SEQUENTIAL MODEL BUILDING FOR PREDICTION IN REGRESSION ANALYSIS, I[1]

By Harold J. Larson[2] and T. A. Bancroft

Stanford Research Institute and Iowa State University

1. Introduction.

1.1 *Completely and incompletely specified linear regression models.* Regression analysis considered here will be concerned with the fitting of linear models for prediction and the analysis of such fitted linear models. Given the specification that the data arose from an investigation adequately represented by a linear model with one predictand (dependent variable) and a definite number of predictors (independent variables), statistical theory provides routine mathematical techniques for obtaining estimates of the population regression coefficients. With the further assumption that the errors are independently normally distributed with mean zero and constant variance, statistical theory provides additional routine mathematical procedures for testing hypotheses and setting confidence intervals.

In the application of the theory of regression analysis to specific data, there is often some uncertainty as to the exact number of predictors to include in the specification of the linear model. Of course, there may also be some uncertainty as to whether the most appropriate specification model should be linear in form, particularly in some new area of research. In this study we shall limit the discussion to linear specification models or to such models which, upon appropriate transformation, may be fitted and analyzed as linear models. In some cases an assumed linear model may provide a good approximation of the true non-linear population model if only a small range of values for the predictors is of interest. We shall be dealing here with a special kind of "incompletely specified models" in the sense that the exact number of predictors to be included in the final fitted linear regression model will be determined by some decision rule based on the data of the investigation.

R. A. Fisher [8] pointed out as early as 1922 that one of the fundamental problems of theoretical statistics is the specification of an appropriate model for an investigation. Even though in more recent times non-parametric or distribution-free statistical inference procedures, not requiring a model specification, have been and are being considerably extended, nevertheless most applications of statistics in scientific research are still parametric in nature.

In introducing the use of analysis of variance techniques in experimental design situations, Fisher stipulated that for every well-designed experiment

Received February 13, 1962; revised September 18, 1962.

[1] Journal Paper No. J-4567 of the Iowa Agricultural and Home Economics Experiment Station, Ames, Iowa, Project 169.

[2] Now at U. S. Naval Postgraduate School, Monterey, California.

there can be only one correct analysis and test(s) of significance are completely determined before the experimental results are available. In this case we may refer to an analysis as being determined by a completely specified model. Presumably the advocates of this philosophy of inference might be expected to extend it to regression analysis, since it is well known that such analysis may be formulated in terms of analysis of variance procedures. However, in the use of regression analysis in experimental design, situations frequently arise in which the model is not completely specified, e.g., in deciding what degree polynomial to fit in a response surface study. Also, with the wider application of regression analysis, particularly in observational (non-experimental situation) investigations, the exact number of predictors in the linear model is often incompletely specified.

A completely specified linear regression model, appropriate to an investigation, may be determined by the investigator in some cases from theoretical considerations in the substantive field and/or from considerable experience with data from previous similar investigations. In many cases, however, particularly in new areas or fields of investigation, the investigator may turn to some objective specification decision rule to supplement scarce a priori theoretical knowledge or experience in the particular field of investigation or as an aid in determining an appropriate model. For the most part, such specification decision rules have been selected on intuitive grounds and have involved the use of preliminary test(s) of significance without proper consideration of distortion (extent of bias and actual versus nominal probability levels) of subsequent inference with regard to the final fitted model.

1.2 *Related papers.* The problems to be discussed here fall in the general class of problems designated as *problems of incompletely specified models involving the use of preliminary tests of significance.* In the discussion presented here the preliminary tests of significance are sequential in nature. Previous related papers in this general area include studies by Bancroft [1], Mosteller [16], Kitagawa [12], [11], Paull [17], Bechhofer [2], Bennett [3], [4], Huntsberger [10], Bozivich, Bancroft and Hartley [6], Larson [13], [14], and McCullough [15]. These studies are concerned primarily with basic investigations of the consequences of the use of certain decision rules involving preliminary tests of significance as aids in determining appropriate model specifications and effects on subsequent inference. It would seem appropriate to designate the class of all such problems of model fitting and subsequent inference as the *analysis of incompletely specified models.* In such investigations one is interested in such characteristics or properties as the magnitudes of specification bias and mean square errors for estimation problems; and effects on Type I and Type II errors for tests of the main hypotheses. It would appear most important to both theoretical and applied statisticians to have these characteristics and properties spelled out even though they may involve nuisance parameters and hence not be directly useable in the inferential process in a simple manner (e.g., as would be a t-test, etc.). Such studies provide insight as to how well specification bias and mean square errors are controlled by the particular decision rule considered under certain assumptions regarding the magnitudes or range of magnitudes of the nuisance parameters involved.

1.3 *Objectives of the present study.* Many different "objective" rules and methods of procedures have been suggested for determining, in situations of uncertainty, the number of predictors to be included in the final fitted linear regression model (see Summerfield and Lubin [20], Fireman and Wadleigh [7], Paperzak [18], and Hollingsworth [9]). Two rules that have proved popular are the following:

(A) The experimenter has measurements available on the $k + 1$ variables $(y, x_1, x_2, x_3, \cdots, x_k)$. He wants to predict the values of y on the basis of the values the x_i's will assume. He arbitrarily decides to test that the coefficient of x_k is 0 (i.e., x_k does not need to be in the equation). If he accepts this hypothesis he deletes x_κ and tests that the coefficient of x_{k-1} is 0. If he accepts the second hypothesis he deletes x_{k-1} from the prediction equation and tests the coefficient of x_{k-2} to be zero, etc. He continues deleting variables in this manner until he rejects a hypothesis that a coefficient is 0, or until he reaches the coefficient of x_r $(r < k)$, then retains in his prediction equation the variable corresponding to that coefficient or that of x_r and all other variables whose coefficients he has not yet tested. Thus, he is acting as if he has an a priori "order of importance" ranking of $x_1, x_2, \cdots, x_k$ and is deleting the "least important" variables, in order, from the end of this equation.

