TECHNOMETRICS©, VOL. 16, NO. 1, FEBRUARY 1974

The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction

David M. Allen

Department of Statistics
University of Kentucky
(Lexington, Kentucky)

We show that data augmentation provides a rather general formulation for the study of biased prediction techniques using multiple linear regression. Variable selection is a limiting case, and Ridge regression is a special case of data augmentation. We propose a way to obtain predictors given a credible criterion of good prediction.

KEY WORDS

Augmentation
Variable Selection
Prediction
Regression

1. INTRODUCTION

In exploratory experiments researchers often collect data on a response variable and several potential predictor variables. Additional predictor variables are frequently created by taking functions of the observed predictor variables. The traditional method of obtaining a prediction equation is multiple linear regression. The hazard of using too many predictor variables is widely known. The reason for the problem is that the addition of a variable to a least squares prediction equation almost always increases (never decreases) the variance of a predicted response [1, 9]. Addition of a variable may decrease the squared bias, but this decrease is often small relative to the increase in variance.

Nearly all of the proposed solutions to this problem can be classified into one of two categories:

(i) Variable selection techniques.
(ii) Augmentation of the variable matrix by dummy observations.

Our first objective is to show that (i) is a limiting case of (ii), and thus (ii) provides a more general framework for study. Secondly we propose a way to obtain predictors given a credible criterion of good prediction.

2. THE RELATIONSHIP BETWEEN VARIABLE SELECTION AND DATA AUGMENTATION

The classical multiple linear regression model is

$$\underset{(n\times 1)}{Y} = \underset{(n\times r)}{X}\ \underset{(r\times 1)}{\beta} + \underset{(n\times 1)}{\epsilon} \qquad (1)$$

Received Nov. 1972; revised June 1973

where Y is a vector of responses; X is a known, full rank matrix of nonstochastic variables; β is the unknown weight vector corresponding to X; and ϵ is a normally distributed vector of random variables having expected value 0 and dispersion matrix $I\sigma^2$.

We first review a technique for estimating β subject to linear restrictions. Let P be a $(p \times r)$ matrix of full rank. The value of β minimizing $(Y - X\beta)'(Y - X\beta)$ subject to $P\beta = 0$ is

$$b - (X'X)^{-1}P'(P(X'X)^{-1}P')^{-1}Pb \qquad (2)$$

where $b = (X'X)^{-1}X'Y$. This can be shown by a straight-forward application of Lagrange's multipliers.

We now consider the effect of augmenting X and Y with dummy observations. The generalized least squares estimate of β based on the augmented matrices

$$\begin{bmatrix} X \\ \hline P \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} Y \\ \hline O \end{bmatrix}$$

is

$$(X'X + P'WP)^{-1}X'Y \qquad (3)$$

where $W = \text{Diag}\ (w_1 w_2 \cdots w_p)$ and $w_i > 0$, $i = 1, 2, \cdots, p$. In this formulation, $P\beta$ represents a set of linear combinations of the elements of β that we have prior belief is null. The matrix W represents the strength of this belief. Specifically, w_i represents the number of observations, having expected value equal to the i-th element of $P\beta$, that we believe necessary to provide information equal to our prior information. If Y is normally distributed, then (3) is the Bayesian estimator of β using a prior distribution of $P\beta$ given σ^2 of $N_r(0, W^{-1}\sigma^2)$. An expression algebraically identical to (3) is

$$b - (X'X)^{-1}P'(W^{-1} + P(X'X)^{-1}P')^{-1}Pb. \qquad (4)$$

Note that (4) approaches (2) as $\min\ (w_1 w_2 \cdots w_p)$ approaches infinity.

Throughout the remainder of this paper we will consider P to have a special structure. Let every row of P contain one "1" and all other elements be "0". Estimator (2) is equivalent to deleting variables corresponding to columns of P that contain a 1; calculating least squares estimators of the remaining coefficients; and then inserting zeros in the positions of the deleted variables. Since (2) represents a prediction equation with deleted variables and is a limiting form of (4), variable selection is said to be a limiting form of data augmentation.

2. Choosing A Predictor

For our special P the matrix $P'WP$ in (3) is diagonal and will be denoted by D. Thus we will consider estimators of the form

$$\hat{\beta} = (X'X + D)^{-1}X'Y \tag{5}$$

where $D = \text{diag}\ (d_1 d_2 \cdots d_r)$ and $d_j \geq 0$, $j = 1, 2, \cdots, r$. For any fixed $D \neq 0$, predictors using $\hat{\beta}$ have variances no larger than predictors using $\underline{b}$.

Hoerl and Kennard [5, 6] present a technique they call ridge regression. Ridge regression is a way to determine D in (5). They require that D be proportional to diag $(X'X)$ and use subjective rules for determining the constant of proportionality.

The procedure we propose depends upon having a credible criterion of good prediction. By credible criterion we mean a criterion that is made small when the predicted values are close to the observed values but also includes a penalty for increasing the variance of the predictors. It is not the purpose of this paper to compare criteria. However, for the purpose of illustration we will describe two credible criteria.

