I Introduction

"In statistical inference proper, the model is never questioned. . . . The methods of mathematical statistics do not provide us with a means of specifying the model." 1

Few would question the profundity of the above quotes. Yet in practice the methodological dictum contained therein is invariably violated. In practice we ordinarily face data with an imperfectly specified model and learn of our imperfection from the data themselves. For an example, a second round estimation procedure to clean up residual serial correlation is often employed after a Durbin-Watson statistic turns out to be insufficiently close to two. For another, we often read of a choice of algebraic form based on computation of $R^2$'s. And few of us can deny deciding whether to drop a variable based on the outcome of a $t$-test. Such procedures seem sensible from an intuitive point of view — data may not speak for themselves but they can raise a voice hard for a sensible man to ignore when a model is badly misspecified.

The opposite extreme, complete empiricism, is just as unattractive — perhaps more so. Theory is worth something. Meaningful research does not come from exhaustively searching through lists of potential explanatory variables. Presumably it is safe to ignore sunspot data and the length of women’s skirts in studying the demand for food. The reader of a research report is unconvinced by even nice results if he is made aware that a large number of trials with alternative specifications led sequentially to those results.

Most would agree to the broad statements made so far. The purpose of this paper is to define more rigorously the balance between two untenable methodological attitudes; that, on the one hand, data should not be allowed to interfere with specification and, on the other, specification must be achieved mainly by experimentation.

The vehicle used in the sequel to investigate costs and returns of sequential investigation is a linear regression model relating a regress and to two potential regressors. The assumption is that the researcher’s interest lies mainly in the effect of the first regressor and the question is whether to include or exclude the second independent variable in regression. Thus, the specific problem considered here is so limited that it arises seldom in practice as a specific case. But the simple problem contains within it features that are common with those of more realistic problems of model construction. The same considerations arise in more complex situations.

The results that follow are rigorous but the mode of presentation is nonrigorous. Those interested in proofs of results are referred to appropriate sources.

II The Nature of a Sequential Estimator in a Regression Model

The illustrative case examined here is that of Bancroft (1944). Let

$$Y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \epsilon_t; \quad t = 1, \ldots, T$$

(1)

represent a tentative specification of a model presumed to reflect a real world process of data generation where it is assumed that $\Sigma y_t = \Sigma x_{1t} = \Sigma x_{2t} = 0$. Suppose further that the error, $\epsilon_t$, is normally and independently distributed with constant variance, that the $x$’s are fixed and that $\beta_1$ is the parameter of interest. Potential inclusion of $x_2$ is for ceteris paribus purposes.

Let us consider two alternative least squares estimators for $\beta_1$, the parameter of interest:

$$b_1 = \frac{\Sigma x_{1t} y_t}{\Sigma x_{1t}^2}$$

(i)

$$b_{1.2} = \frac{\Sigma x_{2t}^2 \Sigma x_{1t} y_t - \Sigma x_{1t} x_{2t} \Sigma x_{1t} y_t}{\Sigma x_{2t}^2 \Sigma x_{1t}^2 - (\Sigma x_{1t} x_{2t})^2}$$

(ii)

Note that $b_1$ is the simple regression estimator that excludes consideration of the second independent variable, while $b_{1.2}$ is the full least squares estimator arising from the multiple re-
gression. The following results are well known:

1) The multiple regression estimator \( b_{1:2} \) is unbiased and has smallest variance among all linear unbiased estimators of \( \beta_1 \). The variance of \( b_{1:2} \) is \( \sigma^2 / \Sigma x_i^2 \) \((1 - r^2)\), where \( r \) is the correlation of \( x_1 \) and \( x_2 \) and \( \sigma^2 \) is the variance of \( \epsilon_i \).

2) The simple regression of \( y \) on \( x_1 \) excluding \( x_2 \) (\( b_1 \)) yields an estimator of \( \beta_1 \) that is biased unless \( r \) is zero or \( \beta_2 \) is zero. The variance of \( b_1 \) is \( \sigma^2 / \Sigma x_i^2 \) and hence is smaller than the variance of \( b_{1:2} \) (unless \( r \) is zero, in which case the two variances are equal).

