σ is positive-definite. The estimator is linear in $\mathbf{y}$, or $\hat{\boldsymbol{\beta}} = \mathrm{W}'\mathbf{y}$.[6] Because it is unbiased for all β, $E\left[\hat{\boldsymbol{\beta}}\right] = E[\mathrm{W}'\mathbf{y}] = \mathrm{W}'\mathrm{X}\beta = \beta$. So matrix W' is constrained according to the equation $\mathrm{W}'\mathrm{X} = \mathrm{I}_k$, which transposes as $\mathrm{X}'\mathrm{W} = \mathrm{I}_k$. The best of the unbiased estimators minimizes $Var\left[\hat{\boldsymbol{\beta}}\right] = Var[\mathrm{W}'\mathbf{y}] = \mathrm{W}'\Sigma\mathrm{W}$. So the problem is to minimize

[6] For a reason immediately to become apparent, we use here the transpose of W.

$\Phi(\mathrm{W}) = \mathrm{W}'\Sigma\mathrm{W} = \mathrm{W}'\left(\Sigma^{-1}\right)^{-1}\mathrm{W}$ subject to $\mathrm{X}'\mathrm{W} = \mathrm{I}_k$. The correspondences between the theorem and this model are:

Theorem	←	Model
Σ	←	Σ^{-1}
W	←	W
A	←	X'
B	←	I_k

X' is of full-row rank, because X is of full-column rank. Hence, according to the theorem, $\mathrm{W}^* = \Sigma^{-1}\mathrm{X}''\left(\mathrm{X}'\Sigma^{-1}\mathrm{X}''\right)^{-1}\mathrm{I}_k = \Sigma^{-1}\mathrm{X}\left(\mathrm{X}'\Sigma^{-1}\mathrm{X}\right)^{-1}$. So $\hat{\boldsymbol{\beta}} = \mathrm{W}^{*\prime}\mathbf{y} = \left(\mathrm{X}'\Sigma^{-1}\mathrm{X}\right)^{-1}\mathrm{X}'\Sigma^{-1}\mathbf{y}$ and $Var\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{W}^{*\prime}\Sigma\mathrm{W}^* = \mathrm{I}_k'\left(\mathrm{X}'\Sigma^{-1}\mathrm{X}''\right)^{-1}\mathrm{I}_k = \left(\mathrm{X}'\Sigma^{-1}\mathrm{X}\right)^{-1}$. Accordingly, $\hat{\boldsymbol{\beta}} = Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{X}'\Sigma^{-1}\mathbf{y}$.

8. PREDICTION IN THE LINEAR STATISTICAL MODEL

The goal of most linear modeling is not to estimate the parameter β, but rather to estimate[7] quantities which eventually will be observed. Although the model makes such quantities dependent on the parameter, the parameter itself is usually hypothetical and never to be observed. With partitioning between the observed $\mathbf{y}_1$ and the to-be-predicted $\mathbf{y}_2$ (hence, containing missing values) the general form of the linear statistical model is:

$$\begin{bmatrix}\mathbf{y}_1\\ \hdashline \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathrm{X}_1 \; {\scriptstyle (t_1\times k)}\\ \hdashline \mathrm{X}_2 \; {\scriptstyle (t_2\times k)}\end{bmatrix}\beta_{k\times 1} + \begin{bmatrix}\mathbf{e}_1\\ \hdashline \mathbf{e}_2\end{bmatrix}, \text{ where } Var\begin{bmatrix}\mathbf{e}_1\\ \hdashline \mathbf{e}_2\end{bmatrix} = \left[\begin{array}{c:c}\Sigma_{11} & \Sigma_{12}\\ \hdashline \Sigma_{21} & \Sigma_{22}\end{array}\right] = \Sigma$$

Not only must Σ be symmetric and non-negative-definite, Σ_{11} must be positive-definite, and X_1 must be of full-column rank. We seek the best linear-in-$\mathbf{y}_1$, unbiased estimator

[7] More accurately, the goal is to predict – we seek the best linear unbiased prediction. But we will continue to call this BLUE, because BLUP already has a different technical meaning in statistics (*Wikipedia*, "Best linear unbiased prediction").

(BLUE) of $\mathbf{y}_2$, i.e., $\hat{\mathbf{y}}_2 = \mathrm{W}'\mathbf{y}_1$ for some matrix W', which depends only on the partitions of the design X and variance Σ matrices. Because the estimator is unbiased for all β, $0 = E[\mathbf{y}_2 - \hat{\mathbf{y}}_2] = E[\mathbf{y}_2 - \mathrm{W}'\mathbf{y}_1] = (\mathrm{X}_2 - \mathrm{W}'\mathrm{X}_1)\beta$. Thus the estimator is unbiased if and only if $\mathrm{W}'\mathrm{X}_1 = \mathrm{X}_2$. By transposition, $\mathrm{X}_1'\mathrm{W} = \mathrm{X}_2'$, where W is $t_1 \times t_2$.

But now there is a complication in being "best." Predicting $\mathbf{y}_2$ as $\hat{\mathbf{y}}_2$, we will err by the amount $\mathbf{y}_2 - \hat{\mathbf{y}}_2$. So it is the prediction-error variance that we must minimize:

$$
\begin{aligned}
Var[\mathbf{y}_2 - \hat{\mathbf{y}}_2] &= Var[\mathbf{y}_2 - \mathrm{W}'\mathbf{y}_1] \\
&= Var\left[\begin{bmatrix} -\mathrm{W}' & I_{t_2} \end{bmatrix}\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}\right] \\
&= \begin{bmatrix} -\mathrm{W}' & I_{t_2} \end{bmatrix} Var\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \begin{bmatrix} -\mathrm{W}' & I_{t_2} \end{bmatrix}' \\
&= \begin{bmatrix} -\mathrm{W}' & I_{t_2} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} -\mathrm{W} \\ I_{t_2} \end{bmatrix} \\
&= \mathrm{W}'\Sigma_{11}\mathrm{W} - \mathrm{W}'\Sigma_{12} - \Sigma_{21}\mathrm{W} + \Sigma_{22}
\end{aligned}
$$

Although we can ignore the constant Σ_{22} in the minimization, we cannot ignore the second and third terms, which are linear in W.

The key here is to apply the one-to-one transform $\mathrm{V} \leftrightarrow \mathrm{W} - \Sigma_{11}^{-1}\Sigma_{12}$. The transformation of the constraint set is $\{\mathrm{W} : \mathrm{X}_1'\mathrm{W} = \mathrm{X}_2'\} = \{\mathrm{V} : \mathrm{X}_1'\mathrm{V} = \mathrm{X}_1'\mathrm{W} - \mathrm{X}_1'\Sigma_{11}^{-1}\Sigma_{12} = \mathrm{X}_2' - \mathrm{X}_1'\Sigma_{11}^{-1}\Sigma_{12}\}$. So expressed in terms of V:

$$
\begin{aligned}
Var[\mathbf{y}_2 - \hat{\mathbf{y}}_2] &= W'\Sigma_{11}W - W'\Sigma_{12} - \Sigma_{21}W + \Sigma_{22} \\
&= \left(V + \Sigma_{11}^{-1}\Sigma_{12}\right)'\Sigma_{11}\left(V + \Sigma_{11}^{-1}\Sigma_{12}\right) - \left(V + \Sigma_{11}^{-1}\Sigma_{12}\right)'\Sigma_{12} - \Sigma_{21}\left(V + \Sigma_{11}^{-1}\Sigma_{12}\right) + \Sigma_{22} \\
&= \left(V' + \Sigma_{21}\Sigma_{11}^{-1}\right)\Sigma_{11}\left(V + \Sigma_{11}^{-1}\Sigma_{12}\right) - \left(V' + \Sigma_{21}\Sigma_{11}^{-1}\right)\Sigma_{12} - \Sigma_{21}\left(V + \Sigma_{11}^{-1}\Sigma_{12}\right) + \Sigma_{22} \\
&= V'\Sigma_{11}V + V'\Sigma_{12} + \Sigma_{21}V + \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
&\quad - V'\Sigma_{12} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \Sigma_{21}V - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{22} \\
&= V'\Sigma_{11}V + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)
\end{aligned}
$$

