Combining Information in Statistical Modeling

Daniel Pena

The American Statistician, Vol. 51, No. 4 (Nov., 1997), 326-332.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28199711%2951%3A4%3C326%3ACIISM%3E2.0.CO%3B2-T

The American Statistician is currently published by American Statistical Association.

Combining Information in Statistical Modeling

Daniel PEÑA

How to combine information from different sources is becoming an important statistical area of research under the name of Meta-Analysis. This paper shows that the estimation of a parameter or the forecast of a random variable can also be seen as a process of combining information. It is shown that this approach can provide some useful insights on the robustness properties of some statistical procedures, and it also allows the comparison of statistical models within a common framework. Some general combining rules are illustrated using examples from ANOVA analysis, diagnostics in regression, time series forecasting, missing value estimation, and recursive estimation using the Kalman filter.

KEY WORDS: Analysis of variance; Diagnostics; Forecasting; Kalman filer; Linear regression; Meta-Analysis; Time series.

1. INTRODUCTION

The proliferation of statistical studies in many areas of research has led to a growing interest in developing methods of combining information from different studies. This area of research was named Meta-Analysis by Glass (1976), and it has received considerable attention in the social sciences (Hedges and Olkin 1985; Wolf 1986). Examples of the use of Meta-Analysis in other scientific areas can be found in Utts (1991), Mosteller and Chalmers (1992), Dear and Begg (1992), Hedges (1992), Draper et al. (1992), and the references included in these papers.

In this paper we show that analyzing estimation problems from the point of view of how information is combined can provide useful insights about the properties of the estimates and their robustness. The paper presents a brief and idiosyncratic review of some aspects of combining information arising in linear models and time series analysis, with emphasis on pooling rules and diagnostics. In particular, it is shown that some simple rules of combining information reveal the effect of different group sizes in ANOVA problems, show the advantage of estimating growth with stochastic instead of deterministic trends, are useful to generalize missing value estimation procedures in linear time series models, suggest new diagnostics in linear regression, and provide a simple understanding on the recursive updating of dynamic linear models estimates.

The process of estimation of an unknown quantity θ, that can be a fixed parameter or a random variable, can always be seen as a process of combining information from the data about θ. Understanding this process is important to evaluate the performance of an estimation rule. Often, we have independent sources of information about θ. For instance, a sample of size n can be considered as a set of j independent samples of size n_j, with $\sum n_j = n$. If we have unbiased and independent estimates of the unknown quantity $\hat{\theta}_1, \ldots, \hat{\theta}_n$, they are usually combined according to the following well-known rule.

Rule 1. Given n unbiased and independent estimates $\hat{\theta}_i$ of a scalar parameter θ with nonzero variances σ_i^2, the best (minimum variance) linear unbiased estimate (BLUE) of $\theta, \hat{\theta}_T$, is given by

$$\hat{\theta}_T = \sum_{i=1}^{n} \frac{\sigma_i^{-2}}{\sum \sigma_j^{-2}} \hat{\theta}_i \tag{1.1}$$

and the variance of the pooled estimate $\hat{\theta}_T$ is given by $(\sum \sigma_j^{-2})^{-1}$.

This rule is commonly applied in Meta-Analysis for the parametric estimation of effect size from a series of experiments (Hedges and Olkin 1985, chap. 6). Note that (1.1) is appropriate only in the fixed-effects model in which the assumption of unbiasedness of the $\hat{\theta}_i$ is crucial. If the possibility of unknown biases is entertained, as in the random effects model $\hat{\theta}_i = \theta_i + e_i, \theta_i = \theta + \varepsilon_i$, variance estimators like (1.1) sharply underestimate the actual uncertainty about θ.

This paper generalizes Rule 1 for dependent and vector-valued unknown quantities, and applies it to several common statistical estimation problems that are presented as particular cases of the general problem of combining different sources of information. The paper is organized as follows. In Section 2 we show that looking at ANOVA from the perspective of Rule 1 allows a simple understanding of the robustness properties of the estimators and of the importance of equal sample size in all groups. Section 3 shows that this approach is useful to compare two time series models for forecasting growth. Section 4 analyzes the estimation of missing values in linear time series, and shows how this approach leads to a simple solution for dealing with the end effects. Section 5 discusses how the structure of an estimator in linear regression can suggest new diagnostics to evaluate the data robustness of the fitted model. Section 6 presents the more general rule for combining information used in the paper, and applies it to derive recursive estimators and diagnostic measures. Finally, Section 7 contains some concluding remarks.

Daniel Peña is Professor of Statistics, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, 28903 Getafe, Spain. This research was supported by the Cátedra BBV de Calidad and DGICYT, Spain under Grants PB-93-0232 and PB-94-0374. The author thanks Victor Guerrero, Irwin Guttman and Ana Justel for very stimulating discussions and comments on a first draft of this paper. The excellent comments of a referee have been very useful in improving the paper.