Thus, if \( \beta_1 \) is zero, the simple regression estimator is the more precise estimator. This argues for the common procedure of deciding whether to keep \( x_2 \) or throw it away on the basis of the \( t \) (or \( F \)) test of the hypothesis \( \beta_2 = 0 \). If the line of argument stopped here, there would appear to be no cost to letting the data dictate the model. Unfortunately, the argument cannot rest here. Statisticians have long recognized that such a procedure gives rise to a third estimator: one based on the testing procedure.

The sequential estimator based on a \( t \)-test can be summarized as follows:

\[
\beta^*_1 = \begin{cases} 
  b_1 & \text{if } |t| \leq t_a \\
  b_{1:2} & \text{if } |t| > t_a
\end{cases}
\]

(2)

The sequential estimator of \( \beta_1 \), \( \beta^*_1 \), is the simple regression estimator \( b_1 \) if the absolute value of computed \( t \) for testing \( \beta_2 = 0 \) is less than a critical value of the \( t \) distribution given a choice of \( a \) and is the multiple regression estimator \( b_{1:2} \) if the computed \( t \) is to the right of the critical point. Thus, whether the sequential estimator \( \beta^*_1 \) is a good one in some sense depends on the probability characteristics of the distribution of \( \beta^*_1 \). As an initial caution that its characteristics are not simple, note that the expectation of \( \beta^*_1 \) is the conditional expectation of \( b_{1:2} \) given that \( |t| > t_a \) plus the conditional expectation of \( b_1 \), given \( |t| \leq t_a \). This expression, as given in Bancroft (1944), can be written only in integral form and involves every parameter thus far introduced, including \( a \), the type one error of the test that \( \beta_2 = 0 \). In general, we note that \( \beta^*_1 \) is biased since its expectation involves a weighted average involving the expectation of the generally biased estimator \( b_1 \). Since the least squares estimator with \( x_2 \) included is unbiased even if \( \beta_2 \) and \( r \) are both zero, the potential bias of \( \beta^*_1 \) is indicative that pretesting is not always a good idea.

Since both bias and variance of alternative estimators are of concern, evaluation of the alternatives calls for a single valued loss (or utility) function that gives weight to both.

### III Mean Squared Error as a Loss Function

An estimator, say, \( \Theta \), of a parameter \( \theta \) is the more desirable, the closer it lies to \( \theta \). Since \( \Theta \) is a random variable, we must speak of \( \Theta \) being close to \( \theta \) in some average (expected) sense. A widely accepted goodness criterion is mean squared error (MSE) of an estimator. Thus we say that \( \Theta \) is better than \( \Theta^* \) if the expected value of the squared difference of \( \Theta \) from \( \theta \) is smaller. I.e., we say that \( \Theta \) is better than \( \Theta^* \) if

\[
E((\Theta - \theta)^2) < E((\Theta^* - \theta)^2),
\]

(3)

where \( E \) is the “averaging” or expectation operator.

Note that MSE differs from variance only if the estimator is biased. One of the ways to characterize MSE is that it is the estimator variance plus the square of the estimator bias.

Another intuitive characterization of mean squared error is that it represents the “average” squared Euclidian distance of the estimator from the parameter value. And in this context there is a nice generalization of the MSE criterion to an estimator vector.

The usefulness of the mean squared error criterion for comparing the three alternative