(B) The experimenter has measurements available on the $k + 1$ variables $(y, x_1, x_2, x_3, \cdots, x_k)$. He also assumes that the first r $(r < k)$ of these k variables are necessary for prediction of y. He then tests that the coefficient of x_{r+1} is zero. If he rejects this hypothesis he adds x_{r+1} to the list of necessary variables and tests that the coefficient of x_{r+2} is 0. If he rejects this second hypothesis he adds x_{r+2} to the list of necessary variables and tests that the coefficient of x_{r+3} is 0, etc. He continues adding variables to his prediction equation in this manner until he arrives at a variable whose coefficient does not differ significantly from 0, at which point he does not add that variable to the equation, nor does he add the variables whose coefficients he has not yet tested. Thus, in this case, the experimenter is beginning with a "basic core" prediction equation and adding on variables in order of importance.

Both of the above rules assume that the investigator has independent knowledge, in advance, of the "order of importance" of the predictor variables. In many cases such knowledge may be available to the investigator from theoretical considerations in the particular substantive field and/or from previous experience with similar data. If such is not the case, the investigator could conduct a preliminary study with independent data, or use a subsample of the available data to decide independently the "order of importance" of the predictor variables. One such ordering from independent data could be constructed by calling the variable (x_1, say) with the highest simple correlation with y the most important, the variable (x_2, say) with the highest partial correlation with y and x_1, the second most important, etc. (See Schultz and Goggans [19]).

The objective of the present study is to provide a means of examining critically the consequences of the decision rules mentioned in (A) and (B) for determining the number of predictors to be included in the linear regression model, with regard to the bias and mean square error of the estimate of the predictant y. This will be accomplished by deriving formulas for the bias and mean square error of the estimates of the predictant y for both of the two rules mentioned, assuming normal theory and that σ^2 (the population variance of the y's) is known (also assuming that an order of importance and hence the order of testing of the predictor coefficients is decided independent of the sample values used in the analysis of the incompletely specified regression model). The bias function in both cases turns out to be a linear function of the "doubtful" x_i values ($i = r + 1, r + 2, \cdots, k$). While future publication of more extensive tables is planned, as an illustration a few tabular values of a realistic range of the nuisance parameters are given in the appendix, permitting numerical evaluation of the two bias functions and the accompanying mean square errors for these particular cases and for any assumed values of the "doubtful" x_i's. The integrals defining the bias for the case of σ^2 unknown are displayed, but no closed form has been found for them.

2. Procedure A. An experimenter has available n measurements on $k + 1$ variables ($y, x_1, x_2, \cdots, x_k$) and is interested in which of the kx-variables will be necessary in a linear model to predict values of y. He arbitrarily decides to test that the coefficient β_k of x_k is zero. If he rejects this hypothesis he uses $x_1, x_2, \cdots, x_k$ to predict y. If he accepts β_k to be zero he tests that β_{k-1}, the coefficient of x_{k-1}, is zero. If he rejects his second hypothesis he uses $x_1, x_2, \cdots, x_{k-1}$ to predict y. If he accepts β_{k-1} to be zero he proceeds to test that the coefficient β_{k-2} of x_{k-2} is zero, etc., continuing in this way until he reaches the variable x_r. At this point he simply stops testing and retains all the variables whose coefficients have not been tested. (Thus the experimenter is acting as if he had a priori knowledge that $x_1, x_2, \cdots, x_r$ are necessary in predicting y and that x_{r+1} is "more important" than x_k.) Thus, mathematically, we have the following situation: (The estimator of the true value of y used in any particular case is denoted by y^*, the subscript on y denotes the number of independent variables included in the fitted linear regression equation.)

Event	y^* =	Situation
A_i	y_{k-i}	Reject $H_i : \beta_{k-i} = 0$; Accept $H_{i-1} : \beta_{k-i+1} = 0; \cdots$ Accept $H_1 : \beta_{k-1} = 0$; Accept $H_0 : \beta_k = 0$. $i = 0, 1, 2, \cdots, k - r - 1$.
A_{k-r}	y_r	Accept $H_{k-r-1} : \beta_{r+1} = 0$; Accept $H_{k-r-2} : \beta_{r+2} = 0; \cdots$; Accept $H_1 : \beta_{k-1} = 0$; Accept $H_0 : \beta_k = 0$.

Hence we see that we have a problem concerning model specification, i.e., the data itself is used to tell us which of these $k - r + 1$ models we should use in predicting values of y.

2.1 *Expected value of* y^*. Assume the true model generating our data is $y = X\beta + e$, where y is the $n \times 1$ vector of observed y values, X is the $n \times k$ matrix of x-values, e is the $n \times 1$ vector or error components. We further assume e is multivariate normal with $Ee = \phi$ (the null vector) $Eee' = \sigma^2 I$, and that $X'X = I$ (the results for the nonorthogonal case can be derived very easily from the orthogonal situation: see Section 4 for proof). The orthogonality assumption plus the assumed distributional properties of e are sufficient for the numerators of our tests of the β_i's ($i = r + 1, r + 2, \cdots, k$) to be independent. Let us first look at the situation in which the error variance σ^2 is known, then turn our attention to the case with σ^2 unknown.

With σ^2 known the test criterion for the hypothesis $\beta_i = 0$ is b_i^2/σ^2, ($i = r + 1, r + 2, \cdots, k$). The corresponding hypothesis is rejected if $b_i^2/\sigma^2 \geqq \lambda$ (the $100\alpha\%$ point of the χ_1^2 distribution), and accepted otherwise. Then the expected value of y^* obviously is $Ey^* = E(y_k \mid A_0)P(A_0) + E(y_{k-1} \mid A_1)P(A_1) + \cdots + E(y_r \mid A_{k-r})P(A_{k-r})$. Further

$$E(y_i \mid A_{k-i})P(A_{k-i}) = [\beta_0 + \beta_1 x_1 + \cdots + \beta_{i-1}x_{i-1} + x_i E(b_i \mid A_{k-i})]P(A_{k-i})$$

for $i = r + 1, r + 2, \cdots, k$ and

$$E(y_r \mid A_{k-r})P(A_{k-r}) = [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r]P(A_{k-r})$$

since all the estimators b_i are mutually independent. Next let us define

$$F_i = 2 \exp\left\{-\frac{1}{2}(\lambda + \beta_i^2/\sigma^2)\right\} \sinh\left(\beta_i(\lambda)^{\frac{1}{2}}/\sigma\right)$$

$$H_i = \left\{\int_{-\infty}^{-\lambda^{\frac{1}{2}}-\beta_i/\sigma} + \int_{\lambda^{\frac{1}{2}}-\beta_i/\sigma}^{\infty}\right\} [1/(2\pi)^{\frac{1}{2}}] \exp[-z^2/2]\, dz,$$