This transformation is a matrix version of completing the square. We can now apply the Gauss-Markov theorem to the problem of minimizing $V'\Sigma_{11}V = V'\left(\Sigma_{11}^{-1}\right)^{-1}V$ subject to $X_1'V = X_2' - X_1'\Sigma_{11}^{-1}\Sigma_{12}$. The correspondences are:

Theorem	←	Model
Σ	←	Σ_{11}^{-1}
W	←	V
A	←	X_1'
B	←	$X_2' - X_1'\Sigma_{11}^{-1}\Sigma_{12}$

As before, the conditions are met; X_1' is of full-row rank since X_1 is of full-column rank. Hence, $V^* = \Sigma_{11}^{-1}X_1\left(X_1'\Sigma_{11}^{-1}X_1\right)^{-1}\left(X_2' - X_1'\Sigma_{11}^{-1}\Sigma_{12}\right)$, and $\hat{\mathbf{y}}_2 = W^{*\prime}\mathbf{y}_1$, where:

$$
\begin{aligned}
W^{*\prime} &= \left(V^* + \Sigma_{11}^{-1}\Sigma_{12}\right)' \\
&= V^{*\prime} + \Sigma_{21}\Sigma_{11}^{-1} \\
&= \left(X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1\right)\left(X_1'\Sigma_{11}^{-1}X_1\right)^{-1}X_1'\Sigma_{11}^{-1} + \Sigma_{21}\Sigma_{11}^{-1}
\end{aligned}
$$

The minimized prediction-error variance is:

$$
\begin{aligned}
Var[\mathbf{y}_2 - \hat{\mathbf{y}}_2] &= V^{*\prime}\Sigma_{11}V^* + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right) \\
&= \left(X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1\right)\left(X_1'\Sigma_{11}^{-1}X_1\right)^{-1}\left(X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1\right)' + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)
\end{aligned}
$$

Introducing the estimator $\hat{\boldsymbol{\beta}} = \left(X_1'\Sigma_{11}^{-1}X_1\right)^{-1}X_1'\Sigma_{11}^{-1}\mathbf{y}_1$ and its variance $Var\left[\hat{\boldsymbol{\beta}}\right] = \left(X_1'\Sigma_{11}^{-1}X_1\right)^{-1}$ allows us to simplify:

$$\begin{aligned}
\hat{\mathbf{y}}_2 &= W^{*'}\mathbf{y}_1 \\
&= \left(X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1\right)\left(X_1'\Sigma_{11}^{-1}X_1\right)^{-1}X_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1 \\
&= \left(X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1\right)\hat{\boldsymbol{\beta}} + \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1 \\
&= X_2\hat{\boldsymbol{\beta}} + \Sigma_{21}\Sigma_{11}^{-1}\left(\mathbf{y}_1 - X_1\hat{\boldsymbol{\beta}}\right)
\end{aligned}$$

$$Var\left[\mathbf{y}_2 - \hat{\mathbf{y}}_2\right] = \left(X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1\right)Var\left[\hat{\boldsymbol{\beta}}\right]\left(X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1\right)' + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Combining this completing-the-square technique with the Gauss-Markov theorem makes for a cleaner and more elegant proof than that in Halliwell [1997, Appendix C, 328-330].

9. LINEAR STATISTICAL MODELS WITH PARAMETER CONSTRAINTS

Here we will impose upon the model of Section 8 a constraint on β, viz., $R_{j\times k}\beta_{k\times 1} = r_{j\times 1}$. The rows of R must be linearly independent, i.e., R must be of full-row rank. The constraint set $\left\{\beta \in \Re^k : R\beta = r\right\}$ is non-empty because right inverses of R exist, most obviously $R'\left(RR'\right)^{-1}$. Hence, $\beta = R'\left(RR'\right)^{-1}r$ exists and satisfies the constraint.

Two procedures are commonly employed to solve β-constrained linear models. The first is to reduce the parameter dimension according to the equation $\beta = \beta_0 + S_{k\times(k-j)}\gamma_{(k-j)\times 1}$, for some matrix S (of full-column rank) such that $RS = 0_{j\times(k-j)}$, as done in Halliwell [1997, Appendix B, 321-324]. This is the purist approach to the problem, but it requires an understanding of eigen-decomposition, cannot be performed in Excel without add-ins, and may suffer from the numerical-analysis problem of deciding when small eigenvalues should be zeroed. The second procedure is to employ the Lagrange multiplier (Judge [1988, §6.2, 235-237]) to minimize $\Lambda(\beta, \lambda_{j\times 1}) = (\mathbf{y} - X\beta)' \Sigma^{-1}(\mathbf{y} - X\beta) + 2\lambda'(R\beta - r)$. But a third procedure (Halliwell [1998, Appendix C]) to us is the most convincing.

This procedure is to treat the β constraint as the limit of $\mathbf{r} = R\beta + \boldsymbol{\eta}$ as $Var[\boldsymbol{\eta}] \to 0_{j\times j}$. We could have specified the variance as $\sigma^2 I_j$, and the limit as $\sigma^2 \to 0$; but for the sake of generality we will let $Var[\boldsymbol{\eta}] = H$ be any positive-definite matrix. So we can form the following augmented linear model, which satisfies the conditions of Section 8:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{r} \\ \hdashline \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ R \\ \hdashline X_2 \end{bmatrix} \beta + \begin{bmatrix} \mathbf{e}_1 \\ \boldsymbol{\eta} \\ \hdashline \mathbf{e}_2 \end{bmatrix}, \text{where } Var \begin{bmatrix} \mathbf{e}_1 \\ \boldsymbol{\eta} \\ \hdashline \mathbf{e}_2 \end{bmatrix} = \left[\begin{array}{cc:c} \Sigma_{11} & 0 & \Sigma_{12} \\ 0 & H & 0 \\ \hdashline \Sigma_{21} & 0 & \Sigma_{22} \end{array}\right]$$

The parameter estimator, which depends on H, is:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(H) &= \left(\begin{bmatrix} X_1' & R' \end{bmatrix} \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & H \end{bmatrix}^{-1} \begin{bmatrix} X_1 \\ R \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1' & R' \end{bmatrix} \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & H \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{r} \end{bmatrix} \\
&= \left(\begin{bmatrix} X_1' & R' \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} X_1 \\ R \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1' & R' \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{r} \end{bmatrix} \\
&= \left(X_1'\Sigma_{11}^{-1}X_1 + R'H^{-1}R \right)^{-1} \left(X_1'\Sigma_{11}^{-1}\mathbf{y}_1 + R'H^{-1}\mathbf{r} \right) \\
&= Var\left[\hat{\boldsymbol{\beta}}(H)\right]\left(X_1'\Sigma_{11}^{-1}\mathbf{y}_1 + R'H^{-1}\mathbf{r} \right)
\end{aligned}$$

