1982

Cautionry Tales in Estimating Variance Components (Or: Throwing the Variance Out with the Bath Water)

Glenn M. MacDonald

Chris Robinson

20075

ISSN: 0318-725X

ISBN: 0-7714-0323-2

RESEARCH REPORT 8205

CAUTIONARY TALES IN ESTIMATING VARIANCE COMPONENTS (OR: THROWING THE VARIANCE OUT WITH THE BATH WATER)

by

Glenn M. MacDonald

and

Chris M. G. F. Robinson

February, 1982

Variance components models are now commonplace in the econometric analysis of pooled time series and cross section data. As usual, the dependent variable is hypothesized to depend on exogenous variables and an error term. But in the variance components model, the error term is assumed to be comprised of two"effects". One is simply a random error, taken to be independent across both time and units of observation (the latter referred to as "individuals"). The other is an *individual effect*, independent across individuals but constant over time. The variances associated with these two effects are termed the *variance components*, and are parameters to be estimated along with the slope coefficients. The variance of the individual effect is usually called the *individual component*.

The variance components model has proved especially useful for the econometric analysis of wage and income data (see Hause, 1977; Lillard and Weiss, 1979; Smith, 1979). The estimated variance components play a particularly important role in studies of income distribution and poverty (for example, see Lillard and Willis, 1978).

Estimation of individual components by classical variance components methods is particularly sensitive to "outliers" if as is typically the case, the length of time over which individuals are observed (panel length) is small relative to the number of individuals in the sample (panel size). The large panel size is typically sufficient to estimate the slope coefficients accurately, but the small panel length creates difficulties in distinguishing between the individual effects and the random disturbances in the relationship.

The problem of outliers is often dealt with by arbitrary elimination of "obviously impossible" values. Indeed, some *truncation* will often have taken place before the econometrician even receives the data. This occurs through the deletion of "out of range" values by the agency providing the data. Additional truncation may follow if particularly extreme values are generated for a variable which is not directly available and which must therefore be constructed. For example, many outliers are created in wage data when they are obtained by dividing annual earnings by annual hours of work, where both are subject to error. Estimates of the population (conditional) mean wage rate (regression coefficient vector) will often be insensitive to elimination of such outliers provided this takes the form of cutting off both tails of the distribution. Thus applied researchers may devote little attention to potential problems resulting from truncation. However, eliminating the tails of the distribution will in general have a marked effect on its variance and hence the estimated variance of the individual effects is potentially very sensitive to such truncation. The implications for analyses of income distribution, which often focus on behavior of the tail of the distribution, are potentially very serious.

Truncation may be viewed as imposing particularly extreme prior beliefs on the distribution of individual effects. For example, an investigator may truncate the value of a wage observation because of his belief that the individual effect cannot possibly be so extreme, and hence that the extreme value is due to measurement error. Lazear (1976, p. 551) who is particularly explicit about his "pre-analysis" incorporation of prior information, provides a good example: "The original sample has records of 5,225 individuals. This had to be reduced to 1,996 observations to meet the following criteria. First, it was necessary that individuals in the sample have wage rates reported in both 1966 and 1969. Second, individuals who reported that their wage rate was either less than 50¢ per hour or greater than $10 were dropped on the grounds that reported wages in these cases were unlikely to be correct. Finally, observations were dropped for which there was incomplete information on the variables used in the analysis." Similarly, Olson, White and Shefrin (1979) explicitly exclude individuals from their sample if they have "obviously miscoded values". Some form of "pre analysis" selection is very common in this area.

Bayesian analysis provides a general framework for incorporating prior information in a statistical analysis. This analysis usually takes the form of the derivation of a "Bayesian" estimator whose properties are then compared with those of a "classical" estimator. In the following sections we follow this pattern in deriving a Bayesian estimator that incorporates flexibly the kind of prior information imposed by truncation in the above examples and comparing its properties with a classical estimator. However, in addition to providing estimators that may improve on classical estimators by some criteria, Bayesian analysis may play the equally important role of clarifying the effects of prior beliefs often imposed implicitly--of necessity in a rigid form--in "classical" analysis. In the present paper we emphasize this role, using the Bayesian framework to provide a method for assessing the "reasonableness" of the prior beliefs implied by any proposed pre-analysis truncation or exclusion of observations. It is shown that in many cases, especially in large samples, truncation implies extremely dogmatic prior beliefs whose imposition substantially affects the resulting estimates.

An almost universal practice in the analysis of large individual data sets is some form of sample censoring. Most investigators follow Lazear in requiring complete observations on all the relevant variables. Keifer (1979), for example, using 6 quarters of data from a longitudinal survey from the Office of Economic Opportunity, restricts his sample to "those who reported wage rates in each of these 6 quarters". Only one-third of the available sample meets this restriction, but as Kiefer notes this is a problem common to most longitudinal data sets used in labour economics, e.g., Michigan Panel Study of Income Dynamics. In general, investigators applying variance components models eliminate an individual with data missing for any single period of the multi-period panel. The effects of such <u>censoring</u>

on the properties of OLS regression coefficients have received a great deal of attention in the recent literature dealing with sample selection bias or attrition bias in panels (see for example, Heckman (1979), Hausman and Wise (1979)). However, the effect of truncation and censoring on the estimation of individual components, or more generally on the validity of conventional inference, has been relatively neglected. An example given below dramatically illustrates the sensitivity of estimated variance components to sample censoring relative to that of regression coefficient estimates. Even when the form of sample censoring does not result in inconsistency of the regression ceofficients, the estimated variance components, and more generally the variance-covariance matrix on which inferences are based, may be inconsistent.

In Section I, a Bayesian approach to incorporating and assessing prior information is presented. Bayesian estimators for the variance components model are proposed. They can incorporate a wide variety of prior beliefs and are very simple to compute. An illustrative computation is presented in Section II. It is based on some typical situations faced by an empirical investigator in possession of prior information such as a survey article on previous estimates of the parameters of interest. These present some evidence that the proposed estimators are well behaved.

In Sections III and IV, an empirical example is provided. The data are from the National Longitudinal Surveys. The classical variance components estimates, under several truncation criteria, are presented along with the Bayesian estimates. The Bayesian estimators are also used to infer the prior information corresponding to several truncation criteria. The implied prior beliefs equivalent to elimination of even a small number of "outliers" are shown to be dogmatic in the extreme. Then, in Section IV, variance component estimation is examined for cases wherein the data are censored, or when some missing data are constructed. Section V contains a summary and suggestions for further work.

I. PRIOR INFORMATION IN VARIANCE COMPONENTS MODELS

Variance components models permit an investigator to distinguish between alternative sources of stochastic disturbance in a relationship. For example, in a wage equation the wage observation on a given individual may include both measurement error and some omitted variables specific to the individual. The investigator may be interested in measuring the individual effects--i.e., isolating them from the neasurement error. Moreover, there may be some prior information on the relative magnitudes of these two forms of disturbance. Thus, knowledge of the labour markets generating the wage data may yield some prior information on the variability in returns that may be paid for various kinds of individual characteristics not included in the data set at hand. Typically there is little prior knowledge on the source of the measurement error. Investigators, therefore, when faced with a particularly large disturbance (i.e., an unusual wage rate) have tended to attribute this all to measurement error and have eliminated the observation from the data set so as to prevent it having any influence on the estimated individual effects. Incorporation of prior information in this way, however, is very crude. A more flexible method of incorporating the notion that most of an extreme disturbance is more likely to be due to measurement error than to variability in the individual effect is the use of an appropriate Bayesian prior distribution on the parameters governing measurement error and individual effects. An appropriate prior distribution is proposed below.