$i = r + 1, r + 2, \cdots, k$. Note that both F_i and H_i are functions only of λ and β_i/σ. It follows from the assumptions above that b_i is distributed normally with mean β_i and variance σ^2, $i = 1, 2, 3, \cdots, k$ and therefore

$$E(b_k|A_0)P(A_0) = [\sigma/(2\pi)^{\frac{1}{2}}]\, F_k + \beta_k H_k$$

$$E(b_{k-i}|A_i)P(A_i) = \prod_{j=k-i+1}^{k} [1 - H_j]\{[\sigma/(2\pi)^{\frac{1}{2}}]\, F_{k-i} + \beta_{k-i} H_{k-i}\}$$

$$i = 1, 2, \cdots, k - r - 1.$$

Making use of the following identities $\sum_{j=0}^{k-r} P(A_j) = 1, \sum_{j=0}^{i} P(A_{k-r-j}) = \prod_{m=r+i+1}^{k} [1 - H_m]$ and combining various terms we can now write

$$E(y^*) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=r+1}^{k} x_i \prod_{j=i+1}^{k} [1 - H_j]\{[\sigma/(2\pi)^{\frac{1}{2}}]\, F_i - \beta_i(1 - H_i)\},$$

where $\prod_{j=k+1}^{k}(1 - H_j) \equiv 1$. If we now define

$$\text{bias} \equiv Ey^* - [\beta_0 + \textstyle\sum_{k=1}^{i} \beta_i x_i]$$

we have

$$\text{bias} = \sum_{i=r+1}^{k} x_i \prod_{j=i+1}^{k} (1 - H_j)\{[\sigma/(2\pi)^{\frac{1}{2}}]F_i - \beta_i(1 - H_i)\}$$

or

$$\frac{\text{bias}}{\sigma} = \sum_{i=r+1}^{k} x_i \prod_{j=i+1}^{k} (1 - H_j)\{[1/(2\pi)^{\frac{1}{2}}]F_i - (\beta_i/\sigma)(1 - H_i)\}.$$

This function can now be fairly easily tabulated, the coefficients of x_i's being dependent only on the parameters β_i/σ, $i = r + 1, r + 2, \cdots, k$, and the significance level α of the preliminary tests of hypotheses.

Although a publication is planned of more extensive tables to be used in finding the coefficients of the x_i's necessary to evaluate the bias/σ function for Procedure A, Table I ($\alpha = .05$, $\lambda = 3.84$), Table II ($\alpha = .10$, $\lambda = 2.71$), Table III ($\alpha = .32$, $\lambda = 1$), given in the appendix, provide a means of finding such coefficients for realistic values of β_i/σ, if we have only two "doubtful" predictors and for the selected values of the significance level α of the preliminary tests of significance. To evaluate the bias/σ function simply enter the desired Table with the assumed values of β_k/σ and β_{k-1}/σ and read off the coefficients of x_k and x_{k-1}. Then for whatever values of x_k and x_{k-1} we wish to predict y we need only substitute them in the bias/σ equation. Since the b_i's are normally distributed with mean β_i and variance σ^2, then a realistic range for β_i/σ is from -3 to $+3$.

Example 1. Given a linear regression model with eight predictors of which x_7 and x_8 are "doubtful". Procedure A is to be used to decide whether or not to retain them, first testing β_8 and then testing β_7 (if necessary) to be zero, at the $\alpha = .10$ significance level. Then no matter what the outcome of the test(s), we find the coefficients of $x_8 = x_k$ (the first predictor variable tested for retention) and $x_7 = x_{k-1}$ (the second one tested) from Table II. Suppose we wish to predict y with $x_7 = -1$ and $x_8 = 1$, assuming $\beta_8/\sigma = \beta_k/\sigma = \frac{3}{2}$, $\beta_7/\sigma = \beta_{k-1}/\sigma = 2$, then

$$\begin{aligned}\text{bias}/\sigma &= (-.193)(-1) + (-.466)(1)\\ &= -.253\end{aligned}$$

while with $\beta_k/\sigma = \frac{1}{2}$, $\beta_{k-1}/\sigma = 3$ we have

$$\begin{aligned}\text{bias}/\sigma &= (-.087)(-1) + (-.262)(1)\\ &= -.135.\end{aligned}$$

(Note that all tabular values are negative for positive values of the β_i/σ, $i = k - 1, k$.) If we wish to evaluate the bias/σ function with one or both of the

(β_i/σ)'s$(i = k - 1, k)$ negative, then the coefficient of the corresponding $x_i(i = k - 1, k)$ is reversed in sign; i.e., if $\beta_k/\sigma < 0$, the coefficient of x_k from the table is positive; similarly if $\beta_{k-1}/\sigma < 0$, the coefficient of x_{k-1} from the table is positive no matter what the value of β_k/σ.)

Next let us consider the expected value of the estimator y^* when σ^2 is unknown. This case differs from the preceding only in that the test criteria for the various hypotheses are different. To test that β_i is zero we would compare b_i^2/v with λ, the $100\alpha\%$ point of the F distribution with one and $n - k - 1$ degrees of freedom (v is the mean square due to deviations from regression). We make the same assumptions concerning linearity of the model, normality of the errors, and orthogonality of the x's as were made in the case with σ^2 known. The same notation is adopted for the various estimators of y that might be used (i.e., y^* may be $y_k, y_{k-1}, \cdots$, or y_r) and for the events corresponding to the various results of the hypotheses tested. v, of course, is distributed as $[\sigma^2/(n - k - 1)]\chi^2_{n-k-1}$, and its distribution function will be called simply $f(v)$. The level of significance for each of the tests is again set at α. Equations (1), (2), (3) and (4) apply here as well as to the case above with σ^2 known. The difference between the cases becomes apparent when we try to evaluate the terms of the form $E(b_i \mid A_{k-1})P(A_{k-1})$.