2. ROBUSTNESS IN ANOVA PROBLEMS

Suppose we have two independent samples $(x_1, \ldots, x_n)$, $(y_1, \ldots, y_m)$ from the same population, and we want to estimate its mean (μ) and variance (σ^2). Assuming normality, and calling $\bar{x}$, $\bar{y}$ the sample means and s_1^2 and s_2^2 the unbiased sample variances, the application of Rule 1 leads to

$$\hat{\mu} = \frac{n}{n+m}\,\bar{x} + \frac{m}{n+m}\,\bar{y} \tag{2.1}$$

and

$$\hat{\sigma}^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}. \tag{2.2}$$

The result in (2.2) follows because in normal samples $\text{var}(s^2) = 2\sigma^4/(n-1)$. When the population is not normal, $\hat{\mu}$ is still the best linear unbiased estimator, whereas $\hat{\sigma}^2$ is not. This happens because the variance of $\bar{x}$ is always σ^2/n, and then Rule 1 always leads to (2.1), whatever the parent population. However, the variance of s_i^2 for nonnormal populations is usually a more complex function of n: for instance, when the population is χ^2, it is given by $\sigma^4/g(n)$, where $g(n)$ is an increasing nonlinear function of n. Therefore, for nonnormal populations the general estimate of σ^2 given by Rule 1 is

$$\hat{\sigma}^2 = \frac{g(n)}{g(n)+g(m)}\,s_1^2 + \frac{g(m)}{g(n)+g(m)}\,s_2^2. \tag{2.3}$$

If $n = m$, (2.2) and (2.3) are both equal to $(s_1^2 + s_1^2)/2$, and the estimate is robust: it is BLUE whatever the population. However, if the sample sizes, n and m, are very different, then (2.2) and (2.3) will produce different answers.

This result will also be true in ANOVA problems. Suppose we have k different groups. Then, under the standard hypothesis of variance homogeneity in all groups, the residual variance estimate is given by

$$s_R^2 = \sum \left(\frac{n_i - 1}{n-k}\right) s_i^2 \tag{2.4}$$

where $s_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ is the unbiased variance estimate in group i. Again, if the population is not normal, (2.4) may be a very bad estimate, and will be in contradiction with Rule 1. However, when the sample size is equal in all groups, it will be BLUE, whatever the population.

3. COMPARING ESTIMATES OF GROWTH IN TIME SERIES

Two procedures often used for forecasting the future growth of a given time series are: 1. detrend the observed data by regressing the observations on time, fit a stationary time series model to the residuals from this regression, and build the forecast as the sum of the deterministic trend and the forecast of the stationary residual; and 2. difference the series, fit a stationary ARMA model in the first difference of the series, and forecast the series using the ARIMA model (Box and Jenkins 1976). Typically, models built in this way include a constant for many economic time series. The decision on which of these two procedures should be used is made by testing whether or not the series has one unit root. However, the available tests are not very powerful, especially for short time series [see, for instance, De Jong et al. (1992)], and therefore it is important to understand the consequences of using these models.

Let y_t be the time series data, and let us assume, for the sake of simplicity, that the sample size is $n = 2m + 1$. Let $t = \{-m, \ldots, 0, \ldots, +m\}$. Then the least squares estimator of the slope in the regression on time

$$y_t = \beta_0 + \beta_1 t + u, \qquad E(u) = 0, \qquad \text{var}(u) = \sigma^2 \tag{3.1}$$

is given by

$$\hat{\beta}_1 = \frac{\sum t y_t}{\sum t^2} = \left(2\sum_{i=1}^{m} i^2\right)^{-1} \sum_{t=1}^{m} t(y_t - y_{-t}). \tag{3.2}$$

Calling $b_t = y_t - y_{t-1}$ the observed growth at time t, and after some straightforward manipulations that are shown in Peña (1995), the estimate of the slope can be written as

$$\hat{\beta}_1 = \sum_{j=1}^{m} \omega_j (b_j + b_{1-j}) \tag{3.3}$$

where the weights ω_j are given by

$$\omega_j = a_0 - a_1(j^2 - j) \qquad j = 1, \ldots, m$$

where $a_0 = 3/(2m+1)$ and $a_1 = 3/m(2m+1)(m+1)$, and add up to one. Therefore, the estimated growth $\hat{\beta}_1$ is a weighted mean of all the observed growths b_j, with decreasing weight from the center of the sample. The maximum weight is given to b_1 and b_0, which corresponds to the observed growth in the middle of the sample period, and the minimum weight is given to b_m and b_{1-m}, the first and last observed growth. The weights decrease quadratically from the middle of the sample.

Note that in the assumption that the linear model (3.1) holds, the $2m$ values $b_t\,(t = -m+1, \ldots, m)$ are unbiased estimates for β. The covariance matrix of these $2m$ estimates is the Toeplitz matrix:

$$V = \begin{bmatrix} 2\sigma^2 & -\sigma^2 & 0 & \cdots & 0 \\ -\sigma^2 & 2\sigma^2 & -\sigma^2 & & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & -\sigma^2 \\ 0 & \cdots & \cdots & -\sigma^2 & 2\sigma^2 \end{bmatrix}. \tag{3.4}$$

Now we can set the following rule [see, for instance, Newbold and Granger (1974)].