Suppose that observations on individual i in period t, y_{it}, are generated as follows:

$$(1) \qquad y_{it} = \theta + \mu_i + \epsilon_{it} \qquad \begin{array}{l} i=1,\ldots,N \\ t=1,\ldots,T \end{array}$$

This is the classic variance components version of the location model where θ is the location parameter, $-\infty < \theta < \infty$; μ_i are the individual effects, assumed to be independent $N(0,\sigma_\mu^2)$ random variables; and ϵ_{it} are independent $N(0,\sigma_\epsilon^2)$ random variables representing measurement error. The location model has no independent variables, contrary to the standard regression models used by economists. It is adopted here since estimating regression coefficients is not of primary interest in the present context. The location model permits us to focus on the structure of the disturbance in the model. The parameter of primary interest is the individual component σ_μ^2--the variance of the individual effects. That is, how much, given "location" (say, schooling, experience, etc.) can one individual's observation (say, wage) differ from another's apart from measurement error? A classical estimator for σ_μ^2 may be obtained by choosing the value of σ_μ^2 (as well as σ_ϵ^2 and θ) that is "most likely" to have generated the data. This involves maximizing the likelihood function, which in this case is given by (see Box and Tiao, 1973, pp. 250-1):

$$(2) \quad \ell(\theta,\sigma_\epsilon^2,\sigma_\mu^2|y) \propto (\sigma_\epsilon^2)^{-\nu_1/2} (\sigma_\epsilon^2 + T\sigma_\mu^2)^{-(\nu_2+1)/2} \cdot$$

$$\exp\left\{-\frac{1}{2}\left[\frac{NT(y_{..}-\theta)^2}{\sigma_\epsilon^2 + T\sigma_\mu^2} + \frac{S_2}{\sigma_\epsilon^2 + T\sigma_\mu^2} + \frac{S_1}{\sigma_\epsilon^2}\right]\right\} ,$$

where

$$y_{..} = \frac{1}{NT} \sum_i \sum_t y_{it},$$

$$y_{i.} = \frac{1}{T} \sum_t y_{it},$$

$$\nu_1 = N(T-1),$$

$$\nu_2 = N - 1,$$

$$S_1 = \sum_i \sum_t (y_{it}-y_{i.})^2,$$

$$S_2 = T \sum_i (y_{i.}-y_{..})^2$$

and $\propto$ indicates "proportional to".

Of course, if some of the data are eliminated by truncation and (2) is maximized over the remaining data, the result is no longer a maximum likelihood estimate.

The Bayesian approach combines the information in the sample with prior information via Bayes's Theorem:

$$(3) \quad p(\theta,\sigma_\epsilon^2,\sigma_\mu^2|y) \propto p(y|\theta,\sigma_\epsilon^2,\sigma_\mu^2)p(\theta,\sigma_\epsilon^2,\sigma_\mu^2)$$

$p(y|\theta,\sigma_\epsilon^2,\sigma_\mu^2)$ is the probability density function (p.d.f.) for the observations, and is algebraically equivalent to the likelihood function. The maximum likelihood estimator follows from choosing θ, σ_ϵ^2 and σ_μ^2 that makes this density largest. $p(\theta,\sigma_\epsilon^2,\sigma_\mu^2|y)$ is called the posterior density for $(\theta,\sigma_\epsilon^2,\sigma_\mu^2)$. It characterizes an investigator's posterior beliefs regarding the parameters given the data, y, and his prior beliefs regarding $(\theta,\sigma_\epsilon^2,\sigma_\mu^2)$ as represented by the prior distribution $p(\theta,\sigma_\epsilon^2,\sigma_\mu^2)$. The likelihood function was presented above. It remains to examine the prior p.d.f.

The parameters $(\theta,\sigma_\epsilon^2,\sigma_\mu^2)$ are fixed, but unknown numbers. However, before the current data is observed the investigator may believe that some values are "more likely" to be the true unknown values than others (perhaps from knowledge of other data sets, or more generally from well corroborated theories relevant to the current problem). One way of viewing this is to consider that there is a "super-population" from which the parameters are drawn. The parameters are then simply realizations of the super-population. Within the Bayesian framework an investigators prior beliefs may be associated with this super-population. This notion of a super-population is used below in comparing alternative estimators.

The location parameter θ is not of primary interest. Thus, we assume that the investigator has "no information" on θ, i.e., cannot consider any one value more or less likely than any other. This notion of no information is captured by a generalization of the uniform distribution by the so-called non-informative prior.[1]

(4) $$p(\theta) \propto C \qquad - \infty < \theta < \infty$$

Since the remaining parameters, σ^2_ϵ and σ^2_μ are both variances, their prior p.d.f's should ensure positive values. A simple p.d.f. that meets this criterion is the inverted χ^2 or χ^{-2} distribution.[2] If Z is a $\chi^{-2}_{(r)}$ random variable, r being the degrees of freedom, the p.d.f. of Z is:

(5) $$p(z) \propto z^{-(r/2+1)} \exp(- \frac{1}{2}z) \qquad z > 0$$

However, since this is a one parameter p.d.f. it cannot be used directly to represent prior beliefs regarding σ^2_ϵ or σ^2_μ since independent statements concerning central tendency and dispersion are ruled out. This is remedied by the transformation $X = aZ$, $a > 0$. Then the p.d.f. for X is:

(6) $$p(x) \propto x^{-(r/2+1)} \exp(-a/2x) \qquad x > 0$$

Given the properties of the inverted χ^2 distribution it is easy to show that the mean and variance of X take the simple forms:

$$E(X) = \frac{a}{r-2} \tag{7}$$

$$V(X) = \frac{2a^2}{(r-2)^2(r-4)} \tag{8}$$

Suppose an investigator's examination of previous research typically finds estimates of σ_μ^2 around 1, but also finds considerable variability. This, coupled with any other prior information could be represented by setting $E(\sigma_\mu^2) = 1$ and $V(\sigma_\mu^2) = 2$, and hence, from (7)-(8), $a_\mu = 3$ and $r_\mu = 5$.[3] This yields the prior p.d.f. $p(\sigma_\mu^2)$ shown in Figure 1 below. The same form of p.d.f. may be chosen for σ_ϵ^2 with appropriate values of a_ϵ and r_ϵ. If there is little prior information about measurement error, a_ϵ and r_ϵ may be chosen so as to yield a flat or uninformative prior distribution.

Combining the prior p.d.f.'s with the likelihood function yields the posterior p.d.f. for the parameters θ, σ_μ^2 and σ_ϵ^2. Since θ is not of interest we remove it from the posterior p.d.f. by integrating it out and obtain:[4]

$$p(\sigma_\epsilon^2,\sigma_\mu^2|y) \propto (\sigma_\epsilon^2)^{-\eta_1}(\sigma_\mu^2)^{-\eta_2}(\sigma_\epsilon^2 + T\sigma_\mu^2)^{-\eta_3} \exp\left\{-\frac{1}{2}\left[\frac{S_2}{\sigma_\epsilon^2 + T\sigma_\mu^2} + \frac{S_1+a_\epsilon}{\sigma_\epsilon^2} + \frac{a_\mu}{\sigma_\mu^2}\right]\right\} ;\ \sigma_\epsilon^2 > 0,\ \sigma_\mu^2 > 0. \tag{9}$$

