Statistical Models and Credibility

Leigh J. Halliwell, FCAS, MAAA

Milliman & Robertson, Inc.
3 Garret Mountain Plaza
West Patterson, NJ 07424

Casualty Actuarial Society
1998 Seminar on Ratemaking
Chicago Hilton and Towers
Friday, March 13, 1998

Outline

1. Matrices as Linear Mappings

2. The Linear Statistical Model

3. Credibility and Prior Information

4. Credibility and the Random-Effects Model

5. Conclusion

All this in seventy-five minutes?

1. Matrices as Linear Mappings

$$\text{Vector } x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in R^n$$

x is a point in *n*-dimensional real-number space. It packages *n* pieces of information.

How to multiply a vector by a scalar and how to add two *n*-dimensional vectors are obvious.

Define the unit vector $u_{n,i}$ as the R^n vector whose i^{th} element is one, the other elements being zeroes.

By the properties of addition and scalar multiplication,

$$
\begin{aligned}
\mathbf{x} &= \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\
&= \begin{bmatrix} x_1 \\ \vdots \\ 0 \end{bmatrix} + \ldots + \begin{bmatrix} 0 \\ \vdots \\ x_n \end{bmatrix} \\
&= x_1 \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} + \ldots + x_n \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix} \\
&= \sum_{i=1}^{n} x_i \mathbf{u}_{n,i}
\end{aligned}
$$

Consider a linear mapping A: $R^n \rightarrow R^m$

Linear as to scalar multiplication: $A(\alpha x) = \alpha A(x)$

Linear as to vector addition: $A(x_1+x_2) = A(x_1)+A(x_2)$.

In general,

$$A\left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right) = A\left(\sum_{i=1}^{n} x_i u_{n,i}\right) = \sum_{i=1}^{n} x_i A\left(u_{n,i}\right)$$

Therefore, a linear mapping is uniquely determined by to where it maps the unit vectors of R^n.

Let $A_i = A(u_{n,i}) \in R^m$. Every linear mapping can be represented by the $m \times n$ matrix $[A_1 \ldots A_n]$. The i^{th} column of the matrix specifies the vector of R^m to which A maps the i^{th} unit vector of R^n.

So,

$$A(x) = \sum_{i=1}^{n} x_i A(u_{n,i}) = \sum_{i=1}^{n} x_i A_i = [A_1 \quad \cdots \quad A_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

This looks like matrix multiplication, although matrix multiplication has not yet been defined (see slide 14).

How to multiply a matrix by a scalar and how to add two $m \times n$ matrices are obvious.

If A and B are two linear mappings from R^n to R^m, then,

$$(A + B)(\mathbf{x}) = [A + B \quad \cdots \quad A + B](\mathbf{x})$$

$$= \sum_{i=1}^{n} x_i (A + B)$$

$$= \sum_{i=1}^{n} x_i (A_i + B_i)$$

$$= \sum_{i=1}^{n} x_i A_i + \sum_{i=1}^{n} x_i B_i$$

$$= A(\mathbf{x}) + B(\mathbf{x})$$

Matrix addition is commutative and associative. There is a zero matrix, and every matrix has an additive inverse. These are the addition characteristics of rings.

But what about matrix multiplication?

Let B ($l \times m$) represent a mapping from R^m to R^l. Let '•' represent the composition of mappings. B•A maps from R^n to R^l. But:

$$
\begin{aligned}
(B \bullet A)(x) &= B(A(x)) \\
&= B\left(\sum_{i=1}^{n} x_i A_i\right) \\
&= \sum_{i=1}^{n} x_i B(A_i) \\
&= \left[B(A_1) \quad \cdots \quad B(A_n) \right] x
\end{aligned}
$$

The columns of B•A make sense. For example, A maps $u_{n,1}$ to A_1, then B maps A_1 to $B(A_1)$. So the i^{th} column of B•A shows to what vector of R^l B•A maps $u_{n,i}$.