---

1. Even if \( \beta_2 \) is not zero, it may be sufficiently small so that one might prefer the biased estimator with the smaller variance.
2. Computed \( t \) is the usual \( t = b_{1:2} / \sqrt{C_{xx} s^2} \), where \( b_{1:2} \) is the least squares estimate of \( \beta_2 \), \( C_{xx} \) is the appropriate element in the inverse of the sums of squares and cross-products matrix and \( s^2 \) is the unbiased estimator of the error variance.
3. One of the common confusions is between the parameter MSE and the random variable EMS (error mean square). The latter is common notation for the denominator of the \( F \) in ANOVA tables, varies from sample to sample and is an observable quantity while the former is an unknown parameter that is averaged over repeated samples.
4. If we take \( \Theta \) as an \( n \times 1 \) vector of parameters and \( \tau \) as an \( n \times 1 \) vector estimator, the average squared distance of the estimator point from the parameter point in \( n \) dimensional space is \( E(\tau - \Theta)^T (\tau - \Theta) \). This is equivalent to the sum of the mean squared errors for each \( \Theta \), (1971).
estimators for $\beta_1$ as discussed in section II lies in its capability for making comparisons of biased and unbiased estimators. Essentially, MSE is a single valued function of bias and variance so that one can conveniently discriminate between unbiased estimators with larger variances and biased estimators which have smaller variances. As will become clear, estimators having smallest variance in the class of unbiased estimators may have variances so large as to make them relatively imprecise.

IV MSE’s for $b_1$ and $b_{1.2}$

From an earlier paper (1964), one can find derivations of the following results:

$$\text{MSE} (b_1) = \frac{\sigma^2}{\sum x_1^2} + \frac{\beta_2^2 r^2}{\sum x_1^2},$$

(4)

where the first term in (4) is the variance of the simple regression estimator of $\beta_1$ and the second term is its squared bias.

Further

$$\text{MSE} (b_{1.2}) = \frac{\sigma^2}{\sum x_1^2 (1 - r^2)},$$

(5)

and to have $b_1$ better in MSE is equivalent to the requirement that

$$\lambda = \frac{\beta_2^2}{\text{var} |b_{2.1}|} < 1,$$

(6)

We are justified in calling the ratio of $\beta_2^2$ to the variance of the full regression estimator of $\beta_2$ by a special symbol, $\lambda$, for several reasons. For one that will suffice now, $\lambda$ turns out to be the critical parameter in comparing not only the regression estimators of $\beta_1$ with $x_2$ in and out, but in comparing both of those with the sequential estimator based on the test of the hypothesis $\beta_2 = 0$.

Note how closely $\lambda$ relates to the usual statistic for testing the hypothesis that $\beta_2$ is zero. The usual test statistic is obtained by replacing the terms of the ratio by their unbiased estimators — e.g.,

$$\mu = \frac{b_{2.1.2}^2}{s^2 b_{2.1}}.$$  

(7)

And under the null hypothesis that $\beta_2$ (or $\lambda$) is zero, $\mu$ is distributed as a Snedecor $F$ with one degree of freedom in the numerator and $T-3$ in the denominator.\(^6\)

Interestingly, there are those who have advocated retaining a variable if the $\mu$ of (7) exceeds unity and discarding it if the value of $\mu$ is smaller than one. In an intuitive way this recommended procedure corresponds to the parametric condition (6) that tells us whether $b_1$ or $b_{1.2}$ is better in the sense of mean squared error.

Unfortunately, $\mu < 1$ in no way assures us that $\lambda < 1$.\(^7\) As a matter of fact, the criterion of choice based on whether $\mu < 1$ falls in the class of sequential estimators. Taking $\mu < 1$ as the rule for dropping $x_2$ gives us the estimator $\beta_\ast_1$, where the significance level, $\alpha$, is residually determined by degrees of freedom. Hence, we can consider the estimator arising from deciding on either $b_1$ or $b_{1.2}$ on the basis of whether $\mu \leq 1$ as a special case of the sequential estimator $\beta_\ast_2$ already introduced, construing the suggestion as an indirect recommendation on a choice of $\alpha$.

The results so far can be summarized usefully in a diagram.

The $b_1$ estimator is best if $\lambda (\beta_2)$ is zero, the intercept of the loss function being its minimum and at that point attaining the value of its variance. As $\lambda$ increases, the bias term adds to the loss function in a linear manner. Alternatively, under the assumption of fixed $x$’s, fixed regression variance and given sample size, the loss function for the full least squares estimator is perfectly flat — its bias is zero. If we had precise prior information on $\lambda$, the procedure is clear. If we want to minimize losses in terms of MSE, we would take $b_1$ for $\lambda < 1$ and $b_{1.2}$ for $\lambda > 1$.