In (9),

$$\eta_1 = \frac{r_\epsilon + \nu_1}{2} + 1,$$

$$\eta_2 = \frac{r_\mu}{2} + 1,$$

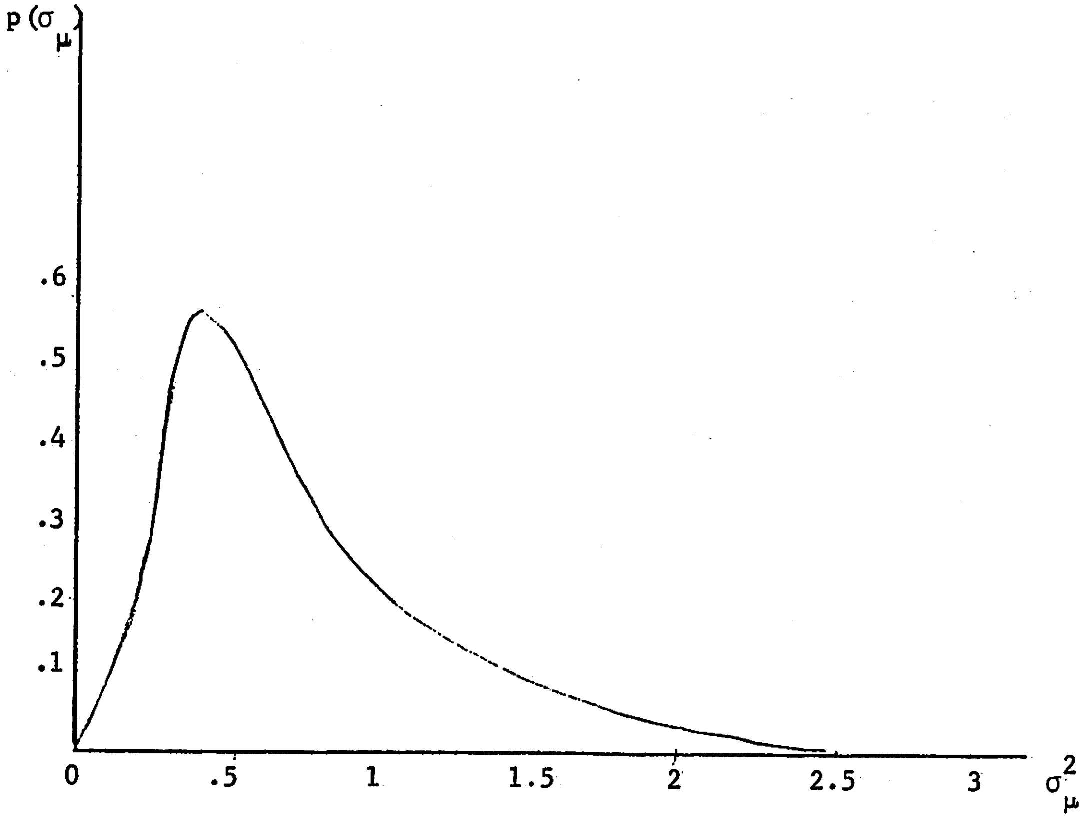
and

$$\eta_3 = \frac{\nu_2}{2} .$$

Figure 1

Prior Distribution for σ_{μ}^{2}

(Informative: $a_{\mu} = 3$; $r_{\mu} = 5$)



This posterior p.d.f. combines both sample and prior information. It represents a compromise between the sample information as represented by the likelihood and the prior information as represented by the prior p.d.f. As the sample size increases, this posterior p.d.f. will be dominated by the likelihood. Almost irrespective of the beliefs the investigator has to begin with, a large enough sample will result in them being changed to conform to the data. Conversely, however, a few "unusual" or "unrepresentative" data points (small sample) will not cause the investigator to modify greatly any informative prior beliefs.

The natural candidates for point estimators of σ_ϵ^2 and σ_μ^2 are expected values or modes of σ_ϵ^2 and σ_μ^2 from (9), i.e., the posterior means or modes. Computation of the means require a bivariate numerical integration. Since this is a strong deterrent to most applied researchers we prefer the posterior modal values of σ_ϵ^2 and σ_μ^2. These are analogous to maximum likelihood except that the values chosen are those that are "most likely" with respect to both the sample points and the prior information. They are obtained by solving the first order conditions for a maximum of (9):

$$\text{(10)}\quad \frac{-\eta_1}{\sigma_\epsilon^2} - \frac{\eta_3}{\sigma_\epsilon^2 + T\sigma_\mu^2} + \frac{S_2}{2(\sigma_\epsilon^2 + T\sigma_\mu^2)^2} + \frac{S_1 + a_\epsilon}{2(\sigma_\epsilon^2)^2} = 0$$

and

$$\text{(11)}\quad \frac{-\eta_2}{\sigma_\mu^2} - \frac{\eta_3 T}{\sigma_\epsilon^2 + T\sigma_\mu^2} + \frac{TS_2}{2(\sigma_\epsilon^2 + T\sigma_\mu^2)^2} + \frac{a_\mu}{2(\sigma_\mu^2)^2} = 0 ,$$

where care must be taken to ensure that the solutions in fact represent a maximum of the posterior. Though (10) and (11) are nonlinear, they are easily solved numerically. Indeed any computer software that will perform nonlinear least squares (e.g., TSP) will solve (10) and (11) easily.[5] Consistency of the resulting estimators, denoted $\hat{\sigma}_\mu^2$ and $\hat{\sigma}_\epsilon^2$, is shown in the appendix.

II. ILLUSTRATIVE COMPUTATION

The estimator $\hat{\sigma}_\mu^2$ is used in a specific example in Section III in a comparison of alternative classical estimators modified by the incorporation of prior information by truncation of observations. Before making these comparisons, however, this section presents illustrative results for $\hat{\sigma}_\mu^2$ for some typical cases of prior information formation. The results suggest that $\hat{\sigma}_\mu^2$ behaves well and can be used as a benchmark estimator that incorporates prior information in a sensible manner.

In order to evaluate the estimators a criterion is required. A commonly used criterion is the unconditional mean squared error (m.s.e.). The estimator $\hat{\sigma}_\mu^2$ is a function of the data as summarized by S_1 and S_2 (see (10)-(11)). Thus $\hat{\sigma}_\mu^2$ takes on different values with probabilities dependent on the sampling distributions of S_1 and S_2. These sampling distributions have as parameters the unknown values of σ_μ^2 and σ_ϵ^2. Conditional on values for σ_μ^2 and σ_ϵ^2, the expected values of say, $\hat{\sigma}_\mu^2$ may be computed. Similarly we may compute the conditional m.s.e. for $\hat{\sigma}_\mu^2$ which is the expected squared deviation of $\hat{\sigma}_\mu^2$ from σ_μ^2 conditional on the chosen values of σ_μ^2 and σ_ϵ^2. This would give us an indication of what our typical (squared) errors would be using $\hat{\sigma}_\mu^2$ as our estimator, provided σ_μ^2 was in fact the true parameter value. However, we do not know the true value of σ_μ^2. One estimator may have a smaller m.s.e. than another for one value of the unknown parameter σ_μ^2; but for a different value of σ_μ^2 the ranking may be reversed. In order to avoid this problem we consider the m.s.e. of all possible values of the unknown parameter σ_μ^2. Using the prior p.d.f. on σ_μ^2 to weight the conditional m.s.e.'s at each possible value of σ_μ^2 and summing yields the unconditional m.s.e. This provides a measure of the typical (squared) error we would make using our estimator, whatever the true value of the parameter happened to be, but taking into account where we think the true value is most likely to be.

The estimators used for comparison with $\hat{\sigma}^2_\mu$ and $\hat{\sigma}^2_\varepsilon$ are the unbiased estimators (see Klotz, Milton and Zacks (1969)):

$$\tau^2_\mu = \frac{1}{T}\left[\frac{S_2}{\nu_2} - \frac{S_1}{\nu_1}\right]$$

$$\tau^2_\varepsilon = \frac{S_1}{\nu_1}$$

Table 1 reports the mean squared errors for the alternative estimators, based on the following parameter values: $N = 50$, $T = 5$, $\theta = 0$, $a_\varepsilon = 12$, $r_\varepsilon = 8$, $a_\mu = 24$, $r_\mu = 8$. This results in priors for σ^2_ε and σ^2_μ with means $E(\sigma^2_\varepsilon) = 2$, $E(\sigma^2_\mu) = 4$ and variances $V(\sigma^2_\varepsilon) = 2$ and $V(\sigma^2_\mu) = 8$. Row (1) reports the results for the unbiased estimators using the above parameters. The remaining rows report the results for the proposed Bayesian estimators, based on three alternative types of prior information.[6] The first set of prior information (Prior 1) gives the investigator prior beliefs that correspond exactly to the true "super populations" from which the realized σ^2_ε and σ^2_μ relevant to any given sample were drawn. The results for this case are given in row (2a). The relative efficiency gains for the Bayesian estimators are modest (columns (3), (6))--approximately 1.7% for $\hat{\sigma}^2_\varepsilon$ and 4.1% for $\hat{\sigma}^2_\mu$.