Composition (•) is always a commutative operator:

$$(C \bullet B) \bullet A = C \bullet (B \bullet A)$$

$$= (B \bullet C) \bullet A$$

$$= B \bullet (C \bullet A)$$

$$= (C \bullet A) \bullet B$$

Composition distributes over matrix addition:

$$(A \bullet (B + C))(x) = A((B + C)(x))$$
$$= A(B(x) + C(x))$$
$$= A(B(x)) + A(C(x))$$
$$= (A \bullet B)(x) + (A \bullet C)(x)$$

$$(B + C) \bullet A)(x) = (B + C)(A(x))$$
$$= B(A(x)) + C(A(x))$$
$$= (B \bullet A)(x) + (C \bullet A)(x)$$

Being commutative and distributive, the composition of linear mappings behaves like multiplication. More accurately, multiplication of two matrices is really the composition of two mappings.

It is really easier to think of an $m \times n$ matrix A as the linear mapping A: $R^n \rightarrow R^m$. The columns of A show to where the unit vectors of R^n are mapped.

Matrices *are* linear mappings.

Partitioned mappings (matrices) compose (multiply) as follows:

$$B \cdot A = B\left[A_1 \;\cdots\; A_n\right] = \left[BA_1 \;\cdots\; BA_n\right]$$

$$B\left[A_1 \;\vdots\; \cdots \;\vdots\; A_p\right] = \left[BA_1 \;\vdots\; \cdots \;\vdots\; BA_p\right]$$

$$\begin{bmatrix} I & Q \\ M & P \\ M & P \\ N & Q \end{bmatrix} \begin{bmatrix} I & Q \\ M & P \\ M & P \\ N & Q \end{bmatrix} = \begin{bmatrix} I & Q \\ M & P \\ M & P \\ N & Q \end{bmatrix}$$

Recall from slide 6:

$$A\mathbf{x} = \sum_{i=1}^{n} x_i A_i = \begin{bmatrix} A_1 & \cdots & A_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$\text{So, } \begin{bmatrix} A_1 & \cdots & A_p \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} = \sum_{i=1}^{p} A_i x_i$$

$$\text{In general, } \begin{bmatrix} B_1 & \cdots & B_p \end{bmatrix} \begin{bmatrix} A_1 \\ \vdots \\ A_p \end{bmatrix} = \sum_{i=1}^{p} B_i A_i$$

Combining the partition rules of this and the previous slide…

As long as the B and A are partitioned conformably, the ij^{th} cell of the mapping B•A, or of the matrix product BA, will be :

$$\sum_k \mathrm{B}_{ik} \bullet \mathrm{A}_{kj}, \text{ or } \sum_k \mathrm{B}_{ik}\mathrm{A}_{kj}$$

Partitionwise multiplication is no different from elementwise multiplication. In fact, the elements *are* just the finest partitions, (1×1) partitions.

Matrix multiplication must have been first defined (by Cayley, Hamilton, Sylvester?) according to the interpretation of matrices as linear mappings. 'B times A' is 'B of A.'

2. The Linear Statistical Model

Information is an *n*-dimensional vector (*n*-tuple).

Models explain the more complicated by the less complicated.

Example: A car moves in a straight line at a constant velocity. At times $t_1,\ldots, t_n$ its position is observed to be $d_1,\ldots, d_n$. Time is the independent variable and distance is the dependent variable. We define our time and distance scales such that at time zero the car is at a distance zero.

We know that there will be some number r such that $d_i = r \cdot t_i$.

The linear model for this example is:

$$\mathbf{d}_{n\times 1} = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} r = \mathbf{t}_{n\times 1} r \quad \mathbf{r}_{1\times 1}$$

If we know t, the independent variable, we can explain the n-dimensional d in terms of the 1-dimensional r. Reducing complexity is the essence of modeling.

Predictive power is a by-product. When we know r, we can predict distances for new t_is.

Here is a model which explains a *t*-dimensional phenomenon in *k*-dimensions (hopefully k < t):

$$y_{(t\times 1)} = f\left(\beta_{(k\times 1)}\right)$$

f is some map from R^k to R^t. If *f* is a *linear* map, then we can express *f* as some $t\times k$ matrix X, and the model becomes:

$$y_{(t\times 1)} = X\left(\beta_{(k\times 1)}\right)$$

$$= X_{(t\times k)}\beta_{(k\times 1)}$$

y may look like a complicated t-dimensional phenomenon; but in reality it's just k-dimensional. This is understanding!