Rule 2. Given a vector $\hat{\theta}$ of unbiased estimators of a parameter θ with covariance matrix V, the best (in the mean-squared sense) linear unbiased estimator of θ is given by

$$\hat{\theta}_T = (1'V^{-1}1)^{-1}1'V^{-1}\hat{\theta} \tag{3.5}$$

where $1' = (1\ 1 \ldots 1)$, and the variance of $\hat{\theta}_T$ is given by

$$\text{var}(\hat{\theta}_T) = (1'V^{-1}1)^{-1}. \tag{3.6}$$

This Rule 2 is a particular case of the Rule 5 that is proved in the Appendix.

The inverse of the Toeplitz matrix (3.4) has been studied by Shaman (1969), who obtained the exact inverse of a first-order moving average process. As V can be interpreted as the covariance matrix of a noninvertible ($\theta = 1$) first-order moving average process, then we have

$$V^{-1} = \frac{1}{\sigma^2(2m+1)} \times \begin{bmatrix} 2m & 2m-1 & 2m-2 & \cdots & 1 \\ 2m-1 & 2(2m-1) & 2(2m-2) & \cdots & 2 \\ 2m-2 & 2(2m-2) & 3(2m-2) & \cdots & 3 \\ \vdots & \vdots & \vdots & & \\ 2 & 4 & 6 & \cdots & 2m-1 \\ 1 & 2 & 3 & \cdots & 2m \end{bmatrix}.$$

It is easy to show that the estimator (3.3) can also be obtained by applying Rule 2 to the unbiased but correlated estimates b_t.

When an ARMA model is fitted to the residuals of the regression model, the equation for the h-step-ahead forecast, where we call $\hat{y}_t(h) = E[y_{t+h}|y_t, y_{t-1}, \ldots]$, is

$$\hat{y}_t(h) = \hat{\beta}_0 + \hat{\beta}_1 h + \hat{n}_t(h) \tag{3.7}$$

where $\hat{n}_t(h)$ is the forecast of the zero-mean stationary process fitted to the residuals. As for a stationary process, the long-run forecast converges to the mean, $\hat{n}_t(h) \to 0$, and the parameter $\hat{\beta}_1$ is the long-run estimated growth of the time series.

Let us compare (3.7) with the growth estimate provided by the integrated ARIMA model

$$\nabla y_t = \beta + n_t \tag{3.8}$$

where $\nabla = 1 - B$, B is the backward shift operator $By_t = y_{t-1}$, and n_t follows a zero-mean stationary and invertible ARMA model. Letting V denote the covariance matrix of n_t, the estimate of β in (3.8) is given by the generalized least squares estimator

$$\hat{\beta} = (1'V^{-1}1)^{-1}1'V^{-1}b \tag{3.9}$$

where the vector b has components $b_t = y_t - y_{t-1}$. Then it is well known (Fuller 1976) that $\bar{b} = (1/(n-1))\sum b_i$ is asymptotically unbiased for β. When n is large, the expected forecast h periods ahead is given by

$$\hat{y}_t(h) = \bar{b}h + \hat{n}_t(h) \tag{3.10}$$

where $\hat{n}_t(h)$ is the h-step-ahead forecast of the stationary process n_t. As for h large, the $\hat{n}_t(h)$ will go to zero; the long-run growth will be estimated by a weighted average with uniform weighing of the observed growths b_t.

In summary, the two models forecast future growth by using a weighted average of the observed growths in the sample. Linear regression gives minimum weight to the last observed growth and maximum weight to the center of the sample period. The ARIMA model gives uniform weighting to all the years in the sample. A comparison of the forecasting structure of these and other models used for forecasting growth can be found in Peña (1995).

4. ESTIMATING MISSING VALUES IN LINEAR TIME SERIES

Suppose a Gaussian stationary time series y_t follows the general representation

$$y_t = \sum_{i=1}^{\infty} \pi_i y_{t-i} + a_t \tag{4.1}$$

where a_t is a white noise process with variance σ_a^2. Then, if the value y_T is missing, we can obtain an unbiased estimate of it by using

$$\hat{y}_T^{(0)} = \sum_{i=1}^{\infty} \pi_i y_{T-i} \tag{4.2}$$

and this estimate will have variance σ_a^2. Also, from (4.1) we can write for all j such that $\pi_j \neq 0$

$$y_T = \pi_j^{-1}\left(y_{T+j} - \sum_{\substack{i=1 \\ i\neq j}}^{\infty} \pi_i y_{T+j-i}\right) - a_{T+j}/\pi_j \tag{4.3}$$

and therefore we can obtain additional "backward" unbiased estimates of y_T from (4.3) by

$$\hat{y}_T^{(j)} = \pi_j^{-1}\left(y_{T+j} - \sum_{i\neq j} \pi_i y_{T+j-i}\right) \tag{4.4}$$

with variance σ_a^2/π_j^2. As all these estimates are unbiased and independent given the observed data, the best linear unbiased estimate of the missing value y_T is readily obtained by applying Rule 1 to yield

$$\hat{y}_T = \sum_{j=0}^{\infty} \frac{\pi_j^2}{\sum \pi_j^2}\, \hat{y}_T^{(j)} \tag{4.5}$$