The term $E(b_k \mid A_0)P(A_0)$ has previously been evaluated by Larson [13]. For the next term to be evaluated we have

$$E(b_{k-1} \mid A_1)P(A_1) = \iiint_R b_{k-1} N(b_{k-1}; \beta_{k-1}, \sigma^2)N(b_k; \beta_k, \sigma^2) \cdot f(v)\, db_{k-1}\, db_k\, dv,$$

where R is the region $b_k^2/\lambda < v < b_{k-1}^2/\lambda$, $|b_k| < |b_{k-1}|$. The last specification of R is implied by the first. The integration on v can be handled very easily with some simple restrictions on the sample size. After completing this integration, however, we are still left with the problem of evaluating a double, noncentral normal integral, where the limits of integration for one variable inescapably involve the second normal variable. So far it has not been found possible to express this double integral in closed form. The other integrals involved in the evaluation of Ey^* will include similar integration difficulties, so no closed form is presented for this case.

The integrals involved in Ey^* above could be evaluated by numerical methods, but such an undertaking would appear to be of a very complex nature.

2.2 *Mean square error of* y^*. The same notation introduced previously shall again be used here. The only new calculation involved in evaluating the mean square error of y^* is to compute $E(y^*)^2$. Here we shall be concerned only with the case σ^2 known, anticipating the fact that $E(y^*)^2$ will lead to the same integration difficulties encountered previously with σ^2 unknown.

We have

$$E(y^*)^2 = E(y_k^2 \mid A_0)P(A_0) + E(y_{k-1}^2 \mid A_1)P(A_1) + \cdots + E(y_r^2 \mid A_{k-r})P(A_{k-r}).$$

Expanding these terms we have

$$E(y_i^2 \mid A_{k-i})P(A_{k-i}) = \Big[\Big(\beta_0 + \sum_{j=1}^{i-1} \beta_j x_j\Big)^2 + \sigma^2\Big(1/n + \sum_{j=1}^{i-1} x_j^2\Big)$$

$$+ 2\Big(\beta_0 + \sum_{j=1}^{i-1} \beta_j x_j\Big) E(b_i \mid A_{k-i})x_i + E(b_i^2 \mid A_{k-i})x_i^2\Big] P(A_{k-i})$$

$$i = r+1, r+2, \cdots, k,$$

$$E(y_r^2 \mid A_{k-r})P(A_{k-r}) = \Big[\Big(\beta_0 + \sum_{j=1}^{r} \beta_j x_j\Big)^2 + \sigma^2\Big(1/n + \sum_{j=1}^{r} x_j^2\Big)\Big] P(A_{k-r}).$$

Thus the only new expectations to be evaluated are those of the appropriate squares of the regression coefficients.

Defining $G_i = 2\exp\{-\frac{1}{2}(\lambda + \beta_i^2/\sigma^2)\}\cosh(\beta_i(\lambda)^{\frac{1}{2}}/\sigma)$, $i = r+1, r+2, \cdots, k$, and applying integration by parts we find

$$E(b_k^2 \mid A_0)P(A_0) = \sigma^2(\lambda/2\pi)^{\frac{1}{2}}G_k + [\sigma/(2\pi)^{\frac{1}{2}}]\beta_k F_k + (\sigma^2 + \beta_k^2)H_k,$$

$$E(b_i^2 \mid A_{k-i})P(A_{k-i}) = \prod_{j=i+1}^{k} [1 - H_j]\{\sigma^2(\lambda/2\pi)^{\frac{1}{2}}G_i + [\sigma/(2\pi)^{\frac{1}{2}}]\beta_i F_i$$

$$+ (\sigma^2 + \beta_i^2)H_i\} \qquad i = r+1, r+2, \cdots, k-1,$$

where F_i and H_i are as defined in Section 2.1. Combining these terms and making use of the identities mentioned earlier we have

$$\frac{\text{mean square error } y^*}{\sigma^2} = \Big(\frac{1}{n} + \sum_{i=1}^{k} x_i^2\Big)$$

$$+ 2\sum_{i=r+1}^{k} x_i\Big\{\frac{\beta_i}{\sigma} - \frac{F_i}{(2\pi)^{\frac{1}{2}}(1 - H_i)}\Big\}\Big(\sum_{j=i}^{k} \frac{\beta_j}{\sigma} x_j\Big)\prod_{m=i}^{k} [1 - H_m]$$

$$+ \sum_{i=r+1}^{k} x_i^2\Big\{\Big(\frac{\lambda}{2\pi}\Big)^{\frac{1}{2}} G_i + \frac{1}{(2\pi)^{\frac{1}{2}}}\frac{\beta_i}{\sigma}F_i - \Big(\frac{\beta_i^2}{\sigma^2} + 1\Big)(1 - H_i)\Big\}\prod_{m=i+1}^{k} (1 - H_m).$$

The (mean square error $y^*)/\sigma^2$ was calculated from the above formula for the two parts of Example 1 in Section 2.1 and found to be $1/n + \sum_{i=1}^{r} x_i^2 + 2.698$, $1/n + \sum x_i^2 + 2.012$, respectively.

3. Procedure B. An experimenter again has available n measurements on $k + 1$ variables $(y, x_1, x_2, \cdots, x_k)$ and is interested in which of the k x-variables will be necessary in a linear model to predict values of y. He arbitrarily decides to test that the coefficient β_{r+1} of x_{r+1} is zero. If he accepts this hypothesis he simply uses $x_1, x_2, \cdots, x_r$ to predict y. If he rejects this hypothesis he tests that β_{r+2}, the coefficient of x_{r+2}, is zero. If he accepts this second hypothesis then he uses $x_1, x_2, \cdots, x_{r+1}$ to predict y. If he rejects this second hypothesis he tests the hypothesis that β_{r+3}, the coefficient of x_{r+3}, is zero, etc. He continues adding variables to his model in this way until he either finally accepts a hypothesis and stops testing, retaining all the variables he found to differ significantly from zero in addition to $x_1, x_2, \cdots, x_r$, or he finally reaches β_k and finds it to differ significantly from zero as well and uses all k variables to predict y. (The experimenter again is acting as though he had a priori information that ranks the variables in order of importance.) Mathematically we have the following (again the estimator used in any particular case is denoted by y^* and the subscript on y denotes the number of independent variables included in the estimator).

Event	$y^* =$	Situation
A_i	y_{r+i}	Accept $H_i : \beta_{r+i+1} = 0$; Reject $H_{i-1} : \beta_{r+i} = 0; \cdots$ Reject $H_0 : \beta_{r+1} = 0$, $i = 0, 1, 2, \cdots, k - r - 1$,
A_{k-r}	y_k	Reject $H_{k-r-1} : \beta_k = 0$; Reject $H_{k-r-2} : \beta_{k-1} = 0; \cdots$ Reject $H_0 : \beta_{r+1} = 0$.