As we saw in the slide 6:

$$\underset{(t\times 1)}{\mathbf{y}} = \underset{(t\times k)}{\mathbf{X}}\,\underset{(k\times 1)}{\boldsymbol{\beta}}$$

$$= \begin{bmatrix} X_1 & \cdots & X_k \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}$$

$$= b_1 X_1 + \ldots + b_k X_k$$

$X\beta$ is a linear combination of the columns of X. This is a *subspace* of R^t of at most k dimensions. The model states that the t observations must fall within this subspace.

In other words, y is operating under k, rather than t, degrees of freedom. We have deepened our understanding, if $k < t$.

When the right β is found, prediction for new Xs is possible.

But reality is usually messy. Models are approximate:

$$\mathbf{y}_{(n\times 1)} \approx f\left(\boldsymbol{\beta}_{(k\times 1)}\right)$$

The equality is regained by adding a random error term; so the model becomes a *statistical* model:

$$\mathbf{y}_{(n\times 1)} = f\left(\boldsymbol{\beta}_{(k\times 1)}\right) + \mathbf{e}_{(n\times 1)}$$

Specifically, a *linear* model becomes a *linear* statistical model:

$$\mathbf{y}_{(n\times 1)} = \mathbf{X}_{(n\times k)}\boldsymbol{\beta}_{(k\times 1)} + \mathbf{e}_{(n\times 1)}$$

($E[\mathbf{e}] = 0$. And let $Var[\mathbf{e}] = \Sigma = \sigma^2\Phi$, which is symmetric $t\times t$.)

Two in-depth papers by the author on estimating the β of the linear statistical model and on predicting:

1. “Loss Prediction by Generalized Least Squares,” *PCAS* LXXXIII (1996), 436-489.

2. “Conjoint Prediction of Paid and Incurred Losses,” CAS Forum, Summer 1997, 241-379.

The author’s “Bible” on the subject:

George G. Judge, *et al.*, Introduction to the Theory and Practice of Econometrics, 2nd edition (Wiley, 1988).

But here follows a “quick and dirty” derivation of the estimator …

$$\mathbf{y}_{(n\times 1)} = \mathbf{X}_{(n\times k)}\,b_{(k\times 1)} + \mathbf{e}_{(n\times 1)}$$

$$\mathbf{W}\,\mathbf{y} = \mathbf{W}_{(k\times t)}\mathbf{X}_{(n\times k)}\,b + \mathbf{W}\,\mathbf{e}$$

If the square ($k\times k$) matrix WX has an inverse (non-singular):

$$\left(\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{W}\mathbf{y} = \left(\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{W}\mathbf{X}b + \left(\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{W}\mathbf{e}$$

$$\left(\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{W}\mathbf{y} = b + \left(\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{W}\mathbf{e}$$

$$E\left[\left(\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{W}\mathbf{y}\right] = E\left[b + \left(\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{W}\mathbf{e}\right]$$

$$= b$$

The last equality holds because β, W, and X are constants, and $E[\mathbf{e}] = 0$.

So we have a ***L***inear-in-**y** and ***U***nbiased ***E***stimator of β:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

$$= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}b + \mathbf{e}\right)$$

$$= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}b + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{e}$$

$$= b + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{e}$$

$$E\left[\hat{\beta}\right] = b$$

$$Var\left[\hat{\beta}\right] = Var\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{e}\right]$$

In the references it is derived that *Var*[A**e**] = A*Var*[**e**]A'. Hence:

$$Var\left[\hat{\beta}\right] = Var\left[(WX)^{-1}W\mathbf{e}\right]$$

$$= (WX)^{-1}W\,Var[\mathbf{e}]\,W'(X'W')^{-1}$$

$$= (WX)^{-1}W\Sigma W'(X'W')^{-1}$$

In the special case that $W\Sigma W' = X'W'$ (or $W = X'\Sigma^{-1}$):

$$Var\left[\hat{\beta}\right] = (WX)^{-1}$$

$$= (X'\Sigma^{-1}X)^{-1}$$

This special W exploits the variance of **e**, so that:

$$\forall W \left(X'\Sigma^{-1}X\right)^{-1} \leq \left(WX\right)^{-}W\Sigma W'\left(WX\right)^{-'}$$

The inequality is meaningful in the context of non-negative definite matrices (Appendix A of paper).