Therefore, according to the formulas of the previous section, the predictor is:

$$\begin{aligned}
\hat{\mathbf{y}}_2(H) &= X_2\hat{\boldsymbol{\beta}}(H) + \begin{bmatrix} \Sigma_{21} & 0 \end{bmatrix} \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & H \end{bmatrix}^{-1} \left(\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{r} \end{bmatrix} - \begin{bmatrix} X_1 \\ R \end{bmatrix} \hat{\boldsymbol{\beta}}(H) \right) \\
&= X_2\hat{\boldsymbol{\beta}}(H) + \Sigma_{21}\Sigma_{11}^{-1}\left(\mathbf{y}_1 - X_1\hat{\boldsymbol{\beta}}(H) \right)
\end{aligned}$$

And the variance of the prediction error is:

$$\begin{aligned}Var\left[\mathbf{y}_2-\hat{\mathbf{y}}_2(\mathrm{H})\right] &= \left(\mathrm{X}_2-\begin{bmatrix}\Sigma_{21} & 0\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{X}_1\\ \mathrm{R}\end{bmatrix}\right)Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\left(\mathrm{X}_2-\begin{bmatrix}\Sigma_{21} & 0\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{X}_1\\ \mathrm{R}\end{bmatrix}\right)'\\ &\quad+\left(\Sigma_{22}-\begin{bmatrix}\Sigma_{21} & 0\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\begin{bmatrix}\Sigma_{12}\\ 0\end{bmatrix}\right)\\ &= \left(\mathrm{X}_2-\Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\left(\mathrm{X}_2-\Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)'+\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)\end{aligned}$$

These two formulas depend on H only insofar as $\hat{\boldsymbol{\beta}}$ depends on H. Therefore, it remains for us to determine $\boldsymbol{\beta}^* = \lim_{\mathrm{H}\to 0}\hat{\boldsymbol{\beta}}(\mathrm{H})$.

We start with $Var\left[\boldsymbol{\beta}^*\right]=\lim_{\mathrm{H}\to 0}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1+\mathrm{R}'\mathrm{H}^{-1}\mathrm{R}\right)^{-1}$. The following proof makes use of the theorem $(\mathrm{A}+\mathrm{BDC})^{-1}=\mathrm{A}^{-1}-\mathrm{A}^{-1}\mathrm{B}\left(\mathrm{D}^{-1}+\mathrm{CA}^{-1}\mathrm{B}\right)^{-1}\mathrm{CA}^{-1}$ (cf. Judge [1988, A.7, 938]; the inverses must exist, as they do here): [8]

$$\begin{aligned}Var\left[\boldsymbol{\beta}^*\right] &= \lim_{\mathrm{H}\to 0}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1+\mathrm{R}'\mathrm{H}^{-1}\mathrm{R}\right)^{-1}\\ &= \lim_{\mathrm{H}\to 0}\left\{\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}-\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\right\}\\ &= \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}-\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\left[0+\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\\ &= \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}-\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\left[\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\\ &= Var\left[\hat{\boldsymbol{\beta}}\right]-Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\end{aligned}$$

The variance of the constrained estimator is neatly expressed in terms of the variance of the unconstrained estimator $Var\left[\hat{\boldsymbol{\beta}}\right]=\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}$. As a check:

$$Var\left[\mathrm{R}\boldsymbol{\beta}^*\right]=\mathrm{R}\,Var\left[\boldsymbol{\beta}^*\right]\mathrm{R}'=\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'-\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'=0$$

[8] In the following formulas the existence of the inverse of $\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'=\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'$ is guaranteed, since the variance matrix is positive-definite and R is of full-row rank.

In order to take the limit of $\hat{\boldsymbol{\beta}}(\mathrm{H})$ we need the following intermediate result:

$$\begin{aligned}
\mathit{Var}\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\mathrm{R}'\mathrm{H}^{-1} &= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\left\{\mathrm{I}_k - \mathrm{R}'\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\right\}\mathrm{R}'\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\left\{\mathrm{R}' - \mathrm{R}'\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left\{\mathrm{I}_j - \left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left\{\mathrm{I}_j - \left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'-\mathrm{H}\right]\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left\{\mathrm{I}_j - \mathrm{I}_j + \left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{H}\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left\{\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{H}\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1} \\
&= \mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{H}+\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}
\end{aligned}$$

Therefore:

$$\begin{aligned}
\boldsymbol{\beta}^* &= \lim_{\mathrm{H}\to 0}\hat{\boldsymbol{\beta}}(\mathrm{H}) \\
&= \lim_{\mathrm{H}\to 0}\mathit{Var}\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \mathrm{R}'\mathrm{H}^{-1}\mathbf{r}\right) \\
&= \lim_{\mathrm{H}\to 0}\mathit{Var}\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \lim_{\mathrm{H}\to 0}\mathit{Var}\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\mathrm{R}'\mathrm{H}^{-1}\mathbf{r} \\
&= \mathit{Var}\left[\boldsymbol{\beta}^*\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \left(\lim_{\mathrm{H}\to 0}\mathit{Var}\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\mathrm{R}'\mathrm{H}^{-1}\right)\left(\operatorname*{\mathit{p\,lim}}_{\mathrm{H}\to 0}\mathbf{r}\right) \\
&= \mathit{Var}\left[\boldsymbol{\beta}^*\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \left(\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\right)\left(\operatorname*{\mathit{p\,lim}}_{\mathit{Var}[\boldsymbol{\eta}]\to 0}\left[\mathrm{R}\beta+\boldsymbol{\eta}\right]\right) \\
&= \mathit{Var}\left[\boldsymbol{\beta}^*\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \left(\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\right)\mathrm{R}\beta \\
&= \mathit{Var}\left[\boldsymbol{\beta}^*\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right] - \mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\right\}\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathrm{I}_k - \mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathrm{I}_k - \mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + \mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}\mathit{Var}\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r}
\end{aligned}$$

As a check:

$$\begin{aligned}
\mathrm{R}\boldsymbol{\beta}^* &= \mathrm{R}\left\{\mathrm{I}_k - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathrm{R} - \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathrm{R} - \mathrm{I}_j\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + \mathrm{I}_j\mathrm{r} \\
&= \mathrm{r}
\end{aligned}$$

And so, to summarize, the solution of a β-constrained linear model is the solution of the unconstrained model with the substitution of $\boldsymbol{\beta}^*$ for $\hat{\boldsymbol{\beta}}$, where:

$$\boldsymbol{\beta}^* = \left\{\mathrm{I}_k - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r}$$

$$Var\left[\boldsymbol{\beta}^*\right] = Var\left[\hat{\boldsymbol{\beta}}\right] - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]$$

10. PARAMETER CONSTRAINTS AS PROJECTIONS

The formulas above for $\boldsymbol{\beta}^*$ and $Var\left[\boldsymbol{\beta}^*\right]$ may seem cumbersome, perhaps even repugnant. However, they become perspicuous when interpreted as a projection. From Section 6 we take the projection formula $P(\mathbf{x};\ \{\mathrm{Ax} = \mathrm{b}\}, \Sigma) = \left\{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\right\}\mathbf{x} + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b}$, where $\Sigma = Var[\mathbf{x}]$. But now let the constraint space Ω be $\left\{\beta \in \Re^{\mathrm{k}} : \mathrm{R}\beta = \mathrm{r}\right\}$. In this case:

$$\begin{aligned}
\boldsymbol{\beta}^* &= \left\{\mathrm{I}_k - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= P\left(\hat{\boldsymbol{\beta}};\ \{\mathrm{R}\beta = \mathrm{r}\}, Var\left[\hat{\boldsymbol{\beta}}\right]\right)
\end{aligned}$$

Hence, the constrained parameter estimator is the projection of the unconstrained estimator according to the metric of the variance of the unconstrained estimator. Just to corroborate, we see that the variance of the constrained estimator,

$Var\left[\boldsymbol{\beta}^*\right] = Var\left[\hat{\boldsymbol{\beta}}\right] - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]$ accords with the projection variance $Var\left[P(\mathbf{x};\Omega,\Sigma)\right] = \Sigma - \Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{A}\Sigma$.