Thus we have a problem of model specification; i.e., the particular model decided upon is the result of a series of results of testing hypotheses.

3.1 *Expected value of* y^*. The same assumptions concerning the linearity of the model, the distribution of the errors and the orthogonality of the x-vectors are made here as were made in Section 2.1. A slight alteration in the proof given in Section 4 illustrates that the expected value or bias of y^* for the non-orthogonal case can easily be derived from the orthogonal situation. The orthogonality assumption is sufficient for our tests of the β_i's $(i = r + 1, r + 2, \cdots, k)$ to be independent. We first look at the situation with σ^2 known. The test criterion for the hypothesis $\beta_i = 0$ is b_i^2/σ^2 $(i = r + 1, r + 2, \cdots, k)$. This quantity is compared with λ, the $100\alpha\%$ point of the χ_1^2 distribution. Then

$$Ey^* = E(y_r \mid A_0)P(A_0) + E(y_{r+1} \mid A_1)P(A_1) + \cdots + E(y_k \mid A_{k-r})P(A_{k-r}),$$

and

$$E(y_i \mid A_{i-r})P(A_{i-r})$$
$$= [\beta_0 + \sum_{j=1}^{r} \beta_j x_j + \sum_{m=r+1}^{i} x_m E(b_m \mid A_{i-r})]P(A_{i-r}), i = r + 1, r + 2, \cdots, k.$$

Adding these expectations together, we find

$$Ey^* = \beta_0 + \sum_{i=1}^{r} \beta_i x_i + x_{r+1} \sum_{j=1}^{k-r} E(b_{r+1} \mid A_j)P(A_j)$$
$$+ x_{r+2} \sum_{j=2}^{k-r} E(b_{r+2} \mid A_j)P(A_j) + \cdots + x_k E(b_k \mid A_{k-r})P(A_{k-r}).$$

Again b_i is distributed as $N(b_i ; \beta_i, \sigma^2)$, $i = 1, 2, \cdots, k$, so

$$\sum_{j=1}^{k-r} E(b_{r+1} \mid A_j)P(A_j) = \frac{\sigma}{(2\pi)^{\frac{1}{2}}} F_{r+1} + \beta_{r+1} H_{r+1},$$

$$\sum_{j=i-r}^{k-r} E(b_i \mid A_j)P(A_j) = \prod_{m=r+1}^{i-1} H_m \left[\frac{\sigma}{(2\pi)^{\frac{1}{2}}} F_i + \beta_i H_i\right],$$

$i = r + 2, r + 3, \cdots, k$, where F_i and H_i are as defined in Section 2.1. Thus

$$Ey^* = \beta_0 + \sum_{i=1}^{r} \beta_i x_i + \sum_{i=r+1}^{k} x_i \prod_{j=r+1}^{i-1} H_j \left[\frac{\sigma}{(2\pi)^{\frac{1}{2}}} F_i + \beta_i H_i\right],$$

where $\prod_{j=r+1}^{r} H_i \equiv 1$. Defining bias/$\sigma$ as in Section 2.1, we find

$$\text{bias}/\sigma = \sum_{i=r+1}^{k} x_i \left[\frac{1}{(2\pi)^{\frac{1}{2}}} F_i \prod_{j=r+1}^{i-1} H_j - \frac{\beta_i}{\sigma}\left(1 - \prod_{j=r+1}^{i} H_j\right)\right].$$

The coefficients of the x_i's in the bias/σ function are fairly easily tabulated and depend only on the values of the (β_i/σ)'s and the level of significance α, as they did with the preceding rule.

Although it is planned to include in the publication mentioned earlier more extensive tables to be used in finding the coefficients of the x_i's necessary to evaluate the bias/σ function for Procedure B, Table IV ($\alpha = .05$, $\lambda = 3.84$), Table V ($\alpha = .10$, $\lambda = 2.71$), Table VI ($\alpha = .32$, $\lambda = 1$), given in the appendix, provide a means of finding such coefficients for realistic values of β_i/σ, if we have only two "doubtful" predictors and for the selected values of the significance level α of the preliminary tests of significance. To evaluate the bias/σ function simply enter the desired Table with the assumed value of β_k/σ and β_{k-1}/σ and read off the coefficients of x_k and x_{k-1}. Then for whatever values of x_k and x_{k-1} we wish to predict y we need only substitute them in the bias/σ equation. Again, since the b_i's are normally distributed with mean β_i and variance σ^2, then a realistic range of β_i/σ is from -3 to $+3$.

EXAMPLE 2. Given a linear regression model with eight predictors of which x_7 and x_8 are "doubtful." Procedure B is to be used to decide whether or not to retain them, first testing β_7 and then β_8 (if necessary) to be zero at the $\alpha = .10$ significance level. Then no matter what the outcome of the test(s), we find the coefficients of $x_7 = x_{k-1}$ (the first predictor variable tested for retention) and $x_8 = x_k$ (the second one tested) from Table V. Suppose we wish to predict y with $x_7 = -1$ and $x_8 = 1$, assuming $\beta_7/\sigma = \beta_{k-1}/\sigma = 2$, $\beta_8/\sigma = \beta_k/\sigma = \frac{3}{2}$, then

$$\begin{aligned}\text{bias}/\sigma &= (-1)(-.345) + (1)(-.825)\\ &= -.480,\end{aligned}$$

while with $\beta_7/\sigma = \beta_{k-1}/\sigma = 3$, $\beta_8/\sigma = \beta_k/\sigma = \frac{1}{2}$ we have

$$\begin{aligned}\text{bias}/\sigma &= (-1)(-.101) + (1)(-.283)\\ &= -.182.\end{aligned}$$

(Note that, as for Procedure A, all tabular values are negative for positive values of the (β_i/σ)'s $(i = k - 1, k)$, for Procedure B. If we wish to evaluate the bias/σ function with one or both of the (β_i/σ)'s $(i = k - 1, k)$ negative, then the coefficient of the corresponding x_i $(i = k - 1, k)$ is reversed in sign; i.e., if $\beta_{k-1}/\sigma < 0$ the coefficient of x_{k-1} from the table is positive; similarly if $\beta_k/\sigma < 0$, the coefficient of x_k from the table is positive, no matter what the value of β_{k-1}/σ.)