The special W makes for the ***B***est ***L***inear ***U***nbiased ***E***stimator:

$$\hat{\beta} = \left(WX\right)^{-}W\mathbf{y}$$

$$= \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}\mathbf{y}$$

$$= Var\left[\hat{\beta}\right]X'\Sigma^{-1}\mathbf{y}$$

The estimator is invariant to the scale of W. For any scalar $\alpha \neq 0$:

$$\left(\alpha \mathbf{W}\mathbf{X}\right)'\left(\alpha \mathbf{W}\mathbf{X}\right)^{-1}\left(\alpha \mathbf{W}\mathbf{X}\right)'\left(\alpha \mathbf{W}\mathbf{y}\right) = \left(\mathbf{W}\mathbf{X}\right)'\left(\mathbf{W}\mathbf{X}\right)^{-1}\alpha^{-1}\alpha\left(\mathbf{W}\mathbf{X}\right)'\left(\mathbf{W}\mathbf{y}\right) = \left(\mathbf{W}\mathbf{X}\right)'\left(\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{W}\mathbf{y}\right)$$

So the BLUE of β is invariant to the scale of *Var*[**e**]:

$$\hat{\beta} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}$$

$$= \left(\mathbf{X}'\left(\sigma^2\Phi\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\sigma^2\Phi\right)^{-1}\mathbf{y}$$

$$= \left(\mathbf{X}'\Phi^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Phi^{-1}\mathbf{y}$$

Later on σ^2 can be estimated (page 69 of paper). Shape, not scale!

The simplest shape is the identity, or $\Phi = I_t$. Then the BLUE is:

$$\hat{\beta} = \left(X' I_t^{-1} X\right)^{-1} X' I_t^{-1} \mathbf{y} = \left(X'X\right)^{-1} X'\mathbf{y}$$

To use the identity shape when the true shape is otherwise — perhaps we're ignoring information — is to settle for a not-as-good estimator; but it's still an unbiased estimator.

This "quick and dirty" approach (with the $k \times t$ matrix W) is related to the interesting subject of instrumental variables. Consult Judge's Econometrics textbook, pages 577-579.

Now follows a geometrical interpretation of the linear statistical model, otherwise known as *least squares* …

$$\hat{\beta} = \left(X'X\right)^{-1}X'y$$

can be shown (Judge, 190-192) to minimize the function:

$$f(b) = (y - Xb)'(y - Xb)$$

which function represents the square of the Euclidean distance between **y** from Xβ. But $X_{(t\times k)}\beta_{(k\times 1)}$ is a *k*-dimensional subspace of the R^t which **y** inhabits. (Recall from slide 18 that Xβ is a linear combination of the columns of X.)

To minimize *f*(β) is to find the point closest to **y** of the subspace spanned by X. At this point, at this particular Xβ, **y** drops a perpendicular to the subspace: $(y - X\hat{\beta}) \perp X\hat{\beta}$

The closer **y** is to the subspace, the more tempted we are to say that **y** *is* $\mathrm{X}\hat{\beta}$ (barring a little randomness, which we can quantify and manage). And a *t*-dimensional phenomenon is more or less reduced to *k* dimensions. To repeat, this is a deeper understanding, if $k < t$.

In general, $\hat{\beta} = (\mathrm{X}'\Phi^{-1}\mathrm{X})^{-1}\mathrm{X}'\Phi^{-1}\mathbf{y}$ minimizes:

$$f(b) = (\mathbf{y} - \mathrm{X}b)'\Phi^{-1}(\mathbf{y} - \mathrm{X}b)$$

This represents a generalized Euclidean distance between **y** and $\mathrm{X}\beta$, since Φ will penalize differences in some directions more heavily than differences in other directions. With a non-identity (and positive definite) Φ, constant distance from a center takes the form of an ellipse, rather than that of a circle.