Prior 1, however, gives the Bayesian investigator the important advantage of priors that are coincident with the super population. One way to relax this is to suppose the investigator to believe that σ^2_ε and σ^2_μ came from a population of the form (6), but to be uncertain about the precise values of a_ε, r_ε, a_μ and r_μ. This hierarchical approach, though

Table 1

Estimator	(1) $E(\phi_\epsilon)$	(2) $mse(\phi_\epsilon)$	(3) $\frac{mse(\tau_\epsilon^2)}{mse(\phi_\epsilon)}$	(4) $E(\phi_\mu)$	(5) $mse(\phi_\mu)$	(6) $\frac{mse(\tau_\mu^2)}{mse(\phi_\epsilon)}$
1. Unbiased	2.0	.060	1.000	4.0	1.122	1.000
2. Unscaled Modal						
a) Prior 1	1.967	.059	1.017	3.724	1.076	1.041
b) Prior 2	1.945	.059	1.017	3.517	1.432	.782
c) Prior 3	1.960	.061	.984	3.634	1.381	.811
3. Scaled Modal						
a) Prior 1	1.985	.054	1.111	3.994	1.042	1.075
b) Prior 2	1.989	.060	1.00	3.723	1.230	.910
c) Prior 3	1.982	.057	1.045	3.893	1.119	1.001

formally appropriate, rapidly becomes very complicated. It is somewhat outside the spirit of the exercise in that the whole point is to find a simple way to incorporate relevant prior information. Instead our approach was to generate some data sets from which the investigator can form his priors. We then ask how the estimators perform if the investigator acts as if his priors are the true super-populations, but they are in fact not. Five values each of σ_ϵ^2 and σ_μ^2 were drawn from the true super-population: $a_\epsilon = 12$, $r_\epsilon = 8$, $a_\mu = 24$, $r_\mu = 8$. Their values are listed in column 1 of Table 2. Five data sets on y_{it} were then randomly generated using these values of σ_ϵ^2 and σ_μ^2, each with $N = 25$ and $T = 5$. The investigator is assumed to have access to (at least reports of) these data sets, and forms his priors based on them. "Prior 2" was obtained by letting the investigator "eyeball" the unbiased estimates available for the data sets (columns 3 and 4, Table 2). In practice for this simulation this was accomplished by one author constructing the data sets and the co-author eyeballing the results without knowledge of the actual values of σ_μ^2 and σ_ϵ^2. This is to represent the situation where prior information is a recollection of other empirical studies similar to the one the investigator is working on. The figures thus obtained were $E(\sigma_\epsilon^2) = 1$, $V(\sigma_\epsilon^2) = .75 = V(\sigma_\mu^2)$, $E(\sigma_\mu^2) = 3$. Finally, "Prior 3" is obtained by a slightly more sophisticated approach: the average values of τ_ϵ^2 and τ_μ^2 were used as $E(\sigma_\epsilon^2)$ and $E(\sigma_\mu^2)$ and the sample variances of τ_ϵ^2 and τ_μ^2 were employed as $V(\sigma_\epsilon^2)$ and $V(\sigma_\mu^2)$. The situation envisioned here is that the investigator might have access to published studies on other data sets similar to his in which τ_ϵ^2 and τ_μ^2 were reported.

Table 2

	(1)	(2)	(3)	(4)
Data Set	σ_{ϵ}^2	σ_{μ}^2	τ_{ϵ}^2	τ_{μ}^2
1	.883	6.234	.424	5.230
2	.546	4.188	1.142	3.522
3	1.463	2.993	2.634	3.025
4	2.837	2.697	2.098	2.110
5	1.540	2.239	1.226	3.584

Returning to Table 1, the simulation results using "Prior 2" and "Prior 3" are reported in rows (2b) and (2c). In this case the Bayesian estimator's performance is worse than that of the unbiased estimators. Since modal estimators often perform badly if the posterior is skewed, particularly under convex loss functions (i.e., criteria like m.s.e.), we also computed Bayesian estimators scaled towards the posterior mean.[7] The results for these estimators are presented in Table 1, columns (3a)-(3c) for the three forms of prior information. The scaled estimator outperformed the unbiased estimator, but the gains are small. Thus, in common with Bayesian estimators presented for other problems, the estimator presented here converges rapidly to the classical unbiased estimator as the sample size increases given relatively vague prior information. Conversely, however, even with relatively short panel lengths, dogmatic priors are required before the estimator differs from the unbiased estimator. At this point therefore we reverse the procedure and use the Bayesian estimator to clarify the nature of the prior information that is imposed by truncation or elimination of "outliers" in classical procedures. The importance of this kind of assessment is dramatically illustrated in the following section. There the results of incorporating prior information via deleting "outliers" are compared with the Bayesian approach. It is demonstrated that deleting "outliers" is an extremely crude way of incorporating prior information and implies unreasonably dogmatic prior beliefs.

III. APPLICATION TO INDIVIDUAL WAGE COMPONENTS

In this section we compute individual wage components using classical and Bayesian methods. We compare the incorporation of prior information via a prior distribution of σ_{μ}^2 with truncation of outliers. We estimate what has become a "standard" wage equation of the form:

(12) $\ell n\ W_{it} = X_{it}\beta + \mu_i + \varepsilon_{it}.$

W_{it} is individual i's wage rate in period t; X_{it} is a vector of (assumed) exogenous characteristics--schooling, experience and experience squared--and a time trend; β is a vector of parameters. The error structure is as follows. μ_i represents an individual effect on the wage rate, invariant across time, and not captured by the characteristics X_{it}. Typically, μ_i is taken to represent omitted ability or productivity characteristics of the individual. ϵ_{it} is a disturbance arising from temporary market phenomena, measurement error, etc. μ_i and ϵ_{it} are assumed to have the properties specified in Section I above.[8] Typical estimates of σ^2_μ obtained in studies using this kind of framework have clustered around 0.15.[9] As a result this has now become a "stylized fact" for this kind of data. In the remainder of this section we examine how the imposition of this stylized fact affects the estimates of σ^2_μ on a "new" set of data.

The data used were six observations on a panel of the husbands of women in the National Longitudinal Survey of Women 30-44, between 1967 and 1976. The number of husbands in the sample was 1187.[10] Thus while the panel length is relatively "small", the panel size is "large". Given the large panel size and the fact that the parameters β are not of primary interest, (12) was estimated by ordinary least squares. The consistent residuals, $\ln W_{it} - X_{it}\hat{\beta}$, were then treated as y_{it} above.

The estimated individual components, given by y_i are plotted in Chart 1 below. The wage measure obtained from these data, as is typically the case, are not available directly. Direct measures on annual earnings, E_i, annual weeks worked, L_i, and "usual" hours per week, H_i are used to compute the hourly wage rate as $W_i = E_i/L_iH_i$. Measurement error in W_i thus depends on the error specification for the components, E_i, L_i, H_i.