In considering the Ey^* with σ^2 unknown the same difficulty is encountered here as was mentioned in Section 2.1. That is, the expectation can be expressed in terms of integrals but the integrals can not be expressed in closed form. These integrals again could be evaluated for particular cases with the aid of numerical integration, but the resulting tables might not justify the necessary expenditures.

3.2 *Mean square error of* y^*. The same notation introduced previously shall again be used here. The only new calculation involved in evaluating the mean square error of y^* is to compute $E(y^*)^2$. Here we shall be concerned only with the case σ^2 known, anticipating the fact that $E(y^*)^2$ will lead to the same integration difficulties encountered in trying to evaluate Ey^* with σ^2 unknown. We have

$$E(y^*)^2 = E(y_r^2 \mid A_0)P(A_0) + E(y_{r+1}^2 \mid A_1) + \cdots + E(y_k^2 \mid A_{k-r})P(A_{k-r}).$$

Expanding these terms we have

$$E(y_r^2 \mid A_0)P(A_0) = \left[\left(\beta_0 + \sum_{i=1}^{r} \beta_i x_i\right)^2 + \sigma^2\left(1/n + \sum_{i=1}^{r} x_i^2\right)\right] P(A_0)$$

$$\begin{aligned} E(y_i^2 \mid A_{i-r})P(A_{i-r}) = &\left\{\left(\beta_0 + \sum_{m=1}^{r} \beta_m x_m\right)^2 + \sigma^2\left(1/n + \sum_{m=1}^{r} x_m^2\right)\right. \\ &+ 2\left(\beta_0 + \sum_{m=1}^{r} \beta_m x_m\right) \sum_{j=r+1}^{i} x_j E(b_j \mid A_{i-r}) \\ &+ 2 \sum_{j=r+2}^{i} \sum_{\substack{m=r+1 \\ m<j}}^{i-1} x_j x_m E(b_j \mid A_{i-r}) E(b_m \mid A_{i-r}) \\ &\left. + \sum_{j=r+1}^{i} x_i^2 E(b_i^2 \mid A_{i-r})\right\} P(A_{i-r}) \qquad i = r+1, r+2, \cdots, k. \end{aligned}$$

Essentially, then, the only new integrations are those involving the expected value of the squares of the regression coefficients. These are evaluated exactly as was done in Section 2.2 and will not be repeated. Combining terms we find

$$\begin{aligned} \frac{\text{mean square error of } y^*}{\sigma^2} = &\left(\frac{1}{n} + \sum_{i=1}^{r} x_i^2\right) - \left(\sum_{i=r+1}^{k} \frac{\beta_i}{\sigma} x_i\right)^2 \\ &- 2\left(\sum_{m=r+1}^{k} \frac{\beta_m}{\sigma} x_m\right) \sum_{i=r+1}^{k} x_i \left\{\prod_{j=r+1}^{i} H_j \left[\frac{F_i}{(2\pi)^{\frac{1}{2}} H_i} + \frac{\beta_i}{\sigma}\right] - \frac{\beta_i}{\sigma}\right\} \\ &+ 2 \sum_{\substack{i=r+1 \\ i<j}}^{k-1} \sum_{j=r+2}^{k} x_i x_j \prod_{m=r+1}^{j} H_m \left[\frac{F_i}{(2\pi)^{\frac{1}{2}} H_i} + \frac{\beta_i}{\sigma}\right]\left[\frac{F_j}{(2\pi)^{\frac{1}{2}} H_j} + \frac{\beta_j}{\sigma}\right] \\ &+ \sum_{i=r+1}^{k} x_i^2 \prod_{j=r+1}^{i-1} H_j \left[\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} G_i + \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{\beta_i}{\sigma} F_i + \left(1 + \frac{\beta_i^2}{\sigma^2}\right) H_i\right]. \end{aligned}$$

The (mean square error)/σ^2 was calculated from the above formula for the two parts of Example 2 in Section 3.1 and found to be $1/n + \sum_{i=1}^{r} x_i^2 + 2.029$, $1/n + \sum_{i=1}^{r} x_i^2 + 1.950$, respectively.

4. Non-orthogonal Cases. All the derivations presented in Sections 2 and 3 made use of the assumption that the independent variables $x_1, x_2, \cdots, x_k$ are orthogonal. It is quite easy to prove, using essentially the algebra of the canonical form, that the bias and mean square error functions for both procedures are independent of the orthogonality assumption.

In a given non-orthogonal case, a reparametrization of the model, consisting of linear combinations of the values of $x_1, x_2, \cdots, x_k$ and the "inverse" linear combinations of the parameters $\beta_1, \beta_2, \cdots \beta_k$, will suffice to transform to a new set of x_i' variables $(x_1', x_2', \cdots, x_k')$ and a new set of parameters $(\beta_1', \beta_2', \cdots, \beta_k')$ such that the orthogonality assumption is satisfied. Then, for any particular values of the x_i's, say $x_{10}, x_{20}, \cdots, x_{k0}$, with such a transformation we have that $\sum x_{i0}\beta_i \equiv \sum x_{i0}'\beta_i'$, $\sum x_{i0}b_i \equiv \sum x_{i0}'b_i'$, where x_{i0}', β_i' and b_i' are the corresponding transformed values. Obviously, then, since the bias and mean square error functions depend on the x_i's, β_i's and b_i's only through such sums of products, both functions remain invariant under the orthogonalizing reparametrization.

5. Discussion of bias and mean square error functions of the two procedures.

5.1 *Evaluation of the bias functions.* We wish to use the results obtained in this study to determine how serious the bias is for Procedure A and Procedure B. Further, if possible, we would like to know whether the bias for either or both of the two procedures will be compensated for by a reduced prediction error.

Let δ_A and δ_B represent respectively the bias/σ function for Procedure A and Procedure B. As pointed out earlier the coefficients of the x_i's $(i = r + 1, r + 2, \cdots, k)$ in both δ_A and δ_B are functions only of the parameters β_i/σ, $(i = r + 1, r + 2, \cdots, k)$ and the significance level α (or λ, the equivalent $100\alpha\%$ point of the χ^2 distribution). Although the β_i/σ, $(i = r + 1, r + 2, \cdots, k)$ are generally unknown, as mentioned earlier, it is unlikely that these values will lie outside the range of -3 to $+3$. Hence, more extensive tables can be prepared of the magnitudes of the coefficients of the corresponding x_i's of δ_A and δ_B for $-3 \leqq \beta_i/\sigma \leqq +3$ and varying values of α or λ. Additional tables are planned for a later publication.