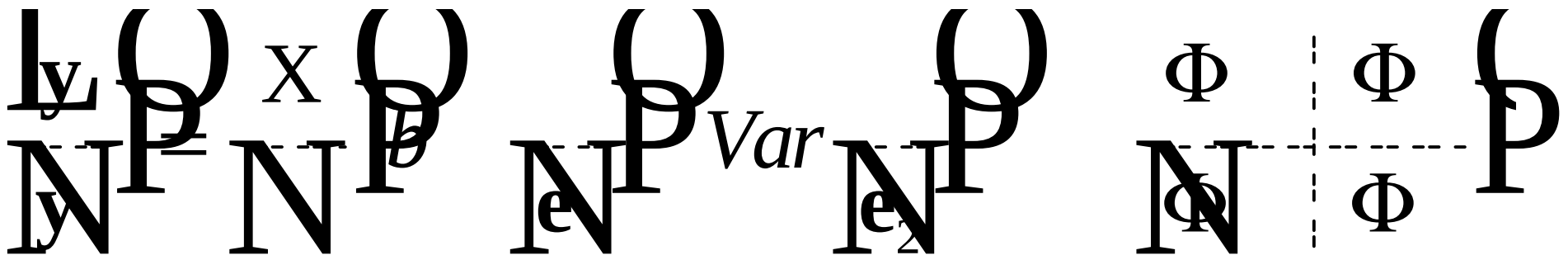
Prediction

Estimating the parameter (β) of a statistical model usually isn't enough. Typically, we'll want to predict new **y**s, given new Xs.

Also, we ought to know how much the phenomenon can vary from our prediction — to know the variance of the prediction from its expected value.

Predictions can be correlated with what we've observed. We can't always have the simple world of i.i.d. (independent, identically distributed)

The formulation, with the help of partitioned matrixes:



$_1$ rows of observations, t_2

$_1$ is t_1 is $t_1\times$ β is $k\times1$, is $t_1\times$, Φ_{12} is $\times t_2$, etc. The Φ matrix is symmetric, so Φ_{12} $_{21}'$.

$\mathbf{y}_1$ s are known (or taken for granted).

$\mathbf{y}_2$ contains missing values. We want to estimate (or predict) it.

$\mathbf{y}_2$ is …

$$\hat{\mathbf{y}}_2 = \mathbf{X}_2\hat{\boldsymbol{\beta}} + \boldsymbol{\Phi}_{21}\boldsymbol{\Phi}_{11}^{-1}\left(\mathbf{y}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}\right)$$

$$\text{where } \hat{\boldsymbol{\beta}} = \left(\mathbf{X}_1'\boldsymbol{\Phi}_{11}^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\boldsymbol{\Phi}_{11}^{-1}\mathbf{y}_1$$

$$Var\left[\mathbf{y}_2 - \hat{\mathbf{y}}_2\right] = s^2\left(\boldsymbol{\Phi}_{22} - \boldsymbol{\Phi}_{21}\boldsymbol{\Phi}_{11}^{-1}\boldsymbol{\Phi}_{12}\right)$$

$$+\left(\mathbf{X}_2 - \boldsymbol{\Phi}_{21}\boldsymbol{\Phi}_{11}^{-1}\mathbf{X}_1\right)Var[\hat{\boldsymbol{\beta}}]\left(\mathbf{X}_2 - \boldsymbol{\Phi}_{21}\boldsymbol{\Phi}_{11}^{-1}\mathbf{X}_1\right)'$$

$$\text{where } Var[\hat{\boldsymbol{\beta}}] = s^2\left(\mathbf{X}_1'\boldsymbol{\Phi}_{11}^{-1}\mathbf{X}_1\right)^{-1}$$

$\Phi_{21} \neq 0$ allows errors in the observations to affect the predictions.

Looks nasty, but really quite gentle. Proof in Appendix C of "Conjoint Prediction." Also see pages 68f. of this paper.

3. Credibility and Prior Information

"Readers who have come this far may conclude from what they've read that casualty actuarial science is the study and application of the theory of credibility, and that's all. Is it all?"

Matthew Rodermund, Foundations of Casualty Actuarial Science, 19.