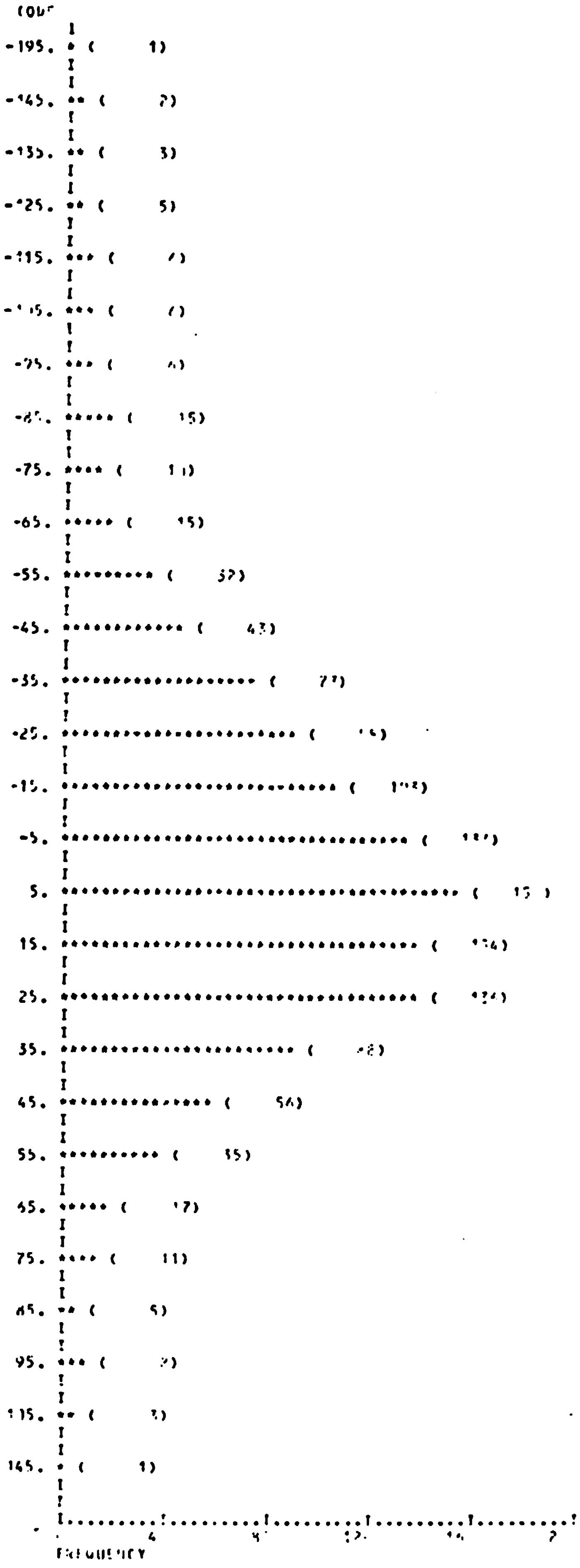


Chart 1

Distribution of Individual Components

The assumed normality of the measurement error in the log wage equation may be induced by assuming multiplicative log-normal disturbances for E_i, L_i and H_i. As usual with this procedure, despite prior removal of out of range values of the variables before the data are made available, some very implausible wage rates are obtained. For example, the residuals for some individuals indicate that given average characteristics, they would be receiving a wage rate less than 10¢; others have wage rates in excess of $40. Prior knowledge of the operation of labour markets for labour of given education and experience level, may suggest that the extreme wage rate observations are the result of an extreme measurement error (ϵ_{it}) rather than an extreme individual component (μ_i). We examine the sensitivity of estimates of individual wage variability, σ_μ^2, to alternative ways of dealing with this prior information. A wage rate cannot be computed without information on three variables: earnings, weeks and hours. Typically, however, there will be individuals in the sample for whom "complete data" are lacking. We postpone consideration of the sensitivity of estimates of σ_μ^2 to alternative ways of dealing with this problem to Section V.

Table 3 presents estimates of σ_μ^2 using the sample of individuals with complete data for all years under various prior specifications. This sample inclusion criterion is standard in the analysis of variance components in labour economics. For example, Lillard and Willis (1978) required that for sample inclusion an individual must have reported positive annual hours and earnings each year. This and other criteria reduced the sample size from approximately 3000 to approximately 1000. The benchmark case is to ignore all prior information and compute the classical estimate τ_μ^2.

Table 3

Estimation of σ_μ^2 Under Alternative Methods of Incorporating Prior Information

(1)	(2)	(3)	(4)	(5)	(6)
Truncation	Panel Length	Panel Size	τ_μ^2	$\hat{\sigma}_\mu^2$	$\tilde{\sigma}_\mu^2$
None	6	1187	0.129	0.129	0.129
None	3	1187	0.137	0.137	0.137
1%	6	1018	0.074	-	-
1%	3	1018	0.072	-	-
5%	6	883	0.050	-	-
5%	3	883	0.050	-	-
Extreme	6	1161	0.113	-	-
Outliers	3	1173	0.118	-	-

This is reported in column 4. In rows 1 and 2 the full sample of individuals with complete data is used for panel lengths of 6 and 3 years, respectively. The remaining rows present values of τ_{μ}^{2} under alternative rules for deleting "outliers". Thus, given the stylized fact of $\sigma_{\mu}^{2} = .15$, all observations with less than a 1% (5%) probability of occurring due to individual effect variability are deleted in rows 3 and 4 (5 and 6).[11] The effect on the estimate τ_{μ}^{2} is dramatic. The 1% truncation cuts the estimate of σ_{μ}^{2} almost in half. The 5% truncation results in an estimate only about one third of its original size. Finally even if only a handful of "very extreme" outliers are eliminated τ_{μ}^{2} is reduced by 12-15%. The "very extreme" outliers were determined by inspection; in the event they turned out to be much more extreme than Lazear's criterion since only observations outside of the 30¢ to $24 range were excluded. Indeed, Lazear's criterion is in practice close to our 1% criterion. These changes are in marked contrast to the effects on the estimated vector of regression coefficients presented in Table 4. Despite the large variation in sample inclusion criteria and sample size, the estimated coefficients in Table 4 exhibit a high degree of stability. This stability may have led investigators to devote little attention to other problems created by sample inclusion criteria.

In columns 5 and 6 of Table 3, Bayesian estimates of σ_{μ}^{2} are presented based on the following prior specification. The labour market we are dealing with is assumed to be stable and hence is assumed to result in approximately the same variance of individual components. The chosen prior means for σ_{μ}^{2} of 0.15 (column 5) and 0.10 (column 6) bracket the estimates quoted earlier. The prior variance was chosen to reflect a moderate degree of confidence based on the tendency of previous estimates to cluster. This is expressed by a coefficient of variation of 0.47 which translates into a prior variance of 0.005.

Table 4

Pooled OLS Estimates of the Regression Coefficients in the Log-Wage Equation for Alternative Samples

Complete Data Samples Truncation	Panel Length	CON	EDHCO	TCO	EXSO CO	EXCO	N	R^2
None	6	-.770863	.0674796	.0150637	-.00039155	.017977	7122	.2017
		(.145335)	(.002119)	(.002168)	(.000078)	(.004186)		
1%	6	-.637076	.066352	.012986	-.000446	.022129	6108	.3017
		(.111560)	(.001684)	(.001667)	(.000060)	(.003206)		
5%	6	-.609521	.0641515	.013748	-.000404	.0192895	5298	.3530
		(.101827)	(.001599)	(.001530)	(.000055)	(.002961)		
None	3	-1.794824	.070637	.028864	-.000522	.023393	3561	.2223
		(.330801)	(.002899)	(.004956)	(.000122)	(.005884)		
1%	3	-1.337310	.070501	.021892	-.000493	.025262	3054	.3148
		(.256021)	(.002322)	(.007838)	(.000095)	(.004538)		
5%	3	-1.154719	.067119	.020826	-.000421	.020884	2649	.3571
		(.234658)	(.002211)	(.003525)	(.000087)	(.004207)		
Extreme Outliers	6	-.716552	.069821	.013828	-.000367	.017878	6966	.2321
		(.134791)	(.001977)	(.002011)	(.000073)	(.003889)		
Extreme Outliers	3	-1.659728	.074982	.025782	-.000464	.023361	3519	.2560
		(.305978)	(.002696)	(.004587)	(.000112)	(.005432)		
Missing Data Samples								
Alternative								
(a)	6	-.200299	.072.99	.005357	-.000437	.018961	13302	.1384
		(.145654)	(.002104)	(.002187)	(.000074)	(.004053)		
(b)	6	-1.09968	.072720	.018442	-.000374	.017737	13302	.1867
		(.120509)	(.001741)	(.001810)	(.000061)	(.003353)		
(a)	3	-1.149956	.077038	.017257	-.000656	.027954	6651	.1591
		(.336265)	(.002902)	(.005050)	(.000115)	(.005681)		
(b)	3	-2.005902	.007735	.030525	-.000446	.020776	6651	.1954
		(.287044)	(.002477)	(.004310)	(.000098)	(.004849)		