However, making use of the tabular values presented in this paper and those additional values in Larson [13] (the tables in Larson [13] present bias coefficients for 4 "doubtful" predictors, significance levels of $\alpha = .32$ $(\lambda = 1)$ and $\alpha = .05$ $(\lambda = 3.84)$ in addition to $\alpha = .10$ $(\lambda = 2.71)$, and $\beta_i/\sigma = \frac{1}{2}(\frac{1}{2})3$) it is found that all coefficients in both bias functions increase as α decreases from .32 to .05. The individual coefficients vary in different ways as their corresponding (β_i/σ)'s increase in value, but generally increasing values of the bias coefficients corresponds to β_i/σ assuming the values $\frac{1}{2}$, 1, $\frac{3}{2}$ or 2 with a decrease in the bias coefficient as its β_i/σ continues to $\frac{5}{2}$ and 3 in magnitude.

With $\alpha = .32$ or $.10$ and assuming up to 4 "doubtful" predictors the coefficients in δ_A are in every case smaller in magnitude than those for the corresponding x_i in δ_B (for the values of β_i/σ considered). With $\alpha = .05$, the relative sizes of the two seem to fluctuate but still the coefficients for δ_A are generally smaller than those for δ_B.

Additional information may be obtained from limiting values. If all β_i/σ $(i = r + 1, r + 2, \cdots, k)$ were actually zero, then δ_A and δ_B would be zero and the predictant y^* would be unbiased for Procedure A or Procedure B. If we set $\lambda = 0$, corresponding to always rejecting H_0 and thus always using all k predictors, we again find δ_A and δ_B to be zero. If we take the limit as $\lambda \to \infty$, corresponding to always accepting each hypothesis and thus always using only the first r independent variables, we would find δ_A and δ_B approaching $-\sum_{i=r+1}^{k} (\beta_i/\sigma)x_i$, the negative of the terms which would always be ignored (here the β_i are divided by σ to be consistent with the definitions of δ_A and δ_B).

Considering *only* the magnitude of the *bias* functions, and the particular tabular values considered for β_i/σ $(i = r + 1, r + 2, \cdots, k)$ and α, it would appear that Procedure A would be preferred to Procedure B for predictions involving the same set of x_i's of like sign. This appears in contradiction to the "practical" recommendation to use Procedure B always, since it may reduce the computing time, i.e., since Procedure B tests for *adding* additional predictors it provides positive information or terminates at any step beginning with the first.

5.2 *Evaluation of the mean square error functions.* Let γ_A^2 and γ_B^2 represent respectively the (mean square error)/σ^2 functions for Procedure A and Procedure B. Now, some information may be obtained from limiting values. For $\lambda = 0$ (corresponding to always rejecting all the hypotheses and hence always using all k predictor variables), γ_A^2 and γ_B^2 reduce to $(1/n + \sum_{i=1}^{k} x_i^2)$. On the other hand in the limit as $\lambda \to \infty$ (corresponding to always accepting all the hypotheses and thus always using only the first r predictors), γ_A^2 and γ_B^2 reduce to $(1/n + \sum_{i=1}^{r} x_i^2) + (\sum_{i=r+1}^{k} (\beta_i/\sigma)x_i)^2$.

As mentioned above, γ_A^2 for the two parts of Example 1 is equal to $1/n + \sum_{i=1}^{r} x_i^2 + 2.698$ and to $1/n + \sum_{i=1}^{r} x_i^2 + 2.012$, respectively, and γ_B^2, for the same two sets of parameters, is equal to $1/n + \sum_{i=1}^{r} x_i^2 + 2.029$ and to $1/n + \sum_{i=1}^{r} x_i^2 + 1.950$, respectively. Thus, for the first set of values we have $1 < \gamma_A^2/\gamma_B^2 < 1.33$ and for the second set $1 < \gamma_A^2/\gamma_B^2 < 1.03$, with the ratio being much more likely to be close to the lower end point rather than the upper, since the larger $\sum_{i=1}^{r} x_i^2$ becomes, the closer the ratio will be to 1. Similarly, if we compare γ_A^2 and γ_B^2 with the (mean square error)/σ^2 of y_k (that is, the usual estimator using all k x_i's for predictions), which for both examples quoted here would be $1/n + \sum_{i=1}^{r} x_i^2 + 2$, we see that both γ_A^2 and γ_B^2 differ negligibly from this quantity, for $\sum_{i=1}^{r} x_i^2$ even moderately large. It is not known, at this time, whether γ_A^2 will in all cases be larger than γ_B^2, but if these two calculated examples are representative of usual values, it would seem that any difference between the two is likely to be negligible, for $\sum_{i=1}^{r} x_i^2$ moderately large.

6. Need for consideration of the role of the analysis of incompletely specified models in the foundations of statistical inference. Birnbaum [5] states on page 274 of the recent paper "On the Foundations of Statistical Inference": "The adequacy of any such model (mathematical-statistical model) is typically supported, more or less adequately, by a complex informal synthesis of previous experimental evidence of various kinds and theoretical considerations concerning both subject-matter and experimental techniques." This statement is followed by: "We deliberately delimit and *idealize* the present discussion by considering only models whose adequacy is postulated and is not in question." Without going into the relative merits or demerits of the proposed philosophy of statistical inference contained in Professor Birnbaum's paper, it is suggested that any proposed *general* theory of statistical inference, useful in the applications of statistics, would need to include provision for statistical inferences based on the analysis of *incompletely specified models.* Here we are referring to statistical inferences based on the analysis of incompletely specified models in general, not just those based on such analysis applied to regression.

APPENDIX

TABLE I

Bias Coefficients for Procedure A. (α = *.05 for both tests,* λ = *3.84. All entries must be multiplied by* -1.)

Coefficient of x_k						
β_k/σ =	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
	.342	.583	.657	.570	.392	.215

Coefficient of x_{k-1}						
β_k/σ	β_{k-1}/σ					
	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
$\frac{1}{2}$	.315	.537	.605	.525	.361	.198
1		.484	.546	.473	.325	.179
$\frac{3}{2}$			.445	.386	.265	.146
2				.276	.190	.104
$\frac{5}{2}$					.115	.063
3						.032

TABLE II

Bias Coefficients for Procedure A. (α = .10 for both tests, λ = 2.71. All entries must be multiplied by -1.)