It's hard to answer "No" to Rodermund's question. Actuaries love the '$Z \times A + (1 - Z) \times E$' credibility formula. It blends observation ('*A*' for 'actual') and prior opinion ('*E*' for 'expected').

But a (linear) statistical model does the blending even better …

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad Var \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$$

Observations only: t_1 rows from source 1, and t_2 rows from source 2.

Simple variance structure in that the sources do not covary (off-diagonal Σs are zero).

The BLUE of β is …

$$\hat{\beta} = \begin{bmatrix} X_1'\Sigma_1^{-1}X_1 & 0 \\ 0 & X_2'\Sigma_2^{-1}X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'\Sigma_1^{-1}y_1 \\ X_2'\Sigma_2^{-1}y_2 \end{bmatrix}$$

$$= \begin{bmatrix} X_1'\Sigma_1^{-1} & 0 \\ 0 & X_2'\Sigma_2^{-1} \end{bmatrix} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'\Sigma_1^{-1} & 0 \\ 0 & X_2'\Sigma_2^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X_1'\Sigma_1^{-1}y_1 \\ X_2'\Sigma_2^{-1}y_2 \end{bmatrix}$$

$$= \begin{bmatrix} X_1 & X_2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} X_1'\Sigma_1^{-1} & 0 \\ 0 & X_2'\Sigma_2^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X_1'\Sigma_1^{-1}y_1 \\ X_2'\Sigma_2^{-1}y_2 \end{bmatrix}$$

$$= \left(X_1'\Sigma_1^{-1}X_1 + X_2'\Sigma_2^{-1}X_2\right)^{-1}\left(X_1'\Sigma_1^{-1}y_1 + X_2'\Sigma_2^{-1}y_2\right)$$

$$= Var\left[\hat{\beta}\right]\left(X_1'\Sigma_1^{-1}y_1 + X_2'\Sigma_2^{-1}y_2\right)$$

Not too bad when one has a feel for partitioned matrices (slide 14).

Next to last equation looks like a (matrix) weighted average. This can be made explicit …

$$\hat{\beta} = \left(X_1'\Sigma_1^{-1}X_1 + X_2'\Sigma_2^{-1}X_2\right)^{-1}\left(X_1'\Sigma_1^{-1}\mathbf{y}_1 + X_2'\Sigma_2^{-1}\mathbf{y}_2\right)$$

$$= \left(X_1'\Sigma_1^{-1}X_1 + X_2'\Sigma_2^{-1}X_2\right)^{-1}\left(X_1'\Sigma_1^{-1}X_1\left(X_1'\Sigma_1^{-1}X_1\right)^{-1}X_1'\Sigma_1^{-1}\mathbf{y}_1 + \cdots\right)$$

$$= \left(X_1'\Sigma_1^{-1}X_1 + X_2'\Sigma_2^{-1}X_2\right)^{-1}\left(X_1'\Sigma_1^{-1}X_1\hat{\beta}_1 + X_2'\Sigma_2^{-1}X_2\hat{\beta}_2\right)$$

$$= \left(Var^{-1}\left[\hat{\beta}_1\right] + Var^{-1}\left[\hat{\beta}_2\right]\right)^{-1}\left(Var^{-1}\left[\hat{\beta}_1\right]\hat{\beta}_1 + Var^{-1}\left[\hat{\beta}_2\right]\hat{\beta}_2\right)$$

The estimator of the two-source model weights the estimators of the one-source models according to the inverses of the variances of those estimators (harmonic average — Appendix A).

Since β is $k\times 1$, this is credibility in k dimensions.

$$Var\left[\hat{\beta}\right] = \left(Var^{-1}\left[\hat{\beta}_1\right] + Var^{-1}\left[\hat{\beta}_2\right]\right)^{-1}$$

$$Var^{-1}\left[\hat{\beta}\right] = Var^{-1}\left[\hat{\beta}_1\right] + Var^{-1}\left[\hat{\beta}_2\right]$$

$$> Var^{-1}\left[\hat{\beta}_1\right],\ Var^{-1}\left[\hat{\beta}_2\right]$$

$$Var\left[\hat{\beta}\right] < Var\left[\hat{\beta}_1\right],\ Var\left[\hat{\beta}_2\right]$$

As Appendix A explains (using positive definite matrices), the two-source estimator is of less variance than either of the one-source estimators. The more knowledge, the better.