Notes: CON is the constant; EDHCO is the coefficient on years of schooling, TCO is the coefficient on time, EXCO is the coefficient on experience and EXSQCO is the coefficient on the square of experience. Standard errors are reported in parentheses.

(It should be noted that a small increase in the prior variance of σ_μ^2 implies a large increase in the potential variability of the μ_i themselves.)

We have much less information on potential measurement error. In particular, we have no economic theory in this area to aid us in forming priors. Thus the assumed prior mean of 0.20 is held with very little confidence reflected in the prior variance of 1.0.[12] Based on this prior information, the Bayesian scaled modal estimate is presented in columns 5 and 6 of Table 3.[13] As expected from the simulation results, the estimates are the same as the unbiased estimates to several decimal places because of the very large sample size.

This result implies that the incorporation of "reasonable" prior information results in no change in the full sample estimate and hence that truncation probably imposes "unreasonable" prior information. Table 5 shows just how unreasonable or dogmatic the implied prior specification must be in order to produce the same estimate of σ_μ^2 as that resulting from truncation. For example, the one percent truncation on a panel length of 6 years yields an estimate of σ_μ^2 much lower than the sample estimate (0.074 vs 0.129). Even if the prior belief regarding individual variability centres on a very low value--0.010--the classical estimate cannot be approached unless the prior is dogmatic in the extreme--i.e., $V(\sigma_\mu^2) = 0.00000015$. Indeed, even if only extreme outliers are deleted, the low prior mean of 0.01 has to be held with a very high degree of confidence ($V(\sigma_\mu^2) = .00000075$) in order to generate the same result as the truncation.[14]

The foregoing results suggest that while the coefficient vector $\hat{\beta}$ is relatively insensitive to the treatment of outliers, estimates of σ_μ^2 are extremely sensitive. Truncation is an extremely crude procedure for incorporating prior information and may unintentionally impose totally unreasonable prior beliefs. The prior distributions suggested in Section II above may be used by the investigator to clarify the nature of his prior beliefs revealed by a willingness to truncate data points and to assess whether or not any proposed truncation accurately reflects those beliefs.

IV. SAMPLE CONSTRUCTION AND VARIANCE COMPONENTS

In this section we broaden the analysis of the sensitivity of variance components estimators to the effects of general sample inclusion criteria or sample construction procedures. Choosing alternative methods of sample construction is not viewed as an attempt to impose prior beliefs, hence this section involves no Bayesian analysis. Thus far individuals have been included in the sample provided they have complete information on hours of work. This potentially involves a problem of sample selection bias or missing data bias. This has received a great deal of attention recently (see for example, Heckman (1979), Hausman and Wise (1979)) and several solutions have been proposed to correct the bias. In practice, however, the censoring involved in this case may have little effect on $\hat{\beta}$. Frequently, therefore, investigators may ignore the problem. However, the effect on the estimate of σ_μ^2 may be substantial, even if there is no effect on $\hat{\beta}$. Thus, when σ_μ^2 is the parameter of interest, great care must be taken in the data preparation or "cleaning" procedure, dealing with missing values, etc. This is dramatically illustrated in the case of the missing data on hours of work.

For the analysis of this problem the sample is increased by retaining individuals with missing data on either hours per week or weeks per year in the sample provided they meet the following criteria: (i) they must have at least one observation on at least one of the three

Table 5

Prior Specifications Approximately Equivalent to 1% and 5% Outlier Truncation

Truncation	Panel Length	τ^2_μ	Equivalent Bayesian Prior Specification		
			$\tilde{\sigma}^2_\mu$	Prior Mean	Prior Variance
1%	6	.074	.074	.01	.00000015
1%	3	.072	.072	.01	.00000016
5%	6	.050	.050	.01	.00000008
5%	3	.050	.050	.01	.00000011
Extreme Outliers	6	.113	.113	.01	.00000075
				.10	.0000085
Extreme Outliers	3	.118	.118	.01	.00000073
				.10	.000012

components of the wage rate (E,L,H) in every year of the panel; (ii) they must have at least one observation on each of the components of the wage rate at some time in the panel. Using this criterion rather than the "positive values in all years" criterion approximately doubles the sample size from 1187 to 2217. The missing values on E, L and H for a given individual are supplied under two alternative procedures. Alternative (a) uses the average of the individual's own values in other years to fill in the missing years. Alternative (b) uses a least squares prediction obtained from an ordinary least squares regression of E, L and H for each year on the exogenous variables, schooling, experience, and experience squared. Alternative (a) retains information specific to the individual; alternative (b) substitutes in effect an average value of all individuals in the sample and contains no individual specific element other than the particular values of the exogenous variables. These procedures are not offered as optimal procedures for supplying missing data. They are presented to illustrate the sensitivity of τ_μ^2 to different "reasonable" methods of sample construction that an applied researcher might use.

Table 6 reports the value of the unbiased estimator τ_μ^2 for the various sample inclusion criteria for panel lengths of 6 and 3 years. The first two rows report the previously obtained value of τ_μ^2 for the sample with complete data in all years. When individuals with missing data in some years are retained by giving them sample averages across the other individual's τ_μ^2 falls somewhat over the 6 year panel and increases marginally for the 3 year panel. Neither change is very marked. On the other hand, when alternative (a) is used to supply the missing values τ_μ^2 more than doubles for both panel lengths. By contrast, the estimated regression coefficient vector, $\hat{\beta}$, reported in the last 4 rows of Table 4 exhibit little change except the coefficients on the time trend and the constants which are common to all the individuals.[15]

Table 6

Sensitivity of τ_{μ}^{2} to Alternative

Criteria for Sample Inclusion

Criterion for Sample Inclusion	Panel Length	τ_{μ}^{2}	Sample Size
Complete data in all years	6	0.129	1187
	3	0.137	1187
Alternative (a)	6	0.302	2217
	3	0.297	2217
Alternative (b)	6	0.122	2217
	3	0.139	2217

It is interesting to note that the high value of τ_μ^2 obtained in this case (.30) relative to the "stylized fact" of approximately .15 is close to the estimate (.25) obtained by Kiefer and Neumann (1981). These authors consider explicitly the relation between individual components and sample inclusion from the perspective of the sample selection bias literature.

The different results for alternatives (a) and (b) are as expected. The individuals for whom some data is missing do not tend to be "average" individuals. Leaving them out altogether would therefore reduce individual variability in the sample. When the individuals are retained in the sample their effect on τ_μ^2 depends on whether their non-average individual components are retained in the procedure for replacing missing values. Clearly in the case of alternative (a) the individual component is retained. On the other hand, with alternative (b) the procedure tends to make the individual more "average". If this averaging tendency is strong enough to outweigh the inherent non-average characteristics of the individuals with missing data, overall individual variability may be reduced.