where $\pi_0 = -1$. It is easy to show (Grenander and Rosenblatt 1957; Peña and Maravall 1991) that this estimate is equivalent to the well-known expression for the missing value estimation in a Gaussian stationary time series

$$\hat{y} = -\sum_{i=0}^{\infty} \rho_i^D (y_{T-i} + y_{T+i}) \tag{4.6}$$

where the ρ_i^D are the inverse autocorrelation coefficients. An advantage of formulation (4.5) is that it provides a clear understanding of how to proceed when the missing value is near the extremes of the series so that the two-sided symmetric filter (4.6) has to be truncated. Then we have to combine (4.2) with the $n - T$ backward estimates (4.4) that are available, and the exact formula for the finite sample interpolator is

$$\hat{y}_{T,F} = \sum_{j=0}^{n-T} \frac{\pi_j^2}{\sum_0^n \pi_j^2}\, \hat{y}_T^{(j)}. \tag{4.7}$$

This idea can be easily extended to groups of missing observations. We will illustrate it here with an example: suppose we have an AR(1) process in which the values y_T and y_{T+1} are missing. Then for y_T we have the two estimates

$$\hat{y}_T^{(0)} = \phi y_{T-1} \tag{4.8}$$

with variance σ_a^2 and the backward estimate

$$\hat{y}_T^{(2)} = \phi^{-2} y_{T+2} \tag{4.9}$$

with variance $\sigma_a^2(1+\phi^2)/\phi^4$. The best linear unbiased estimate will be

$$\hat{y}_T = \frac{\phi(1+\phi^2)}{1+\phi^2+\phi^4}\, y_{T-1} + \frac{\phi^2}{1+\phi^2+\phi^4}\, y_{T+2} \tag{4.10}$$

which agrees with the general formula obtained by a different approach in Peña and Maravall (1991). The estimate of $\hat{y}_{T+1}$ will be similar to (4.10), but with the roles of y_{T-1} and y_{T-2} reversed.

5. SENSITIVITY ANALYSIS IN LINEAR REGRESSION

It is well known that in the linear regression model

$$y = \beta_0 + \beta_1 x + u, \qquad E(u) = 0, \qquad \text{var}(u) = \sigma^2 \tag{5.1}$$

the least square estimate of the slope is given by

$$\hat{\beta} = \sum w_i b_i \tag{5.2}$$

where $w_i = (x_i - \bar{x})^2 / \sum (x_i - \bar{x})^2$ is a set of weights $(w_i \geq 0, \sum w_i = 1)$ and the b_i are estimates of the slope that can be built up by using the sample data:

$$b_i = \frac{y_i - \bar{y}}{x_i - \bar{x}} \tag{5.3}$$

where $w_i b_i = 0$ if $x_i = \bar{x}$. These estimates are not independent because $a_x' b = 0$, where $a_x' = ((x_1 - \bar{x}) \cdots (x_n - \bar{x}))$ and $b = (b_1, \ldots, b_n)$. They have a singular covariance matrix:

$$S_b = D_x^{-1}(I - 1/n1\ 1')D_x^{-1}\sigma^2 \tag{5.4}$$

where D_x is a diagonal matrix such that the ith diagonal element is the ith element of a_x, that is, $\text{diag}(D_x) = a_x$. Then we can use the following rule.

Rule 3. Given n dependent estimates $\hat{\theta}_i$ with singular covariance matrix S_θ, the best linear unbiased estimator of θ is given by

$$\hat{\theta}_T = (1' S_\theta^- 1)^{-1} 1' S_\theta^- \hat{\theta} \tag{5.5}$$

where S_θ^- is a generalized inverse of S_θ, and the variance of the pooled estimator $\hat{\theta}_T$ is

$$\text{var}(\hat{\theta}_T) = (1' S_\theta^- 1)^{-1}.$$

It is straightforward to check that a generalized inverse of (5.4) is given by

$$S_b^- = D_x D_x \sigma^{-2} \tag{5.6}$$

because $1' D_x = 0$, and if we apply (5.5) to (5.6) and (5.3) as $1' S_b^- 1' = 1/\sigma^2 \sum (x_i - \bar{x})^2$, we obtain (5.2). In summary, (5.2) is again the BLUE estimate given the estimates b_i.

Equation (5.2) shows that the leverage $(x_i - \bar{x})^2 / \sum (x_i - \bar{x})^2$ determines the potential influence of an observation on the estimated slope of the regression line, whereas the observed effect also depends on b_i. Because $\hat{\beta}$ is the sum of n components $w_i b_i$, a measure of the relative weight of each term on determining $\hat{\beta}$ can be built by $|w_i b_i - \bar{w}b|$ where $\bar{w}b = \hat{\beta}/n$. Assuming $\hat{\beta} \neq 0$, this measure can be written in relative terms as

$$\delta_i = \left| \frac{n w_i b_i}{\hat{\beta}} - 1 \right| = \left| \frac{(x_i - \bar{x})(y_i - \bar{y})}{S_{xy}} - 1 \right| \tag{5.7}$$

where $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$. Note that δ_i is a measure of the influence of a point (x_i, y_i) on the slope, whereas the usual statistics of influence, as the one due to Cook (1977), tries to identify both outliers and influential points. Also, Cook's statistic can be very affected by masking (Peña and Yohai 1995), whereas δ_i is not. For instance, Table 1 presents a set of artificial data with three large influential observations that are not identified either by D_i (Cook's statistics) or by the studentized residual (t_i) as extremes, but they are indicated as the most influential on the slope by the statistic (5.7).