The second source doesn't actually have to be observed. It can be theory, opinion, or guess — anything on which you're willing to rely.

The second source, or the prior information, doesn't even have to be a complete model, so that a β_2 could be estimated:

$$\hat{\beta} = \left(X'\Sigma^{-1}X + R'V^{-1}R\right)^{-1}\left(X'\Sigma^{-1}y + RV^{-1}r\right)$$

If $k \times k$ $R'V^{-1}R$ is of rank $j \leq k$, the variance of the two-source estimator will be improved along j orthogonal axes. If $j<k$, then $R'V^{-1}R$ is singular and the second source cannot produce its own estimate for β. But it still improves the mixed estimator in j out of k dimensions. See Appendix A.

Sorry, no example is given in this presentation. But the paper works out several examples.

Benefits of statistical modeling to credibility:

- Provides an systematic and orderly framework.
- Furnishes variances of estimates and predictions, as well as means.
- Extends credibility from one to *k* dimensions.

4. Credibility and Random Effects

Hardest part of the paper — Sections 9 and 10, and Appendix E.

Given n related groups, with non-covarying **e**s and **v**s:

$$\mathbf{y}_i \underset{t_i\times 1}{} = \mathrm{X}_i \underset{t_i\times k}{} b_i \underset{k\times 1}{} + \mathbf{e}_i \underset{t_i\times 1}{} \quad Var[\mathbf{e}_i] = \Sigma_i \underset{t_i\times t_i}{}$$

$$\text{But } b_i = b_0 + \mathbf{v}_i,\ Var[\mathbf{v}_i] = \mathrm{V} \underset{k\times k}{}$$

$$\text{So } \mathbf{y}_i = \mathrm{X}_i b_0 + \mathrm{X}_i \mathbf{v}_i + \mathbf{e}_i$$

$$= \mathrm{X}_i b_0 + \tau_i,\ Var[\tau_i] = \mathrm{X}_i \mathrm{V} \mathrm{X}_i' + \Sigma_i$$

This is just a linear statistical model, so we can estimate β_0, the β_is, and any predictions built on the β_is.

If V is unknown, it may be estimated by the method of variance components (ML method also possible — Appendix F).

If V is large, the β_is are free and the groups have much credibility. If V is small, the β_is are close to β_0 and the groups have little credibility.

Appendix E discusses the random-effects model in detail. A beautiful result is that the simple average of the estimates of the β_is must equal the estimate of β_0. In effect, credibility democratizes the groups.

Section 10 presents a random-effects trend model, a two-dimensional credibility problem.

5. Conclusion

Arthur Bailey challenged (1945-1950) classical statistics with three problems which justified his "greatest accuracy credibility:"

• Use of prior information in estimation ($Z \times A + (1 - Z) \times E$)

• Estimating for an individual that belongs to a heterogeneous population (merit and experience rating — a fruitful subject for Bayesian credibility. See Appendix B)

• Estimating for groups together, which is more accurate than estimating each separately (the random-effects model. See Sections 9 and 10, and Appendix E)

The paper shows how modern statistics solves these problems, to the legitimization and enrichment of credibility.

Corrections

Page:Line	Text	Change
74:14	0.904	0.113
74:15	4.240	0.606
87:1	ends of the introduction and	end
118:3	element	variance
121:1	subscript *i* is ranges	subscript *i* ranges
137:2	$\sum_1^n((\mathbf{y}_i-\hat{\mu})-(\nu-\hat{\mu}))((\mathbf{y}_i-\hat{\mu})-(\nu-\hat{\mu}))$	$\sum_1^n((\mathbf{y}_i-\hat{\mu})-(\nu-\hat{\mu}))((\mathbf{y}_i-\hat{\mu})-(\nu-\hat{\mu}))'$
147:11 and 14 (three times)	$X_i T_i^{-1} X_i'$	$X_i' T_i^{-1} X_i$

Vice President of Actuarial Research and Development
American Re-insurance Company (as of March 16, lhalliwell@amre.com)