V. CONCLUSIONS AND FUTURE WORK

A frequent practice in empirical work is to "pre-analyze" the data via various sample inclusion rules. For example, truncation of "outliers" or "impossible values" is common. A requirement of "complete data" for all observations is almost universal.[16] Estimates of regression coefficients are relatively insensitive to these practices. However, estimates of variance-components are markedly affected. Truncation of "outliers" may be viewed as a crude way of incorporating prior information about variance components. We presented a Bayesian estimator that is simple to compute and incorporates prior information in a more flexible manner. In addition we reversed the usual Bayesian procedure and used our Bayesian estimator to assess the prior beliefs that an investigator imposes by any proposed truncation of outliers. We showed that especially in large sample, extremely dogmatic prior beliefs may, inadvertantly, be imposed when outliers are eliminated. We also showed that variance components are sensitive to general sample inclusion criteria. Thus, if assessing individual effects is the focus of analysis much greater attention than is usually the case must be paid to how the final sample is arrived at.

An obvious extension of our analysis is to the effect of crudely imposed prior beliefs, e.g., truncation, on inference. In many cases, for example, regression coefficients may be insensitive to sample construction. The estimated standard errors may, however, be sensitive. Thus while an investigator may encounter little risk of losing consistency of $\hat{\beta}$ when alternative methods of sample construction are used, the risk of invalid inference may be large.

Footnotes

[1] See Box and Tiao (1973) for a discussion of non-informative priors.

[2] See Box and Taio (1973) for a description of this distribution.

[3] In general the natural form of prior information will be in terms of the mean and variance. These may be turned into the parameters a and r by solving (7)-(8) as follows:

$$a = 2E(\frac{E^2}{V} + 1)$$

$$r = 2(\frac{E^2}{V} + 2).$$

[4] For the procedure of integrating out so-called nuisance parameters or parameters not of interest see Box and Tiao (1973).

[5] This is accomplished via estimating the nonlinear model (in the standard notation)

$$y_t = \frac{x_{1t}}{\sigma_\epsilon^2} + \frac{x_{2t}}{\sigma_\mu^2} + \frac{x_{3t}}{\sigma_\epsilon^2 + T\sigma_\mu^2} + \frac{x_{4t}}{(\sigma_\epsilon^2 + T\sigma_\mu^2)} + \frac{x_{5t}}{2(\sigma_\epsilon^2)^2} + \frac{x_{6t}}{2(\sigma_\mu^2)^2} + \text{error.}$$

The data are entered as

$$y = \binom{0}{0}$$

$$x = (x_{1t}, \ldots, x_{6t})$$

$$= \begin{bmatrix} -\eta_1 & 0 & -\eta_3 & s_2 & s_1 + a_\epsilon & 0 \\ 0 & -\eta_2 & -\eta_3 T & s_2 T & 0 & a_\mu \end{bmatrix}$$

The regression has zero degrees of freedom. Accordingly, the sum of squared errors is zero and the regression provides precise values for $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\mu^2$.

[6] The m.s.e. for the unbiased estimators are derived analytically in the appendix. The m.s.e. for the Bayesian estimators involves a four-fold integral. Because of computational considerations m.s.e. was obtained via sampling rather than (4-dimensional) numerical integration. M values of σ^2_ϵ and σ^2_μ were drawn from $p(\sigma^2_\epsilon)$ and $p(\sigma^2_\mu)$. For each pair, y_{it} was generated according to (1), yielding s_1 and s_2 and hence $\hat{\sigma}^2_\epsilon$ and $\hat{\sigma}^2_\mu$. The expected value and mean squared error were then calculated, e.g., $E(\hat{\sigma}^2_\mu) = \frac{1}{M} \sum_{j=1}^{M} \hat{\sigma}^2_{\mu_i}$. The large sample size (M = 150,000) ensures a very small error.

[7] The scaling factors for $\hat{\sigma}^2_\epsilon$ and $\hat{\sigma}^2_\mu$ respectively:

$$F_\epsilon = \frac{\eta_1 + \eta_3 + 2}{\eta_1 + \eta_3 - 2} = \frac{r_\epsilon + NT + 1}{r_\epsilon + NT - 3}$$

$$F_\mu = \frac{\eta_2 + \eta_3 + 2}{\eta_2 + \eta_3 - 2} = \frac{r_\mu + N + 1}{r_\mu + N - 3}$$

[8] Some writers use a more elaborate error structure, e.g., allowing for serial correlation in ϵ_{it} (Lillard and Willis, 1978). This is referred to as the "autocorrelated individual component model".

[9] For example, Carliner's (1980) estimate is 0.150; Lillard and Willis (1978) report 0.124 for whites, 0.146 for blacks and 0.125 for the pooled sample.

[10] This is a sub-sample obtained from the full sample by imposing sample inclusion criteria of the type usually employed in this area. This sample censoring itself has important consequences for the estimation of variance components which are discussed in Section IV.

[11] Deletion of one observation on an individual results in the individual being dropped from the sample.

[12] The results were insensitive to the prior specification on σ_ϵ^2 as long as it was not dogmatic.

[13] Because of the large sample size F_ϵ, $F_\mu \rightarrow 1$ hence the scaled and unscaled estimators become equivalent.

[14] Of course, an estimate equivalent to a truncation between 1-5% could be obtained with a prior distribution which is uniform on [0,.06]. In this case the prior mean is .03 and the variance.0003, which appears less dogmatic. However, in assigning a zero probability to a variance of the individual component greater than .06 is very extreme since it ensures that the investigator always ignores the data entirely when it falls outside a limited range he is willing to accept.

[15] Since the constant captures the part of individual components common to all individuals this will be affected when the sample changes to reflect the different average individual component. Similarl remarks apply to the time trend.

[16] An exception to this is the literature on sample selection bias. A good example that addresses the problem of missing data is Hausman and Wise (1979).

References

Box, G. E. P., and Tiao, G. C., Bayesian Inference in Statistical Analysis, Reading: Addison Wesley, 1973.

Carliner, G., "Permanent and Transitory Wage Effects in a Multiperiod Family Labour Supply Model," University of Western Ontario, Department of Economics Working Paper #8003, 1980.

Dhrymes, P. J., Econometrics: Statistical Formulations and Applications, New York: Springer-Verlag, 1974.

Hausman, J. A. and Wise, D. A., "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," Econometrica, 47, 455-474, 1979.

Heckman, J. J., "Sample Selection Bias as a Specification Error," Econometrica, 47, 153-162.

Kiefer, N. M. and Neumann, G. R., "Individual Effects in a Non Linear Model: Explicit Treatment of Heterogeneity in the Empirical Job Search Model," Econometrica, 49, 965-980, 1981.

Klotz, J. H., Milton, R. C., and Zacks, S., "Mean Square Efficiency of Variance Component Estimators," Journal of the American Statistical Association, 64, pp. 1383-1402, 1969a.

Lazear, E., "Age, Experience, and Wage Growth," American Economic Review, 66, 548-558, 1976.

Lillard, L. E. and Willis, R. J., "Dynamic Aspects of Earning Mobility," Econometrica, 46, 985-1012, 1978.

Portnoy, S., "Formal Bayes Estimation with Application to a Random Effects Model," The Annals of Mathematical Statistics, 42, pp. 1379-1402, 1971).

Smith, J. P. and Welch, F., "Inequality: Race Differences in the Distribution of Earnings," International Economic Review, 20, 515-526, 1979.