Instead of looking for a summary, we may look at the whole distribution of the $w_i b_i$ terms in order to find those terms that seem to be very different from the others. Also, if the estimate $\hat{\beta}$ is close to zero, the statistics (5.7) will be very unstable, and it will be more useful to consider $|w_i b_i - \bar{w}b|$ or to study directly the distribution of the terms $w_i b_i$.

Consider now the multiple regression model

$$Y = X\beta + U \tag{5.8}$$

where X is $n \times p$, $E(U) = 0$, $\text{var}(U) = \sigma^2 I$, and I is the identity matrix. We suppose, to simplify the presentation and without loss of generality, that all the variables have zero mean. Then it is well known that each of the components of $\hat{\beta}$ can be written as

$$\hat{\beta}_j = \sum_{i=1}^n w_{ij} b_{ij} \tag{5.9}$$

where

$$b_{ij} = \frac{y_i - \bar{y}}{e_{ij \cdot R}} \tag{5.10}$$

Table 1.

x_i	1	2	3	4	5	6	7	8	9	10	17	17	17
y_i	2	3	4	5	6	7	8	9	10	11	25	25	25
D_i	.25	.10	.03	.01	.00	.01	.02	.05	.09	.15	.16	.16	.16
t_i	1.4	1.0	.6	.3	−0.1	−0.4	−0.7	−1.1	−1.5	−2.0	.9	.9	.9
δ_i	.42	.08	.21	.46	.66	.82	.93	.99	1.01	.99	1.85	1.85	1.85

and $e_{ij\cdot R}$ is the ith component of the vector of residuals $e_{j\cdot R}$ obtained by regressing x_j on all of the other explanatory variables. That is, if $X_{(j)}$ is a matrix without the jth column x_j and $\hat{\gamma}_j = (X'_{(j)}X_{(j)})^{-1}X'_{(j)}x_j$ is the least square estimate of this regression, then $e_{j\cdot R} = x_j - X_{(j)}\hat{\gamma}_j$. The weight w_{ij} is given by $e^2_{ij\cdot R}/\sum e^2_{ij\cdot R}$.

Suppose that we are mainly interested in some regression coefficient $\hat{\beta}_j$. Then the usual diagnostic statistics that look at the change on the whole vector of parameter estimates may not be useful. However, the weights w_{ij} provide a natural and simple way of looking at the potential effect of an observation. These weights can be computed from

$$w_{ij} = \frac{(x_{ij} - x'_{i(j)}(X'_{(j)}X_{(j)})^{-1}X'_{(j)}x_j)^2}{x'_j(I - H_{(j)})x_j} \tag{5.11}$$

where $x_{i(j)}$ is the ith row of $X_{(j)}$ and $H_{(j)}$ is the hat or projection matrix built without variable X_j. A plot of the variables w_{ij} can be useful to judge the robustness of one estimate to the given sample.

As in the simple regression case, a measure of the influence of point (x_iy_i) on the estimation of $\hat{\beta}_j$ can be built by

$$\delta_i = \left| \frac{nw_{ij}b_{ij}}{\hat{\beta}_j} - 1 \right| \tag{5.12}$$

and we can apply to this measure the comments indicated for (5.7).

A different problem occurs when we have two independent samples of n_i data points $(X_iY_i), i = 1, 2$ in which we have obtained $\hat{\beta}_i = (X'_iX_i)^{-1}X'_iY_i$ with covariance $s_i^2(X'_iX_i)^{-1}$, and we want to combine both estimates to obtain the BLUE. Then we can use the following rule.

Rule 4. If $\hat{\theta}_1$ is an unbiased estimator of θ with covariance matrix V_1 and $\hat{\theta}_2$ is also unbiased for θ with covariance V_2 and these two estimates are independent, the best linear unbiased estimator (minimizing the trace of the variance covariance matrix) is given by

$$\hat{\theta}_T = (V_1^{-1} + V_2^{-1})^{-1}V_1^{-1}\hat{\theta}_1 + (V_1^{-1} + V_2^{-1})^{-1}V_2^{-1}\hat{\theta}_2 \tag{5.13}$$

and the covariance matrix of the pooled estimator is

$$V_T^{-1} = V_1^{-1} + V_2^{-1}. \tag{5.14}$$

This rule is a particular case of Rule 5 that will be proved in the Appendix, and generalizes Rule 1 to the vector case. For instance, the BLUE estimate of $\hat{\beta}$, when combining two independent samples of two regression equations with the same parameter β but different error variances, is given by

$$\hat{\beta} = (X'_1X_1/s_1^2 + X'_2X_2/s_2^2)^{-1}(X'_1Y_1/s_1^2 + X'_2Y_2/s_2^2).$$