Similarly to how we argued in Section 4, $Var\left[\hat{\boldsymbol{\beta}}\right]$ can be Cholesky-decomposed as $Var\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{QQ}'$ for some non-singular $\mathrm{Q}_{k\times k}$. So the constrained variance can be factored as $Var\left[\boldsymbol{\beta}^*\right] = \mathrm{Q}\left\{\mathrm{I}_k - \mathrm{Q}'\mathrm{R}'\left(\mathrm{RQQ}'\mathrm{R}'\right)^{-1}\mathrm{RQ}\right\}\mathrm{Q}' = \mathrm{QMQ}'$, and its rank is that of the idempotent matrix M, whose rank equals its trace. Again, to continue as in Section 4:

$$Tr(\mathrm{M}) = Tr(\mathrm{I}_k) - Tr\left(\mathrm{Q}'\mathrm{R}'\left(\mathrm{RQQ}'\mathrm{R}'\right)^{-1}\mathrm{RQ}\right) = k - Tr\left(\left(\mathrm{RQQ}'\mathrm{R}'\right)^{-1}\mathrm{RQQ}'\mathrm{R}'\right) = k - j$$

Therefore, $rank\left(Var\left[\boldsymbol{\beta}^*\right]\right) = k - j = rank\left(Var\left[\hat{\boldsymbol{\beta}}\right]\right) - j$. The parameter constraint reduces the parameter variance by j degrees of freedom. In words, $\mathrm{R}\boldsymbol{\beta}^*$ is a degenerate random variable, or a constant. Certainly it is, since by the constraint $\mathrm{R}\boldsymbol{\beta}^* = \mathrm{r}$.

All this shows that the solution of a parameter-constrained model is equivalent to the projection of the solution of an unconstrained model. There seems to be a certain commutativity between constraining/projecting and solving.

11. INFORMATION AS PROJECTION

We start with the equation of Section 7: $\mathbf{y} = \mathrm{X}_{t\times k}\beta_{k\times 1} + \mathbf{e}$, where $Var[\mathbf{e}] = \Sigma_{t\times t}$. However, let us suppose that β is known and needs no estimation. Our best prediction of **y** is Xβ, whose prediction-error variance is $Var[\mathbf{y} - \hat{\mathbf{y}}] = Var[\mathbf{y} - \mathrm{X}\beta] = Var[\mathbf{e}] = \Sigma$. At this stage we are saying nothing more than $\mathbf{y} \sim (X\beta, \Sigma)$. But furthermore suppose that we have observed $A_{m \times t}\mathbf{y}$, where A is of full-row rank. Name the observed value $\mathbf{b}_{m \times 1}$. The problem is to predict **y** after the observation.

Since $\mathbf{y} = X\beta + \mathbf{e}$, $\mathbf{b} = A\mathbf{y} = AX\beta + A\mathbf{e}$, where $Var[A\mathbf{e}] = A\Sigma A'$. Since **y** is the same in the observation as in the prediction, the observation covaries with the prediction; in fact, $Cov[A\mathbf{y}, \mathbf{y}] = Cov[A\mathbf{e}, \mathbf{e}] = A\Sigma$. We can predict **y** according to the parameter-constrained model of Section 9:

$$\begin{bmatrix} \mathbf{b} \\ \beta_0 \\ \hline \mathbf{y} \end{bmatrix} = \begin{bmatrix} AX \\ I_k \\ \hline X \end{bmatrix} \beta + \begin{bmatrix} A\mathbf{e} \\ 0 \\ \hline \mathbf{e} \end{bmatrix}, \text{ where } Var \begin{bmatrix} A\mathbf{e} \\ 0 \\ \hline \mathbf{e} \end{bmatrix} = \left[\begin{array}{cc|c} A\Sigma A' & 0 & A\Sigma \\ 0 & 0 & 0 \\ \hline \Sigma A' & 0 & \Sigma \end{array}\right]$$

Although this will work, a simpler and more appealing model can be constructed if one allows for zero-dimensional matrices.[9] Because all *m*×0 and 0×*n* matrices are of rank zero, $A_{m \times 0} B_{0 \times n} = 0_{m \times n}$. This is nothing more than the nullity of the empty summation operator, i.e., $(AB)_{ij} = \sum_{k=1}^{0} (A)_{ik} (B)_{kj} = 0$. The simpler model is:

$$\begin{bmatrix} \mathbf{b} - AX\beta \\ \hline \mathbf{y} - X\beta \end{bmatrix} = \begin{bmatrix} X_1 \; (m \times 0) \\ \hline X_2 \; (t \times 0) \end{bmatrix} \gamma_{0 \times 1} + \begin{bmatrix} A\mathbf{e} \\ \hline \mathbf{e} \end{bmatrix}, \text{ where } Var \begin{bmatrix} A\mathbf{e} \\ \hline \mathbf{e} \end{bmatrix} = \left[\begin{array}{c|c} A\Sigma A' & A\Sigma \\ \hline \Sigma A' & \Sigma \end{array}\right]$$

Its solution begins with:

[9] It is a windfall for a matrix language to allow for zeros in the dimensions of its arrays, as do APL, J, and R. SAS/IML does not; at least it did not in the late 1990s (version 7), when the author last used it.

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{0\times1} &= Var[\hat{\boldsymbol{\gamma}}]\mathrm{X}_1'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathbf{b}-\mathrm{AX}\beta)\\ &= \left(\mathrm{X}_1'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathbf{b}-\mathrm{AX}\beta)\\ &= (0\times0)^{-1}(0\times m)\cdot(m\times m)^{-1}(m\times1)\\ &= (0\times0)^{-1}(0\times1)\end{aligned}$$

The only thing to give pause here is the inverse of the 0×0 matrix. But the space of real 0-vectors, $\Re^0$, contains just one element, viz., the origin. It is closed under addition and multiplication (0+0 = 0×0 = 0), and 0 serves as its identity element. So in $\Re^0$, $0^{-1}=0$. Hence, $(0\times0)^{-1}=(0\times0)$. Therefore, $Var[\hat{\boldsymbol{\gamma}}]=(0\times0)$ and $\hat{\boldsymbol{\gamma}}_{0\times1}=(0\times1)$.[10] Finally:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathrm{X}\beta+\widehat{\mathbf{y}-\mathrm{X}\beta}\\ &= \mathrm{X}\beta+\mathrm{X}_2\hat{\boldsymbol{\gamma}}+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\{\mathbf{b}-\mathrm{AX}\beta\}-\mathrm{X}_1\hat{\boldsymbol{\gamma}})\\ &= \mathrm{X}\beta+0_{t\times1}+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\{\mathbf{b}-\mathrm{AX}\beta\}-0_{m\times1})\\ &= \mathrm{X}\beta+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathbf{b}-\mathrm{AX}\beta)\end{aligned}$$

The variance of its prediction error is:

$$\begin{aligned}Var[\mathbf{y}-\hat{\mathbf{y}}] &= Var[(\mathbf{y}-\mathrm{X}\beta)-(\hat{\mathbf{y}}-\mathrm{X}\beta)]\\ &= \left(\mathrm{X}_2-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{X}_1\right)Var[\hat{\boldsymbol{\gamma}}]\left(\mathrm{X}_2-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{X}_1\right)'+\Sigma-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\\ &= (t\times0)(0\times0)(0\times t)+\Sigma-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\\ &= 0_{t\times t}+\Sigma-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\\ &= \Sigma-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\end{aligned}$$

Except for slight notational differences, this solution is the same as that of Section 6. As long as no parameter needs to be estimated (or the parameter dimension is 0×1), the linear statistical model treats "*m* dimensions" of prior information as a projection into a subspace of $t-m$ dimensions.