Appendix

1. Consistency of the Bayesian Estimators

To see that $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\mu^2$ are consistent estimators of σ_ϵ^2 and σ_μ^2, first recall that the unbiased estimators of σ_ϵ^2 and σ_μ^2 are

$$\tau_\epsilon^2 \equiv \frac{S_1}{\nu_1}$$

and

$$\tau_\mu^2 \equiv \frac{1}{T}\left[\frac{S_2}{\nu_2} - \frac{S_1}{\nu_1}\right] .$$

τ_ϵ^2 and τ_μ^2 converge in mean squared error to σ_ϵ^2 and σ_μ^2 as $N \to \infty$ for any T. Dividing (10) by η_1 yields

$$\text{(A1)} \qquad -\frac{1}{\hat{\sigma}_\epsilon^2} - \frac{\eta_3}{\eta_1} \cdot \frac{1}{\hat{\sigma}_\epsilon^2 + T\hat{\sigma}_\mu^2} + \frac{S_2}{2\eta_1(\hat{\sigma}_\epsilon^2 + T\hat{\sigma}_\mu^2)^2} + \frac{S_1 + a_\epsilon}{2\eta_1(\hat{\sigma}_\epsilon^2)^2} = 0.$$

Now, since $\eta_3/\eta_1 = (N-1)/[r_\epsilon + N(T-1) + 2]$, $\lim\limits_{\substack{N\to\infty \\ T\to\infty}} \frac{\eta_3}{\eta_1} = 0$. Also

$$\lim_{\substack{N\to\infty \\ T\to\infty}} \frac{S_2}{2\eta_1(\hat{\sigma}_\epsilon^2 + T\hat{\sigma}_\mu^2)^2} = \lim_{\substack{N\to\infty \\ T\to\infty}} \frac{S_2}{\nu_2} \cdot \frac{1}{2T} \cdot \frac{1}{\hat{\sigma}_\epsilon^2/T + \hat{\sigma}_\mu^2} \cdot \frac{\nu_2}{\eta_1} \cdot \frac{1}{T} \cdot \frac{1}{\hat{\sigma}_\epsilon^2/T + \hat{\sigma}_\mu^2} = 0$$

Further $\lim\limits_{\substack{N\to\infty \\ T\to\infty}} a_\epsilon/2\eta_1(\hat{\sigma}_\epsilon^2)^2 = 0$. Finally, since $f(\hat{\sigma}_\epsilon^2) = \frac{1}{(\hat{\sigma}_\epsilon^2)^2}$ is continuous at each point in the parameter space it follows that $\text{plim}\,(\frac{1}{(\hat{\sigma}_\epsilon^2)^2}) = 1/(\text{plim}\,\hat{\sigma}_\epsilon^2)^2$ (see Dhrymes, 1974, p. 110), hence

$$\text{plim}\,\frac{S_1}{2\eta_1(\hat{\sigma}_\epsilon^2)^2} = \text{plim}\,\frac{S_1}{\nu_1} \cdot \lim \frac{\nu_1}{r_\epsilon + \nu_1 + 2} \cdot \text{plim}\,\frac{1}{(\hat{\sigma}_\epsilon^2)^2} = \sigma_\epsilon^2 \times \frac{1}{(\text{plim}\,\hat{\sigma}_\epsilon^2)^2} .$$

Thus for large N and T, (41) becomes

$$\text{(A2)} \qquad \frac{1}{\text{plim}\,\hat{\sigma}_\epsilon^2}\left(\frac{\sigma_\epsilon^2}{\text{plim}\,\hat{\sigma}_\epsilon^2} - 1\right) = 0 ,$$

or plim $\hat{\sigma}_{\epsilon}^{2} = \sigma_{\epsilon}^{2}$. Division of (11) by η_3 yields plim $\hat{\sigma}_{\mu}^{2} = \sigma_{\mu}^{2}$ in a similar fashion.

Finally, as both F_{ϵ} and F_{μ} converge to unity, the scaled estimators $\tilde{\sigma}_{\epsilon}^{2}$ and $\tilde{\sigma}_{\mu}^{2}$ are also consistent at each point in the sample space.

2. Mean Squared Errors for the Bayesian and Unbiased Estimators

Let $\phi_{\mu}(S_1,S_2)$ be an estimator of σ_{μ}^{2}. The unconditional mean of $\phi_{\mu}(\cdot)$ is

$$E(\phi_{\mu}) = \int_0^{\infty}\int_0^{\infty}\int_0^{\infty}\int_0^{\infty} \phi_{\mu}(s_1,s_2)p(s_1|\sigma_{\epsilon}^{2})p(s_2|\sigma_{\epsilon}^{2},\sigma_{\mu}^{2})p(\sigma_{\epsilon}^{2})p(\sigma_{\mu}^{2})ds_1 ds_2 d\sigma_{\epsilon}^{2} d\sigma_{\mu}^{2} \, . \tag{A3}$$

Similarly, the uncondition mean squared error is

$$\begin{aligned} \text{m.s.e. } (\phi_{\mu}) &= E[(\phi_{\mu} - \sigma_{\mu}^{2})^{2}] \\ &= \int_0^{\infty}\int_0^{\infty}\int_0^{\infty}\int_0^{\infty} [\phi_{\mu}(s_1,s_2)-\sigma_{\epsilon}^{2}]^{2}\, p(s_1|\sigma_{\epsilon}^{2})p(s_2|\sigma_{\epsilon}^{2},\sigma_{\mu}^{2}) \\ &\quad p(\sigma_{\epsilon}^{2})p(s_{\mu}^{2})ds_1 ds_2 d\sigma_{\epsilon}^{2} d\sigma_{\mu}^{2} \, . \end{aligned} \tag{A4}$$

Analogous formulae apply to an estimator of σ_{ϵ}^{2}, $\phi_{\epsilon}(S_1,S_2)$.

For the unbiased estimators, τ_{ϵ}^{2} and τ_{μ}^{2}, since $S_1/\sigma_{\epsilon}^{2} \sim \chi^{2}(\nu_1)$ and $S_2/(\sigma_{\epsilon}^{2}+ T\sigma_{\mu}^{2}) \sim \chi^{2}(\nu_2)$, where S_1 and S_2 are independent, it can be shown that

$$E[(\tau_{\epsilon}^{2}- \sigma_{\epsilon}^{2})^{2} |\sigma_{\epsilon}^{2}] = \frac{2(\sigma_{\epsilon}^{2})^{2}}{N(T-1)} \, ,$$

and

$$E[(\tau_{\mu}^{2}- \sigma_{\mu}^{2})^{2} |\sigma_{\epsilon}^{2},\sigma_{\mu}^{2}] = \frac{1}{T^{2}} \left[\frac{2(\sigma_{\epsilon}^{2}+ T\sigma_{\mu}^{2})^{2}}{N - 1} + \frac{2(\sigma_{\epsilon}^{2})^{2}}{N(T-1)}\right].$$

Integrating under the priors $p(\sigma_{\epsilon}^{2})$ and $p(\sigma_{\mu}^{2})$ yields the unconditional means and variances

$$E(\tau_{\epsilon}^{2}) = \frac{a_{\epsilon}}{r_{\epsilon} - 2} \, ,$$

$$E(\tau_{\mu}^{2}) = \frac{a_{\mu}}{r_{\mu} - 2} \, ,$$

$$E[(\tau_\epsilon^2 - \sigma_\epsilon^2)^2] = \frac{2a_\epsilon^2}{N(T-1)(r_\epsilon - 2)(r_\epsilon - 4)}$$

and

$$E[(\tau_\mu^2 - \sigma_\mu^2)^2] = \frac{2a_\epsilon^2}{(r_\epsilon-2)(r_\epsilon-4)} \frac{1}{T^2} \left(\frac{1}{N-1} + \frac{1}{N(T-1)}\right) + \frac{4a_\mu a_\epsilon}{T(N-1)(r_\epsilon-2)(r_\mu-2)}$$

$$+ \frac{2a_\mu^2}{(r_\mu-2)(r_\mu-4)} \cdot \frac{1}{N-1} \, .$$

For the Bayesian estimators (A4) has to be integrated numerically.