6. RECURSIVE ESTIMATION

Suppose we have a parametric model $y_t = f(x_t, \theta, a_t)$ that relates a vector of responses to a set of explanatory variables x_t, a vector θ of p location parameters, and a set of unobserved random variables a_t. We will say that a *basic estimate* of the location parameter θ is an estimate obtained from a sample of size p, and an *elemental estimate* of θ is an estimate obtained from a sample of size one. We will say that an estimate is proper when it is obtained from a sample of at least size p. In the standard regression model where β is $p \times 1$ the elemental estimate of β given a sample (y_i, x_i) of size 1 is obtained from $x'_i\hat{\beta}_i = y_i$. Using the Moore–Penrose generalized inverse (Guttman 1982) and calling A^- the generalized inverse of A, the solution to this equation can be written as

$$\hat{\beta}_i = (x'_i)^- y_i = (x'_ix_i)^{-1}x_iy_i \tag{6.1}$$

where x_i is a $p \times 1$ column vector. This elemental estimate has a singular covariance matrix.

Sometimes we need to combine a proper and an elemental estimate of θ. For instance, in regression recursive estimation where we have an estimate $\hat{\beta}_{(n)}$ of (5.8) based on n data points, we observe y_{n+1} and need to revise $\hat{\beta}_{(n)}$ to obtain $\hat{\beta}_{(n+1)}$. In general, given a $p \times 1$ vector of parameters θ we will say that $\hat{\theta}_i$ is an elemental unbiased estimator of θ if: 1. the covariance matrix of $\hat{\theta}_i, V_i$, is such that rank $(V_i) = 1{:}2$. given p independent estimates $\hat{\theta}_i$ with covariance matrices V_i, the matrix $V_1^- + \cdots + V_p^-$, where V_i^- is a generalized inverse of V, is nonsingular; and 3. combining these p estimates by

$$\hat{\theta}_T = \sum_{i=1}^{p} \left(\sum_{1}^{p} V_i^- \right)^{-1} V_i^-\hat{\theta}_i \tag{6.2}$$

we obtain a basic unbiased estimator of $\hat{\theta}$. For instance, in linear regression the estimate (6.1) is elemental unbiased because: 1. The $p \times p$ covariance matrix of the estimate $\hat{\beta}_i, V_i$ is $V_i = x_ix'_i(x'_ix_i)^{-2}\sigma^2$, and it has rank equal to $1{:}2$. $V_i^- = x_ix'_i1/\sigma^2$ and

$$\left| \sum_{i=1}^{p} V_i^- \right| = \left(\frac{1}{\sigma^2} \right) \left| \sum_{i=1}^{p} x_ix'_i \right| = \left(\frac{1}{\sigma^2} \right) |X'X| \neq 0;$$

and 3. combining the elemental unbiased estimates $\hat{\beta}_i$ by

$$\hat{\beta}_T = \sum_{i=1}^{p} (\sum x_ix'_i)^{-1}x_iy_i = (X'X)^{-1}X'Y \tag{6.3}$$

we obtain the basic BLUE estimate. We can generalize (6.1) as follows.

Rule 5. Given n independent estimates $\hat{\theta}_i$ unbiased or elemental unbiased with covariance matrices V_i, that may be singular, the best (minimizing the trace of the covariance matrix) unbiased estimate is given by

$$\hat{\theta}_T = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} V_j^- \right)^{-1} V_i^-\hat{\theta}_i \tag{6.4}$$

where V_i^- is the Moore–Penrose generalized inverse of V_i, and where we have assumed that $\sum V_j^-$ is nonsingular. The

covariance matrix of $\hat{\theta}_T$ is then easily seen to be

$$V_T^{-1} = \sum_{i=1}^{n} V_i^{-}. \tag{6.5}$$

This rule is proved in the Appendix.

The application of this Rule 5 to recursive estimation leads directly to the Kalman filter (see, for instance, West and Harrison (1989) for a presentation of this estimation procedure in dynamic linear models). To show this, let us consider the standard state space formulation of a dynamic linear model with observation equation

$$y_t = A_t\theta_t + \varepsilon_t \tag{6.6}$$

and state equation

$$\theta_t = \Omega_t\theta_{t-1} + u_t \tag{6.7}$$

where y_t is $r \times 1$, A_t is $r \times p$ with rank $(A_t) = r, \varepsilon_t$ is $N_r(0, C_t), \Omega_t$ is $p \times p$, and $u_t \sim N_p(0, R_t)$. In this model, at any time t we may consider two independent estimates of θ. The first is the forecast of the state that comes from (6.7)

$$\hat{\theta}_t^{(1)} = \Omega_t\hat{\theta}_{t-1} \tag{6.8}$$

and whose covariance matrix can be obtained from

$$\theta_t - \hat{\theta}_t^{(1)} = \Omega_t(\theta_{t-1} - \hat{\theta}_{t-1}) + u_t. \tag{6.9}$$