[10] To elaborate on the previous footnote, we have verified that APL, J, and R yield these results. Therefore, they correctly treat $(0\times0)^{-1}$ as (0×0).

12. COMBINING ESTIMATES

It is not uncommon for an actuary linearly to combine two or more unbiased estimators of the same quantity. Of course, it is desirable for the combination to be best. In the simplest situation of independent scalar estimators, the best combination uses weights inversely proportional to the variances of the estimators. But with Gauss-Markov theorem one can determine the best linear combination of vector estimators, even if they are not independent.

To frame the problem, suppose that we have *n* unbiased estimators $\hat{\mathbf{y}}_i$ of the same *t*×1 vector **y**, as well as their *t*×*t* prediction-error variances $\Sigma_{ii} = Var[\mathbf{y} - \hat{\mathbf{y}}_i]$. Suppose also that we have the *t*×*t* prediction-error covariances $\Sigma_{ij} = Cov[\mathbf{y} - \hat{\mathbf{y}}_i, \mathbf{y} - \hat{\mathbf{y}}_j]$. Frequently the covariances are $0_{t\times t}$, but there are realistic exceptions. Stack the estimators and block their (co)variances:

$$\hat{\mathbf{Y}}_{nt\times 1} = \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \vdots \\ \hat{\mathbf{y}}_n \end{bmatrix}, \quad Var\left[\mathbf{Y} - \hat{\mathbf{Y}}\right]_{nt\times nt} = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix}$$

The variance matrix must be non-negative-definite; but we will assume it to be positive definite, hence invertible. If the weight given to $\hat{\mathbf{y}}_i$ is the *t*×*t* matrix W_i', the combined estimator will be $\hat{\mathbf{y}} = \sum_{i=1}^{n} \mathrm{W}_i'\hat{\mathbf{y}}_i = \begin{bmatrix} \mathrm{W}_1' & \cdots & \mathrm{W}_n' \end{bmatrix}_{t\times nt} \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \vdots \\ \hat{\mathbf{y}}_n \end{bmatrix} = \mathrm{W}'\hat{\mathbf{Y}}$. In order for this estimator to be unbiased, $\begin{bmatrix} \mathrm{W}_1' & \cdots & \mathrm{W}_n' \end{bmatrix} \begin{bmatrix} \mathrm{I}_t \\ \vdots \\ \mathrm{I}_t \end{bmatrix} = \mathrm{W}' \begin{bmatrix} \mathrm{I}_t \\ \vdots \\ \mathrm{I}_t \end{bmatrix} = \mathrm{I}_t$. The transpose of this constraint is

$\left[\mathrm{I}_t \quad \cdots \quad \mathrm{I}_t\right]_{t\times nt} \mathrm{W}_{nt\times t} = \mathrm{I}_t$. The best combination will minimize the combined prediction-error variance $\mathrm{W}' Var\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\mathrm{W}$. Posing the problem in the proper form, we seek to minimize $\mathrm{W}' Var\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\mathrm{W} = \mathrm{W}'\left(Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\right)^{-1}\mathrm{W}$ subject to $\left[\mathrm{I}_t \quad \cdots \quad \mathrm{I}_t\right]_{t\times nt} \mathrm{W}_{nt\times t} = \mathrm{I}_t$. According to the Gauss-Markov theorem:

$$\begin{aligned}\mathrm{W}^* &= Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\begin{bmatrix}\mathrm{I}_t\\ \vdots\\ \mathrm{I}_t\end{bmatrix}\left(\left[\mathrm{I}_t \quad \cdots \quad \mathrm{I}_t\right] Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\begin{bmatrix}\mathrm{I}_t\\ \vdots\\ \mathrm{I}_t\end{bmatrix}\right)^{-1}\mathrm{I}_t\\ &= Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\begin{bmatrix}\mathrm{I}_t\\ \vdots\\ \mathrm{I}_t\end{bmatrix}\left(\left[\mathrm{I}_t \quad \cdots \quad \mathrm{I}_t\right] Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\begin{bmatrix}\mathrm{I}_t\\ \vdots\\ \mathrm{I}_t\end{bmatrix}\right)^{-1}\end{aligned}$$

If the covariances $\Sigma_{i\neq j}$ are zero, this simplifies to:

$$\mathrm{W}^* = \begin{bmatrix}\Sigma_{11}^{-1}\\ \vdots\\ \Sigma_{nn}^{-1}\end{bmatrix}\left(\sum_{i=1}^{n}\Sigma_{ii}^{-1}\right)^{-1}$$

It is recognizable as the matrix version of the well known rule of weighting independent scalar estimates inversely proportionally to their variances.[11] Appendix B will provide a simple example of covarying estimates, and will outline its importance to conjoint modeling, or to modeling in which ultimate paid and incurred losses must be equal.

[11] Unlike scalar weighting, a matrix-weighted average can fall outside its extremes, e.g.: $\begin{bmatrix}0.50 & 0.25\\ 0.25 & 0.40\end{bmatrix}\begin{bmatrix}400\\ 420\end{bmatrix} + \begin{bmatrix}0.50 & -0.25\\ -0.25 & 0.60\end{bmatrix}\begin{bmatrix}440\\ 400\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}425\\ 398\end{bmatrix}$, yet $398 \notin [400, 420]$. This is due to non-zero off-diagonal amounts (± 0.25) in the weighting matrices. In practice, such amounts are relatively small, and the matrix-weighted averages lie within their extremes. Cf. Judge [1988, 287].

13. CONCLUSION

The Gauss-Markov theorem is truly profound. It provides a lucid basis for solving a wide range of modeling and estimation problems, even within the rudimentary matrix functionality of Excel. As the many sections of this paper have demonstrated, it deserves to be liberated from being an appendage to the least-squares approach to linear statistical modeling.[12]

[12] For a brief history of the least-squares method and the true relation of the Gauss-Markov theorem to it see Appendix C.

REFERENCES

[1.] Halliwell, Leigh, "Conjoint Prediction of Paid and Incurred Losses," *1997 Loss Reserving Discussion Papers*, Casualty Actuarial Society, 1997, 241-379, www.casact.org/pubs/forum/97sforum/97sf1241.pdf.

[2.] Halliwell, Leigh, "Statistical Models and Credibility," *CAS Forum* (Winter 1998), www.casact.org/pubs/forum/98wforum/98wf061.pdf.

[3.] Halliwell, Leigh, "Chain-Ladder Bias: Its Reason and Meaning," *Variance*, 1:2, 2007, 214-247, www.variancejournal.org/issues/01-02/214.pdf.

[4.] Judge, George G., Hill, R. C., *et al.*, *Introduction to the Theory and Practice of Econometrics* (Second Edition), New York, John Wiley & Sons, 1988.