Coefficient of x_k

β_k/σ =	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
	.262	.426	.446	.345	.211	.101

Coefficient of x_{k-1}

β_k/σ	β_{-1}/σ					
	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
$\frac{1}{2}$	.225	.366	.383	.296	.181	.087
1		.315	.329	.254	.155	.075
$\frac{3}{2}$			.249	.193	.118	.057
2				.124	.076	.036
$\frac{5}{2}$					.041	.020
3						.009

TABLE III

Bias Coefficients for Procedure A. (α = .32 for both tests, λ = 1. *All entries must be multiplied by -1.)*

Coefficient of x_k

β_k/σ =	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
	.090	.132	.119	.077	.038	.015

Coefficient of x_{k-1}

β_k/σ	β_{k-1}/σ					
	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
$\frac{1}{2}$	.056	.083	.074	.048	.024	.009
1		.063	.057	.037	.018	.007
$\frac{3}{2}$			.036	.023	.012	.004
2				.012	.006	.002
$\frac{5}{2}$					.003	.001
3						.000

TABLE IV

Bias Coefficients for Procedure B. (α = .05 for both tests, λ = 3.84. All entries must be multiplied by −1.)

Coefficient of x_{k-1}

β_{k-1}/σ =	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
	.342	.583	.657	.570	.392	.215

Coefficient of x_k

β_{k-1}/σ	β_k/σ					
	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
$\frac{1}{2}$	.488					
1	.473	.929				
$\frac{3}{2}$	.449	.865	1.228			
2	.419	.785	1.065	1.262		
$\frac{5}{2}$	.389	.706	.906	.991	1.013	
3	.366	.645	.783	.783	.706	.631

TABLE V

Bias Coefficients for Procedure B. (α = .10 for both tests, λ = 2.71. All entries must be multiplied by −.1)

Coefficient of x_{k-1}

β_{k-1}/σ =	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
	.262	.426	.446	.345	.211	.101

Coefficient of x_k

β_{k-1}/σ	β_k/σ					
	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
$\frac{1}{2}$	.467					
1	.438	.850				
$\frac{3}{2}$	.395	.747	1.035			
2	.348	.632	.825	.940		
$\frac{5}{2}$	.309	.538	.652	.667	.657	
3	.283	.476	.538	.489	.409	.353

TABLE VI

Bias Coefficients for Procedure B. (α = .32 for both tests, λ = 1. All entries must be multiplied by −1.)

Coefficient of x_{k-1}

β_{k-1}/σ =	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
	.090	.132	.119	.077	.038	.015

Coefficient of x_k

β_{k-1}/σ	β_k/σ					
	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
$\frac{1}{2}$	.346					
1	.286	.546				
$\frac{3}{2}$	.214	.395	.536			
2	.154	.269	.336	.380		
$\frac{5}{2}$	.117	.190	.211	.206	.203	
3	.099	.152	.150	.121	.094	.083

REFERENCES

[1] BANCROFT, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Ann. Math. Statist.* **15** 190–204.

[2] BECHHOFER, R. E. (1951). The effect of preliminary tests of significance on the size and power of certain tests of univariate linear hypotheses. Unpublished Ph.D. thesis, Columbia Univ.

[3] BENNETT, B. M. (1952). Estimation of means on the basis of preliminary tests of significance. *Ann. Inst. Statist. Math* **4** 31–43.

[4] BENNETT, B. M. (1956). On the use of preliminary tests in certain statistical procedures. *Ann. Inst. Statist. Math.* **8** 45–47.

[5] BIRNBAUM, ALLAN (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 269–326.

[6] BOZIVICH, HELEN, BANCROFT, T. A. and HARTLEY, H. O. (1956). Power of analysis of variance test procedures for certain incompletely specified models, I. *Ann. Math. Statist.* **27** 1017–1043.

[7] FIREMAN, M. and WADLEIGH, C. H. (1951). A statistical study on the relation between pH and the exchangeable sodium percentage of western soils. *Soil Sci.* **71** 273–285.

[8] FISHER, R. A. (1922). Mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London. Ser. A* **222** 309–368.

[9] HOLLINGSWORTH, T. H. (1959). Using an electronic computer in a problem of medical diagnosis. *J. Roy. Statist. Soc. Ser. A* **122** 212–223.

[10] HUNTSBERGER, D. V. (1955). A generalization of a preliminary testing procedure for pooling data. *Ann. Math. Statist.* **26** 734–743.

[11] KITAGAWA, TOSIO (1951). Successive process of statistical inference. *Mem. Fac. Sci. Kyusyu Univ. Ser. A* **6** 55–95.

[12] KITAGAWA, TOSIO (1959). Successive process of statistical inference applied to linear regression analysis and its specialization to response surface analysis. *Bull. Math. Statist.* **8** 80–114.

[13] LARSON, HAROLD J. (1957). Biases in prediction by regression for certain incompletely specified models. Unpublished M.S. thesis, Iowa State Univ.

[14] LARSON, HAROLD J. (1960). Sequential model building for prediction in regression analysis. Unpublished Ph.D. thesis, Iowa State Univ.

[15] MCCULLOUGH, ROGER (1961). Testing equality of means under variance heterogeneity. Unpublished Ph.D. thesis, Iowa State Univ.

[16] MOSTELLER, FREDERICK (1948). On pooling data. *J. Amer. Statist. Assoc.* **43** 231–242.

[17] PAULL, A. E. (1950). On a preliminary test for pooling mean squares in the analysis of variance. *Ann. Math. Statist.* **21** 539–556.

[18] PEPERZAK, PAUL (1956). Correlation of selected soil indices with plant growth on highway backslopes. Iowa Highway Research Board, Bull. No. 10.

[19] SCHULTZ, E. FRED, JR., and GOGGANS, JAMES F. (1961). A systematic procedure for determining potent independent variables in multiple regression and discriminant analysis. Agr. Exp. Sta. Bull. 336, Auburn University.

[20] SUMMERFIELD, A. and LUBIN, A. (1951). A square root method of selecting a set of variables in multiple regression: I. The method. *Psychometrika* **16** 271–284.