Calling $I_t = \{y_t, \ldots, y_1\}$ the information until time t, and defining

$$V_{t/t-1} = E[(\hat{\theta}_t - \theta_t)(\hat{\theta}_t - \theta_t)'|I_{t-1}] \tag{6.10}$$

and letting $V_t = V_{t/t}$, we have from (6.9) that the covariance matrix of (6.8) is given by

$$V_{t/t-1} = \Omega_t V_{t-1}\Omega_t' + R_t. \tag{6.11}$$

The second estimate of θ at time t is obtained from (6.6) when y_t is observed. Assuming $p > r$ and A_tA_t' nonsingular, this estimate is given by

$$\hat{\theta}_t^{(2)} = A_t'(A_tA_t')^{-1}y_t \tag{6.12}$$

and using (6.6) it can be written

$$\hat{\theta}_t^{(2)} = A_t'(A_tA_t')A_t\theta_t + A_t'(A_tA_t')^{-1}\varepsilon_t \tag{6.13}$$

which shows that it is not unbiased for θ_t. However, it is easy to see that it is elemental unbiased, with singular covariance matrix

$$V_t^{(2)} = A_t'(A_tA_t')^{-1}C_t(A_tA_t')^{-1}A_t. \tag{6.14}$$

This matrix has a generalized inverse

$$(V_t^{(2)})^{-} = A_t'C_t^{-1}A_t.$$

Therefore, following Rule 5 the BLUE estimate will have a pooled covariance matrix

$$V_t^{-1} = V_{t/t-1}^{-1} + A_t'C_t^{-1}A_t \tag{6.15}$$

and the estimate will be given by

$$\hat{\theta}_T = (I - V_tA_t'C_t^{-1}A_t)\Omega_t\hat{\theta}_{t-1} + V_tA_t'C_t'y_t \tag{6.16}$$

or, as it is normally written,

$$\hat{\theta}_T = \Omega_t\hat{\theta}_{t-1} + V_tA_t'C_t^{-1}(y_t - A_t\Omega_t\hat{\theta}_{t-1}). \tag{6.17}$$

Equations (6.15) and (6.17) constitute the Kalman filter, which appears as a particular case of Rule 5.

It is interesting to stress that Equation (6.7) provides a clear ground for building influence measures of the last observed data in recursive estimation. Calling $\hat{\theta}_{t/t-1} = \Omega_t\hat{\theta}_{t-1}$ the forecast of θ_t with information until y_{t-1}, the change to the parameter vector due to observing y_t is given by

$$\hat{\theta}_t - \hat{\theta}_{t/t-1} = V_tA_t'C_t^{-1}e_{t/t-1} \tag{6.18}$$

where $e_{t/t-1} = y_t - A_t\hat{\theta}_{t/t-1}$ is the predicted residual. The Mahalanobis change to $\hat{\theta}_t$ will be given by

$$D_t = (\hat{\theta}_t - \hat{\theta}_{t/t-1})'V_t^{-1}(\hat{\theta}_t - \hat{\theta}_{t/t-1}) \tag{6.19}$$

that can be written as

$$D_t = e_{t/t-1}'C_t^{-1}A_tV_tA_t'C_t^{-1}e_{t/t-1}. \tag{6.20}$$

This diagnostic measure can be built for any statistical model in the state space form (6.6), (6.7) and estimated with the Kalman filter. It is straightforward to show that for linear regression models this statistic for the last observed point is equivalent to the one introduced by Cook (1977), whereas in ARIMA and transfer function models it is equivalent to the statistic introduced by Peña (1990, 1991).

7. CONCLUDING REMARKS

Any estimation or forecasting procedure can be seen as a way to combine the available information. In Bayesian statistics the prior information is combined with the posterior using Bayes's Theorem. In classical statistics the different pieces of sample information are weighted to obtain the final estimate. When we have unbiased estimators (or elemental unbiased) they are linearly combined to obtain the best linear unbiased estimate using as weights the (generalized) inverse covariance matrices. We have shown that analyzing estimates from this point of view can provide some useful insights on the properties of some statistical procedures.

APPENDIX: PROOF OF RULE 5

To prove Rule 5 let us consider the class of unbiased estimators

$$\hat{\theta}_T = \sum_{i=1}^{n-1} A_i\hat{\theta}_i + \left(I - \sum_{i=1}^{n-1} A_i\right)\hat{\theta}_n \tag{A.1}$$

such that if the $\hat{\theta}_i$ $(i = 1, \ldots, n)$ are unbiased, $\hat{\theta}_T$ will also be unbiased. The covariance matrix of $\hat{\theta}_T$ is

$$S_T = \sum_{i=1}^{n-1} A_i V_i A_i' + V_n - \sum_{i=1}^{n-1} A_i V_n - \sum_{i=1}^{n} V_n A_i' + \sum\sum A_i V_n A_j'$$

and the trace of this matrix is

$$m = \operatorname{tr}(S_T) = \sum_{i=1}^{n} \operatorname{tr}(A_i V_i A_i') + \operatorname{tr}(V_n) - 2\sum_{i=1}^{n} \operatorname{tr}(A_i V_n) + \sum_{i=1}^{n-1}\sum_{i=1}^{n-1} \operatorname{tr}(A_i V_n A_j').$$