Stating the results in a slightly altered fashion, one might consider two cases:

1) If $\lambda > 1$ and the researcher ignores $x_2$, he is bearing a cost for underspecifying his model. Ignoring $x_2$ in this case yields an estimator for $\beta_1$ that is so highly biased that the bias outweighs in MSE any gains in reduced variance.

2) If $\lambda < 1$ and the researcher includes $x_2$,

\(^6\) Equivalently, the square root of $\mu$ is a $t$ with $T-3$ d.f. (degree of freedom) under the hypothesis that $\lambda = 0$.

\(^7\) Dropping $x_2$ if $\mu < 1$ does assure us of maximizing $R^2$, the coefficient of determination adjusted for degrees of freedom. See Haitovsky (1969). However, we are guaranteed nothing about “goodness” of the resulting $\beta_1$ estimator in any parametric context.
he is paying a cost for overspecifying his model. Including $x_2$ eliminates the risk of bias but the loss in terms of increased variance is more than offsetting.

It is the usual case for the researcher to have only vague prior knowledge of $\lambda$. Certainly he feels that $\beta_2$ is not zero — otherwise he would not have included $x_2$ in the first specification. However, the researcher can observe $\mu$ and may choose in a sequential fashion to make adjustments in his initial specification.

As a brief digression, consider what happens to the graphs of the loss functions in figure 1 as we change the $x_2$ variable in such a way as to increase $r^2$. The loss function for the full least squares estimator would shift upward (i.e., the variance of $b_{1.2}$ increases). The intercept of the loss function for $b_1$ remains fixed, since the $b_1$ estimator is unbiased for $X = 0$ and its variance is independent of $r^2$. And the point of intersection of the two loss functions remains at $X = 1$, hence the slope of MSE($b_1$) necessarily increases as $r^2$ increases. Thus for small $\lambda$ ($\lambda < 1$), an increase in the correlation of the independent variables makes for larger regret.

These arguments generalize in various ways. See Bock et al. (1971) (Torro-Vizcarrondo and Wallace (1968). In particular, if one starts with a general linear model and considers a set of $m$ linear restrictions and an average squared distance criterion as in footnote 5, the requirement that the restricted estimator be better than unrestricted least squares is to have $\lambda < m$ where $\lambda$ is a generalized parameter involving the restrictions. See Wallace (1971).

\[ \beta^*_{1,2} = \delta(\mu) b_{1,2} + (1 - \delta(\mu)) b_1 \]

where $\delta(\mu) = 1$ if $\mu$ is in the rejection region of the $F$ distribution for a particular choice of $\alpha$.

We now turn to the loss function of the sequential estimator.

\section*{V Loss Function for $\beta^*_{1,2}$}

The probability characteristics of the sequential estimator are rather messy. However, quite a bit is known about the nature of the loss function for $\beta^*_{1,2}$ and some intuitive justification can be provided. Most of the discussion here is based on Ashar (1970), although some of the results can be found in a paper by Larson and Bancroft (1963).

First, note that we can express the sequential estimator as a (random) convex linear combination of the two estimators already considered, where the weights take the values zero or one depending on a test outcome.

\begin{figure}[h]
  \centering
  \caption{Determination of the Weights of the Sequential Estimator}
  \includegraphics[width=\textwidth]{figure2.png}
  \label{fig:2}
\end{figure}

In particular, if one starts with a general linear model and considers a set of $m$ linear restrictions and an average squared distance criterion as in footnote 5, the requirement that the restricted estimator be better than unrestricted least squares is to have $\lambda < m$ where $\lambda$ is a generalized parameter involving the restrictions. See Wallace (1971).
we see that given $a$, the critical rejection region lies to the right of $u_a$ and if $\mu$ falls to the right of $u_a$, $\beta^*_1$ would take the value $b_{1.2}$. If $\mu$ were to the left of $u_a$, $\delta(\mu)$ is zero and hence $b_1$ would be the estimator chosen.\(^9\) Thus, the "messiness" of the probability characteristics of $\beta^*_1$ can be appreciated — $\beta^*_1$ is the sum of products of random variables and $b_{1.2}$ is not independent of $\mu$, hence one of the products involves non-independent random variables.