[5.] Wikipedia contributors, "Best linear unbiased prediction," *Wikipedia*, http://en.wikipedia.org/wiki/Best_linear_unbiased_prediction (accessed September 2015).

[6.] Wikipedia contributors, "Gauss–Markov theorem," *Wikipedia*, http://en.wikipedia.org/wiki/Gauss-Markov_theorem (accessed September 2015).

[7.] Wikipedia contributors, "Least squares," *Wikipedia,* http://en.wikipedia.org/wiki/Least_squares (accessed September 2015)

[8.] Wikipedia contributors, "Mahalanobis distance," *Wikipedia*, http://en.wikipedia.org/wiki/Mahalanobis_distance (accessed September 2015).

APPENDIX A

GEOMETRIC MATTERS CONCERNING VECTORS IN $\Re^n$

In this appendix we will interpret the vector dot product and prove the triangle inequality. For $\mathrm{x}, \mathrm{y} \in \Re^n$ the dot product $\mathrm{x} \bullet \mathrm{y} = x'y = \sum_{i=1}^{n} x_i y_i$. However, this is easily generalized with a Σ metric as $\mathrm{x}'\Sigma^{-1}\mathrm{y}$. The Σ-metric triangle inequality is:

$$\sqrt{(\mathrm{x}+\mathrm{y})'\Sigma^{-1}(\mathrm{x}+\mathrm{y})} \le \sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}} + \sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}$$

As for the dot product, let $\hat{\mathrm{y}}$ be the Σ projection of y onto x. If $\mathrm{x}_{n\times 1} \neq 0$, then x is of full-column rank, $\mathrm{x}'\Sigma^{-1}\mathrm{x}$ is 1×1 positive-definite, and $(\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}$ exists. So vector y will Σ-project as some multiple of x, or $\hat{\mathrm{y}} = \mathrm{x}\beta$. From Section 7, $\beta = (\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}\mathrm{x}'\Sigma^{-1}\mathrm{y}$. Hence, $\hat{\mathrm{y}} = \mathrm{x}(\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}\mathrm{x}'\Sigma^{-1}\mathrm{y}$. Accordingly, $\mathrm{x}'\Sigma^{-1}\mathrm{y} = \mathrm{x}'\Sigma^{-1}\mathrm{x}(\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}\mathrm{x}'\Sigma^{-1}\mathrm{y} = \mathrm{x}'\Sigma^{-1}\hat{\mathrm{y}}$. So the Σ-metric dot product of two vectors is equal to Σ-metric dot product of one vector and the Σ projection of the other onto it. Although $(\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}$ does not exist if $\mathrm{x} = 0$, we know that the projection of any vector onto 0 is 0. Hence, our geometric interpretation of the dot product is valid for all x and y. For the Euclidean metric $\Sigma = \mathrm{I}_n$, the projection is the perpendicular, and $|\hat{\mathrm{y}}| = |\mathrm{y}|\cos\theta$, where θ is the angle between the two vectors with vertex at 0. From this follows the well-known formula $\mathrm{x} \bullet \mathrm{y} = x'y = |\mathrm{x}||\mathrm{y}|\cos\theta$.

In preparation for the triangle inequality, since Σ is positive-definite, the following 2×2 symmetric matrix is non-negative-definite:

$$\begin{bmatrix} \mathrm{x}_{n\times 1} & \mathrm{y}_{n\times 1} \end{bmatrix}' \Sigma^{-1} \begin{bmatrix} \mathrm{x} & \mathrm{y} \end{bmatrix} = \begin{bmatrix} \mathrm{x}' \\ \mathrm{y}' \end{bmatrix} \Sigma^{-1} \begin{bmatrix} \mathrm{x} & \mathrm{y} \end{bmatrix} = \begin{bmatrix} \mathrm{x}'\Sigma^{-1}\mathrm{x} & \mathrm{x}'\Sigma^{-1}\mathrm{y} \\ \mathrm{y}'\Sigma^{-1}\mathrm{x} & \mathrm{y}'\Sigma^{-1}\mathrm{y} \end{bmatrix}$$

It is a theorem that the determinant of a non-negative-definite matrix is non-negative; but we can readily prove it here for the 2×2 case. Such a matrix can be Cholesky factored as $\begin{bmatrix} a & 0 \\ b & d \end{bmatrix}\begin{bmatrix} a & b \\ 0 & d \end{bmatrix}$, for real numbers *a*, *b*, and *c*. This equals $\begin{bmatrix} a^2 & ab \\ ba & b^2 + d^2 \end{bmatrix}$, whose determinant is $a^2\left(b^2 + d^2\right) - a^2b^2 = a^2d^2$, which must be greater than or equal to zero.

Now let '~' stand for the relationship in the triangle inequality:

$$\sqrt{(\mathrm{x}+\mathrm{y})'\Sigma^{-1}(\mathrm{x}+\mathrm{y})} \sim \sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}} + \sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}$$

Because the quantities under all the radical signs are non-negative, the following transformations will not affect the relationship:

$$\begin{aligned}
(\mathrm{x}+\mathrm{y})'\Sigma^{-1}(\mathrm{x}+\mathrm{y}) &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x} + 2\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}} + \mathrm{y}'\Sigma^{-1}\mathrm{y} \\
\mathrm{x}'\Sigma^{-1}\mathrm{x} + \mathrm{x}'\Sigma^{-1}\mathrm{y} + \mathrm{y}'\Sigma^{-1}\mathrm{x} + \mathrm{y}'\Sigma^{-1}\mathrm{y} &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x} + 2\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}} + \mathrm{y}'\Sigma^{-1}\mathrm{y} \\
\mathrm{x}'\Sigma^{-1}\mathrm{y} + \mathrm{y}'\Sigma^{-1}\mathrm{x} &\sim 2\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}} \\
2\,\mathrm{x}'\Sigma^{-1}\mathrm{y} &\sim 2\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}} \\
\mathrm{x}'\Sigma^{-1}\mathrm{y} &\sim \sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}} \\
\left(\mathrm{x}'\Sigma^{-1}\mathrm{y}\right)^2 &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x} \cdot \mathrm{y}'\Sigma^{-1}\mathrm{y} \\
\mathrm{x}'\Sigma^{-1}\mathrm{y} \cdot \mathrm{y}'\Sigma^{-1}\mathrm{x} &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x} \cdot \mathrm{y}'\Sigma^{-1}\mathrm{y} \\
0 &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x} \cdot \mathrm{y}'\Sigma^{-1}\mathrm{y} - \mathrm{x}'\Sigma^{-1}\mathrm{y} \cdot \mathrm{y}'\Sigma^{-1}\mathrm{x}
\end{aligned}$$

But the expression on the right of the last line is the determinant of our 2×2 non-negative-definite matrix. Therefore '~' is '≤'. Thus have we proven the triangle inequality in $\Re^n$ for every valid Σ metric.