Now, if V is symmetric we have that $\partial \operatorname{tr}(AVA')/\partial A = 2AV, \partial \operatorname{tr}(AV)/\partial A = V$, and $\partial \operatorname{tr}(AVB)/\partial A = B'V$ so that

$$\frac{\partial m}{\partial A_i} = 2A_i V_i - 2V_n + 2\sum_{j=1}^{n-1} A_j V_n = 0$$

and so

$$A_i V_i = \left(I - \sum_{j=1}^{n-1} A_j\right) V_n$$

$$A_i = \left(I - \sum_{j=1}^{n-1} A_j\right) V_n V_i^{-1}. \qquad \text{(A.2)}$$

Adding the $n-1$ equations (A.2) we obtain

$$\sum_{i=1}^{n-1} A_i = V_n \sum_{i=1}^{n-1} V_i^{-1} - \sum_{j=1}^{n-1} A_j V_n \sum_{i=1}^{n-1} V_i^{-1}$$

$$\sum_{i=1}^{n-1} A_i \left(I + V_n \sum_{i=1}^{n-1} V_i^{-1}\right) = V_n \sum_{i=1}^{n-1} V_i^{-1}$$

$$\sum_{i=1}^{n-1} A_i V_n \left(\sum_{i=1}^{n} V_i^{-1}\right) = V_n \sum_{i=1}^{n-1} V_i^{-1}$$

$$\sum_{i=1}^{n-1} A_i V_n = V_n \left(\sum_{i=1}^{n-1} V_i^{-1}\right)\left(\sum_{i=1}^{n} V_i^{-1}\right)^{-1}$$

and inserting this result in (A.2)

$$A_i = V_n \left(I - \left(\sum_{i=1}^{n-1} V_i^{-1}\right)\left(\sum_{i=1}^{n} V_i^{-1}\right)^{-1}\right) V_i^{-1}$$

$$A_i = V_n \left(\sum_{i=1}^{n} V_i^{-1} - \sum_{i=1}^{n-1} V_i^{-1}\right)\left(\sum_{i=1}^{n} V_i^{-1}\right)^{-1} V_i^{-1}$$

$$A_i = \left(\sum_{i=1}^{n} V_i^{-1}\right)^{-1} V_i^{-1}.$$

We have assumed in the proof that all the inverse matrices involved exist; the proof is similar when some of these matrices are singular by replacing the inverse by the generalized inverse of the matrix.

[Received April 1995. Revised August 1996.]

REFERENCES

Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis, Forecasting and Control*, San Francisco: Holden-Day.

Cook, D. R. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

Dear, K. B. G., and Begg, C. B. (1992), "An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis," *Statistical Science*, 7(2), 237–245.

DeJong, D. N. et al. (1992), "Integration versus Trend Stationarity in Time Series," *Econometrica*, 60(2), 423–433.

Draper, D. et al. (1992), *Combining Information. Statistical Issues and Opportunities for Research*, Washington, DC: National Academy Press.

Fuller, W. A. (1976), *Introduction to Statistical Time Series*, New York: Wiley.

Glass, G. V. (1976), "Primary, Secondary, and Meta-Analysis of Research," *Educational Researcher*, 5, 3–8.

Grenander, V., and Rosenblatt, M. (1957), *Statistical Analysis of Stationary Time Series*, New York: Wiley.

Guttman, I. (1982), *Linear Models*, New York: Wiley.

Hedges, L. V. (1992), "Modeling Publication Selection Effects in Meta-Analysis," *Statistical Science*, 7(2), 246–255.

Hedges, L. V., and Olkin, I. (1985), *Statistical Methods for Meta-Analysis*, New York: Academic Press.

Mosteller, F., and Chalmers, T. C. (1992), "Some Progress and Problems in Meta-Analysis of Clinical Trials," *Statistical Science*, 7(2), 227–236.

Newbold, P., and Granger, C. W. J. (1974), "Experience with Forecasting Univariate Time Series and the Combination of Forecasts," *Journal of Royal Statistical Society*, Ser. A, 131–146.

Peña, D. (1990), "Influential Observations in Time Series," *Journal of Business and Economic Statistics*, 8(2), 235–241.

——— (1991), "Measuring Influence in Dynamic Regression Models," *Technometrics*, 33(1), 93–104.

——— (1995), "Forecasting Growth with Time Series Models," *Journal of Forecasting*, 97–105.

Peña, D., and Maravall, A. (1991), "Missing Observations, Additive Outliers and Inverse Autocorrelation Function," *Communications in Statistics (Theory and Methods)*, A20(10), 3175–3186.

Peña, D., and Yohai, V. (1995), "The Detection of Influential Subsets in Linear Regression Using an Influence Matrix," *Journal of Royal Statistical Society*, Ser. B, 57(1), 145–156.

Shaman, P. (1969), "On the Inverse of the Covariance Matrix of a First Order Moving Average," *Biometrika*, 56, 595–600.

Utts, J. (1991), "Replication and Meta-Analysis in Parapsychology," *Statistical Science*, 6(4), 363–403.

West, M., and Harrison, J. (1989), *Bayesian Forecasting and Dynamic Models*, Berlin: Springer-Verlag.

Wolf, F. M. (1986), *Meta-Analysis, Quantitative Methods for Research Synthesis*, Beverly Hills, CA: Sage Publications.