To provide some additional intuitive basis for understanding the nature of $\beta^*_1$, note that if we choose $a = 0$, the critical region vanishes. Hence for $a = 0$ we never reject the hypothesis that $\beta_2 = 0$ and in this case the sequential estimator is identical with $b_1$. Alternatively, if we make $a = 1$, the critical region is the entire range of $u$ in figure 2 — thus we would always reject the hypothesis $\beta_2 = 0$. So, for a choice of $a = 1$, the $\beta^*_1$ estimator is always the full least squares estimator $b_{1.2}$. This provides information that for extreme choices of $a$, the loss function for $\beta^*_1$ becomes identical to the loss function of either $b_1$ or $b_{1.2}$. The loss functions of figure 1 are in some sense binding for the loss function of $\beta^*_1$, depending on $a$. Intuitively then, the loss function for $\beta^*_1$ is critically dependent on a choice of $a$.\(^10\)

The loss function for $\beta^*_1$ can best be discussed diagrammatically and in reference to the loss functions for $b_1$ and $b_{1.2}$.

In reference to figure 3, it is important to first state that the loss function as drawn for the sequential estimator is for a particular choice of $a, 0 < a < 1$. Varying $a$ systematically changes the MSE ($\beta^*_1$) in ways to be noted later, but certain characteristics of the function are true for every $a$ in the open interval from zero to one.

1) The minimum loss of the sequential estimator is for $\lambda = 0$, but that minimum is above the minimum loss for $b_1$. Intuitively, if $\lambda$ is identically zero, chances are good $(1 - a)$ that the pre-testing procedure will tell us to do the right thing — throw away $x_2$. But occasionally $(100a$ per cent of the time) we will erroneously keep $x_2$, hence the average loss over all possible trials reflects the occasional wrong decision.

2) The maximum loss of the sequential estimator lies to the right of the loss function for $b_1$. It (the maximum) can be shown to be always to the right of $\lambda = 5/2$.

3) The loss function for $\beta^*_1$, increases monotonically from $\lambda = 0$ to its maximum value and then declines monotonically as $\lambda$ increases beyond the maximum. Intuitively, for any choice of $a$, $\lambda$ can be so big that we would reject the hypothesis that $\lambda = 0$ in a large number of cases. This is a counterpart in sequential estimation to the well-known fact that the power of the test increases the larger is $X$. However, for any finite $\lambda$, there is a finite probability that we will accept a false hypothesis (type two error) so the loss function for $\beta^*_1$ remains above the loss function for the full least squares estimator beyond the point where the two curves intersect. However, the two approach each other as $\lambda$ gets larger.

4) For $\lambda < 1/2$, the sequential estimator dominates (has smaller loss than) the estimator that retains $x_2$ ($b_{1.2}$). This range of $\lambda$ is comforting to those who pre-test, as is the range of $\lambda$ greater than $\lambda^*(a)$.\(^11\) However, there is no free lunch. There is a finite range of $\lambda$, $1/2 < \lambda < \lambda^*(a)$, where pre-testing is the worst of the three alternatives.

\(^9\) Figure 2 is drawn to represent the Snedecor $F$, thus the null hypothesis is $\lambda = 0$ (or $\beta_2 = 0$). Alternatively, one could test $\lambda < 1$ (1968), but in the context of sequential estimators and given the nature of the non-central $F$, this can be characterized as an alternative choice of $a$. (See (1968) or (1964) for explanation.)

\(^10\) Ordinary tabulations of the $F$ statistic are for $a = .01$, $a = .05$. Uncritical use of such tables for purposes of rejecting (keeping) variables in regression indicates a lack of awareness of the result just stated.