APPENDIX B

COVARYING ESTIMATORS AND CONJOINT PREDICTION

This appendix furnishes a simple, but not too contrived, example of combining estimators that are not independent. Let $X_i \sim [\mu, \sigma^2]$ be independent random variables. Our task will be to estimate the mean μ. However, we must estimate it from two known statistics, $Y_1 = (X_1 + X_2 + X_3)/3$ and $Y_2 = (X_3 + X_4)/2$. Four *X* variables have been melded into two *Y* variables: $Y_1 \sim [\mu, \sigma^2/3]$ and $Y_2 \sim [\mu, \sigma^2/2]$. But since X_3 is common to both, they are not independent; rather, $Cov[Y_1, Y_2] = Cov[X_3/3, X_3/2] = \sigma^2/6$. So the first two moments of the **y** vector are:

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \sigma^2 \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix} \right)$$

The problem is to minimize $\mathrm{W}'\sigma^2 \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix} \mathrm{W}$ subject to $[1 \quad 1]\mathrm{W}_{2\times 1} = \mathrm{I}_1$. By the Gauss-Markov theorem (the σ^2 cancels, so it's omitted):

$$\mathrm{W}^* = \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left([1 \quad 1] \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-1} \mathrm{I}_1$$

$$= \frac{1}{5/36} \begin{bmatrix} 1/2 & -1/6 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left([1 \quad 1] \frac{1}{5/36} \begin{bmatrix} 1/2 & -1/6 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-1}$$

$$= \begin{bmatrix} 1/2 & -1/6 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left([1 \quad 1] \begin{bmatrix} 1/2 & -1/6 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 1/3 \\ 1/6 \end{bmatrix} (1/2)^{-1} = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$$

So the minimal variance results from combining in a 2:1 = 10:5 ratio. One who ignored the covariance would weight them in a 3:2 = 9:6 ratio, underweighting the first and overweighting the second. The minimal variance itself is:

$$\begin{bmatrix} 2/3 & 1/3 \end{bmatrix} \sigma^2 \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix} \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix} = \sigma^2 \frac{5}{18} = \sigma^2 \cdot 0.2\overline{7}$$

Since $\sigma^2/4 < \sigma^2 \cdot 0.2\overline{7} =< \sigma^2/3$, the informational value of the two *Y* statistics lies in between the informational values of three and four independent *X* statistics.

As for conjoint prediction, the following model combines submodels *a* and *b*:

$$\begin{bmatrix} \mathbf{y}_{a1} \\ \mathbf{y}_{b1} \\ \hdashline \mathbf{y}_{a2} \\ \mathbf{y}_{b2} \end{bmatrix} = \begin{bmatrix} X_{a1} & 0 \\ 0 & X_{b1} \\ \hdashline X_{a2} & 0 \\ 0 & X_{b2} \end{bmatrix} \begin{bmatrix} \beta_a \\ \beta_b \end{bmatrix} + \begin{bmatrix} \mathbf{e}_{a1} \\ \mathbf{e}_{b1} \\ \hdashline \mathbf{e}_{a2} \\ \mathbf{e}_{b2} \end{bmatrix}, \text{where } Var \begin{bmatrix} \mathbf{e}_{a1} \\ \mathbf{e}_{b1} \\ \hdashline \mathbf{e}_{a2} \\ \mathbf{e}_{b2} \end{bmatrix} = \left[\begin{array}{cc:cc} \Sigma_{a11} & 0 & \Sigma_{a12} & 0 \\ 0 & \Sigma_{b11} & 0 & \Sigma_{b12} \\ \hdashline \Sigma_{a21} & 0 & \Sigma_{a22} & 0 \\ 0 & \Sigma_{b21} & 0 & \Sigma_{b22} \end{array}\right]$$

One who works through the formulas of Section 8 will find that the solution of the combination is identical to the combination of the separate solutions (Halliwell [1998, Appendix A]). Were it not for this good fortune, one would have to model everything in order to model anything. So this combination is trivial; although the submodels are written down together, they interact neither in the design matrix nor in the variance structure. But conjoint prediction (Halliwell [1997]) makes use of the fact that paid losses (model *a*) and incurred losses (model *b*) must ultimately be equal by exposure period. This constrains the variance matrix; the sums of the paid and the incurred errors of each exposure period must be equal. But additionally, it imposes a restriction on the parameters. The *a priori*, or prior-to-any-observation, expected values are $E[\mathbf{y}_a] = X_a\beta_a$ and $E[\mathbf{y}_b] = X_b\beta_b$. The exposure-period sums of these paid and incurred vectors must also be equal. A "semi-conjoint" model adds the appropriate β constraint to the trivial combination: $R\beta = \left(Q\begin{bmatrix} X_a & 0 \\ 0 & X_b \end{bmatrix} \right) \cdot \begin{bmatrix} \beta_a \\ \beta_b \end{bmatrix} = 0 = r$; consequently, $\hat{\boldsymbol{\beta}}_a$ and $\hat{\boldsymbol{\beta}}_b$ will covary. This parameter covariance will introduce covariance, or off-diagonal blocks, into $Var\begin{bmatrix} \mathbf{y}_{a2} - \hat{\mathbf{y}}_{a2} \\ \mathbf{y}_{b2} - \hat{\mathbf{y}}_{b2} \end{bmatrix}$.

According to Section 12, one may best combine the semi-constrained solutions $\begin{bmatrix} \hat{\mathbf{y}}_{a2} \\ \hat{\mathbf{y}}_{b2} \end{bmatrix}$ and $Var\begin{bmatrix} \mathbf{y}_{a2} - \hat{\mathbf{y}}_{a2} \\ \mathbf{y}_{b2} - \hat{\mathbf{y}}_{b2} \end{bmatrix}$ according to the linear constraint that exposure-period sums of paid and incurred losses are equal. Equivalently, in terms of Section 11, one can project the semi-constrained solution into the subspace of the constraint. Although a proof of this has so far eluded us, it works with examples. So the Gauss-Markov theorem seems to allow modeling temporarily to ignore variance restrictions in order to arrive at a tentative solution that can rather easily be collapsed by the hitherto ignored restrictions into the desired solution. This is the meaning of the sentence at the end of Section 10: "There seems to be a certain commutativity between constraining/projecting and solving." Conjoint prediction by collapsing a semi-conjoint model is much easier than fully conjoint prediction; it requires no eigen-decomposition, and is amenable to a spreadsheet solution.

APPENDIX C

LEAST-SQUARES VERSUS GAUSS-MARKOV

Many, probably most, actuaries think in terms of linear regression, rather than in terms of linear modeling. The standard linear-regression problem begins with *t* observed quantities *y_j*. Each observation is associated with a *k*-tuple of known variables $(x_{j1},\dots,x_{jk})$, on which the observation is believed linearly to depend, i.e., $y_j = x_{j1}\beta_1 + \dots + x_{jk}\beta_k$. Of course, if $t = k$ and the *k*-tuples are linearly independent, one is merely solving simultaneous equations for the β_j. The regression problem arises when $t > k$, and the equations are approximate: $y_j \approx x_{j1}\beta_1 + \dots + x_{jk}\beta_k$. One then needs to find the values of β_j that make $x_{j1}\beta_1 + \dots + x_{jk}\beta_k$ most closely approximate the *y_j*. A reasonable method, called "least squares," is to find the β_j that minimize the sum of the squared errors, i.e., to minimize $f(\beta_1,\dots,\beta_k) = \sum_{j=1}^{t}(y_j - x_{j1}\beta_1 + \dots + x_{jk}\beta_k)^2$. This is a problem well within the capability of a first-year calculus student.

The least-squares criterion for fitting, or "regressing," the best line to data first appeared in print in 1805, when Legendre published his Nouvelles méthodes pour la détermination des orbites des comètes. Earlier, in 1801, Gauss had applied the method to predict the reappearance of Ceres, which had just been discovered and then lost. However, he did not publish the method until 1806 in his Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium. Apparently, he did not refer to Legendre; and in the ensuing controversy over priority Gauss insisted that he had worked out the method at least as early as 1795, when at the age of eighteen he entered the University of Göttingen.[13] The relevant point of this interesting story is that in this early astronomical setting, the least-squares method was not statistical modeling. It was applied to deterministically moving objects (comets and the newly discovered asteroids). All uncertainty stemmed from the imprecision of the astronomers. But for the first time it was realized that many economical but "fuzzy" observations could be more useful than one costly but "sharp" observation.

Gradually the approximate equations were turned into exact ones with random error terms: $\boldsymbol{y}_j = x_{j1}\beta_1 + \ldots + x_{jk}\beta_k + \boldsymbol{e}_j$. Gauss himself in 1822 stated the optimality of the least-squares method, an early form of BLUE. So today we talk of the "Gauss-Markov" theorem because Gauss started it. But the linear algebra and statistical theory that developed after his death in 1855 culminated in the work of Andrey Markov (1856-1922). Even today it is common for students to be introduced into linear modeling by way of least squares; many texts still refer to the matrix formula $\boldsymbol{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$ as the OLS ("Ordinary Least Squares") estimator.

How does the least-squares method differ from our version of the Gauss-Markov theorem? To put it in modern terms, both deal with estimating the β parameter in the model of Section 7: $\mathbf{y} = \mathrm{X}\beta + \mathbf{e}$, where $Var[\mathbf{e}] = \Sigma$. But instead of finding the $t \times k$ matrix W that will make estimator $\hat{\boldsymbol{\beta}} = \mathrm{W}'\mathbf{y}$ unbiased and of minimal variance, as per the Gauss-Markov theorem, the least-squares method seeks the value of β for which Xβ most closely

[13] Wikipedia "Least squares" gives an excellent account of this history, which is also recounted in many histories of mathematics. However, there is slight disagreement about some of the dates. Most historians cede the priority to Gauss.

approximates **y**. "Closeness" here requires distance as measured by the Σ metric. So the least-squares problem is to minimize $f(\beta)=(\mathbf{y}-X\beta)'\Sigma^{-1}(\mathbf{y}-X\beta)$. As with the Gauss-Markov theorem, X must be of full-column rank and Σ must be positive-definite. But the two approaches are not logically equivalent. Although they yield the same answer, BLUE is *a posteriori* to the least-squares answer, whereas it is *a priori* to the Gauss-Markov.

The usual approach to the minimization is by means of multivariate calculus:

$$\frac{\partial f}{\partial \beta}=-2X'\Sigma^{-1}(\mathbf{y}-X\beta)$$

$$\frac{\partial^2 f}{\partial \beta \partial \beta'}=2X'\Sigma^{-1}X$$

Setting the first derivative to $0_{k\times 1}$, we derive $\tilde{\boldsymbol{\beta}}=(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\mathbf{y}$. Since the second derivative is positive-definite, the critical value $\tilde{\boldsymbol{\beta}}$ is a minimum, as desired. However, since vector differentiation is unfamiliar to many (cf. Judge [1988, Appendix A.16]), we will solve the problem algebraically:

$$\begin{aligned}
f(\beta) &= (\mathbf{y}-X\beta)'\Sigma^{-1}(\mathbf{y}-X\beta) \\
&= (\mathbf{y}-X\tilde{\boldsymbol{\beta}}+X[\tilde{\boldsymbol{\beta}}-\beta])'\Sigma^{-1}(\mathbf{y}-X\tilde{\boldsymbol{\beta}}+X[\tilde{\boldsymbol{\beta}}-\beta]) \\
&= (\mathbf{y}-X\tilde{\boldsymbol{\beta}})'\Sigma^{-1}(\mathbf{y}-X\tilde{\boldsymbol{\beta}})+(\mathbf{y}-X\tilde{\boldsymbol{\beta}})'\Sigma^{-1}X[\tilde{\boldsymbol{\beta}}-\beta]+[\tilde{\boldsymbol{\beta}}-\beta]'X'\Sigma^{-1}(\mathbf{y}-X\tilde{\boldsymbol{\beta}})+[\tilde{\boldsymbol{\beta}}-\beta]'X'\Sigma^{-1}X[\tilde{\boldsymbol{\beta}}-\beta] \\
&= f(\tilde{\boldsymbol{\beta}})+2[\tilde{\boldsymbol{\beta}}-\beta]'X'\Sigma^{-1}(\mathbf{y}-X\tilde{\boldsymbol{\beta}})+[\tilde{\boldsymbol{\beta}}-\beta]'X'\Sigma^{-1}X[\tilde{\boldsymbol{\beta}}-\beta] \\
&= f(\tilde{\boldsymbol{\beta}})+2[\tilde{\boldsymbol{\beta}}-\beta]'\{X'\Sigma^{-1}\mathbf{y}-X'\Sigma^{-1}X\tilde{\boldsymbol{\beta}}\}+[\tilde{\boldsymbol{\beta}}-\beta]'X'\Sigma^{-1}X[\tilde{\boldsymbol{\beta}}-\beta] \\
&= f(\tilde{\boldsymbol{\beta}})+2[\tilde{\boldsymbol{\beta}}-\beta]'\{X'\Sigma^{-1}\mathbf{y}-X'\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\mathbf{y}\}+[\tilde{\boldsymbol{\beta}}-\beta]'X'\Sigma^{-1}X[\tilde{\boldsymbol{\beta}}-\beta] \\
&= f(\tilde{\boldsymbol{\beta}})+2[\tilde{\boldsymbol{\beta}}-\beta]'\{X'\Sigma^{-1}\mathbf{y}-X'\Sigma^{-1}\mathbf{y}\}+[\tilde{\boldsymbol{\beta}}-\beta]'X'\Sigma^{-1}X[\tilde{\boldsymbol{\beta}}-\beta] \\
&= f(\tilde{\boldsymbol{\beta}})+[\tilde{\boldsymbol{\beta}}-\beta]'X'\Sigma^{-1}X[\tilde{\boldsymbol{\beta}}-\beta] \\
&\geq f(\tilde{\boldsymbol{\beta}})
\end{aligned}$$

As in Section 2, the last line is to be taken in a matrix-definite sense. Moreover, since $X'\Sigma^{-1}X$ is positive-definite, the inequality is strict except for $\beta = \tilde{\boldsymbol{\beta}}$. Therefore, $\tilde{\boldsymbol{\beta}}$ uniquely minimizes $f(\beta) = (\mathbf{y} - X\beta)' \Sigma^{-1} (\mathbf{y} - X\beta)$. Geometrically, the least-squares method drops a Σ perpendicular from **y** to the linear subspace swept by Xβ.

But now that we have a "least-squares" estimator, we must check its "BLUE-ness." This is the meaning of the sentence above, that BLUE is *a posteriori* to the least-squares answer. Of course, from our *a priori* Gauss-Markov approach, we already know it to be BLUE, since it is identical to the Section 7 formula $\hat{\boldsymbol{\beta}} = (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}\mathbf{y}$. If $\tilde{\boldsymbol{\beta}}$ were not identical to $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$ would be either biased or not as good as $\hat{\boldsymbol{\beta}}$; it would lack either the 'B' or the 'U' of BLUE.

Finally, despite the historical development from least squares to Gauss-Markov, this is neither a "distinction without a difference" nor a matter of taste. Developing the theory of linear statistical modeling from our Gauss-Markov theorem allows us cleanly to solve problems that the least-squares approach can solve only with difficulty, if at all – such problems as predicting (Section 8), constraining (Section 9), projecting (Section 10), incorporating prior information (Section 11), and combining estimates (Section 12).