\(^11\) The argument $a$ is a reminder that the specific value $\lambda^*$ beyond which the sequential estimator $\beta^*$, dominates $b_1$ is influenced by the choice of $a$. 

---

**Figure 3. — Typical Loss Function for the Sequential Estimator for Fixed $X$'s, Fixed $a^2$ and Given Sample Size**
5) To bring us back to the Malinvaud quote, there is no range of \( \lambda \) over which the sequential estimator is better than both \( b_1 \) and \( b_{1,2} \). Thus the rank empiricist can never hope to beat an enlightened theorist.

As regards the relationship of the loss function for \( \beta^{*}_{1} \) to the choice of \( a \), some observations can be made that have intuitive appeal. First, the larger is \( a \), the flatter the loss function for \( \beta^{*}_{1} \). It can be seen from figure 2 that a larger value of \( a \) corresponds to greater dominance of \( b_{1,2} \) in the sense of average weighting in the sequential estimator \( \beta^{*}_{1} \). As a consequence, there is small gain over least squares \( b_{1,2} \) even for the maximum MSE(\( \beta^{*}_{1} \)). For \( a \) close to zero, the loss function for the sequential estimator more nearly approximates the loss function for \( b_{1,2} \). The gains over \( b_{1,2} \) are bigger if \( \lambda \) is small — the costs relative to \( b_{1,2} \) are larger for intermediate \( \lambda \).

Many of the results stated here for the sequential estimator in the simple case of a two-variable model with a potential zero restriction on \( \beta_{2} \) have been generalized in two recent papers by Bock, Judge and Yancey (1971). They investigate two alternative generalized loss functions in the context of whether to impose general linear restrictions in a general linear model.

VI Summary and Conclusions

A natural question that arises is whether, in the class of pre-test estimators, there is in some sense an optimal choice of the critical point \( u_{a} \) or, equivalently, an optimal choice of \( a \). Given the nature of the loss functions as depicted in figure 3, one might define regret as

\[
R_{1}(a) = \text{MSE} \beta^{*}_{1} - \text{MSE} b_{1,2}.
\]

As noted by Gun (1965) and others, defining regret in this way yields a minimax choice for \( \beta^{*}_{1} \). That is, for \( a \) close to zero, the loss function for the sequential estimator more nearly approximates the loss function for \( b_{1,2} \). The gains over \( b_{1,2} \) are bigger if \( \lambda \) is small — the costs relative to \( b_{1,2} \) are larger for intermediate \( \lambda \).

As a practical guide to empirical research, the results reported here have the following implications: (1) If one has a strong prior that \( \lambda \) is either greater or less than 1/2, then no pre-testing is called for. One simply treats \( \beta_{2} = 0 \) in the latter and \( \beta_{2} \neq 0 \) in the former if the mean squared error loss function is taken as a guide. (2) Any intermediate prior about \( \lambda \) can be reflected thru choice of \( a \). The stronger the belief that \( \beta_{2} \neq 0 \), the smaller should be the choice of \( a \) and conversely, \( a \) should be larger the stronger the prior doubts about inclusion of \( x_{2} \). Of course, if one could cast priors into a precise distributional form, there are both classical and Bayesian procedures to by-pass the pre-testing altogether (1961). The question that motivated this paper was: How much should one rely on statistical testing to arrive at the specification of a model? The answer that emerges is that there are no precise rules. Yet what emerges is sensible. There are costs to ignorance that cannot be avoided by statistical procedures. If we are not sure a priori whether a particular variable(s) belongs in a model, pre-testing in the sense discussed here or in more elaborate step-wise regression selector programs will, on average, yield estimators that are sure to be worse than least squares estimators derived from an accurate prior specification.13 And the other side of the coin is that data can tell us something valuable about specification.

12 There are related estimators for which this statement will not hold. See Ashar (1970).
Taking heed of standard tests in certain instances can eliminate costly errors in specification. Work remains to be done to establish optimal rules for pre-testing, even in the simplest case.

REFERENCES