The total mean square error of prediction is

$$n\sigma^2 + \sum_{i=1}^{n} \text{Var}\ (\hat{Y}_i) + \sum_{i=1}^{n} (E(Y_i - \hat{Y}_i))^2. \tag{6}$$

If $\hat{Y}_i = x_i'\hat{\beta}$ where x_i' is the i-th row of X and $\hat{\beta}$ is as given by (5) with fixed D, then (6) is

$$n\sigma^2 + \text{tr}[(X'X + D)^{-1}X'X]^2\sigma^2 + \beta'X'(I - X(X'X + D)^{-1}X')^2X\beta. \tag{7}$$

It would be desirable to choose D such that (7) is minimum. However since β and σ^2 are unknown, an estimator of (7) can be used as a criterion. The quantity

$$M(D) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + 2\,\text{tr}\,(I + (X'X)^{-1}\,D)^{-1}S^2, \tag{8}$$

where S^2 is an estimator of σ^2, is an estimator of (7). The criterion (8) can be considered an extension of Mallow's Cp Statistic [4, 7, 8] to a more general class of estimators.

Another credible criterion is the Prediction Sum of Squares (PRESS). This criterion is defined by

$$\text{PRESS}\ (D) = \sum_{i=1}^{n} (Y - \hat{Y}_{(i)})^2$$

where $\hat{Y}_{(i)}$ is the estimator of $E(Y_i)$ *excluding* the i-th observation. Explicitly, $\hat{Y}_{(i)} = (X'X - x_ix_i' + D)^{-1}(X'Y - x_iY_i)$. In other words, each observation is "predicted" using the other $n - 1$ observations. The resulting "errors of prediction" are squared and summed to form PRESS. PRESS is appealing because it simulates prediction. It does not use an observation to aid in the "prediction" of itself. PRESS as defined here is an extension of PRESS defined in [2] to a more general class of predictors.

The proposed procedure is to choose a credible criterion and then find the value of D such that the criterion is minimum. The coefficients of the Xs in the prediction equation are given by (5) evaluated at this value of D.

4. Discussion

A local minimum of the criteria presented here can be found by a standard nonlinear regression technique. We recommend that the user apply this technique several times using initial values of D corresponding to different subsets having technical appeal. Indeed, our experience has shown the existence of distinct prediction equations having nearly equal values of the criterion. A discussion of computation and a numerical example are presented in [3].

The usefulness of the proposed procedure depends upon the statistical properties of predictors associated with the criterion used. While little is known about these properties, the Cp statistic has apparently gained wide acceptance.

References

[1] Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics 13*, 469–75.

[2] Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting predictor variables. Univ. of Ky. Dept. of Statistics, *Tech. Report 23*.

[3] Allen, D. M. (1972). Biased prediction using multiple linear regression. Univ. of Ky. Dept. of Statistics, *Tech. Report 36*.

[4] Gorman, J. W. and Toman, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics, 8* 27–51.

[5] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression. Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

[6] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression. Applications to nonorthogonal problems. *Technometrics 12*, 69–82.

[7] MALLOWS, C. L. (1964). Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.

[8] MALLOWS, C. L. (1966). Choosing a subset regression. Presented at the Joint Statistical Meeting, Los Angeles, Calif.

[9] WALLS, R. C. and WEEKS, D. L. (1969). A note on the variance of a predicted response in regression. *Amer. Statistician* 23(3):24–26.

The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means

Morton B. Brown* and Alan B. Forsythe

Department of Biomathematics†
University of California at Los Angeles
Los Angeles, California

Four statistics which may be used to test the equality of population means are compared with respect to their robustness under heteroscedasticity, their power, and the overlap of their critical regions. The four are: the ANOVA F-statistic; a modified F which has the same numerator as the ANOVA but an altered denominator; and two similar statistics proposed by Welch and James which differ primarily in their approximations for their critical values.

The critical values proposed by Welch are a better approximation for small sample sizes than that proposed by James. Both Welch's statistic and the modified F are robust under the inequality of variances. The choice between them depends upon the magnitude of the means and their standard errors. When the population variances are equal, the critical region of the modified F more closely approximates that of the ANOVA than does Welch's.

KEY WORDS

Robust
ANOVA
Heteroscedasticity
Equality of Means

1. INTRODUCTION

The testing of the equality of several means from independent samples is a common statistical problem. Unfortunately, the one-way analysis of variance (ANOVA) F statistic yields a test that is sensitive to a lack of homogeneity of within group variances. That is, the actual size of a test may differ greatly from the nominal or selected size when the groups have different underlying population variances [4].

For possibly unequal population variances, Welch [7] and James [2] extended the Welch [6] asymptotic confidence interval approach for the two-means problem to more than two groups. Although deriving statistics which are quite similar, Welch approximates the distribution of his statistic by that of the Snedecor F and James by that of the chi-square. An alternative approach is to change the denominator of the usual ANOVA F statistic so that it has the same expected value as the numerator when the population means are equal. Approximate critical values may be found for this statistic by using the Satterthwaite [3] method of combining degrees of freedom.

Except for the two-means problem [7], the behavior of the above statistics have not been studied for small group sizes. Using a Monte Carlo sampling study, the size and power of these statistics are compared for varying sample sizes and population variances. In the case of homogeneity of variances, the similarity of the critical regions of these statistics are compared to that of the ANOVA F statistic.

2. NOTATION

Let x_{ij} represent the jth observation in the ith group, where $j = 1, \cdots, n_i$ and $i = 1, \cdots, g$. The x_{ij} are independent normal variates with expected value μ_i and variance σ_i^2.

The four statistics considered are

1) ANOVA

$$F = \frac{\sum_i n_i(x_{i.} - x_{..})^2/(g-1)}{\sum_i (n_i - 1)s_i^2/(N-g)}$$

where

$$N = \sum_i n_i$$

$$x_{i.} = \sum_j x_{ij}/n_i$$

$$x_{..} = \sum_i \sum_j x_{ij}/N = \sum_i n_i x_{i.}/N$$

$$s_i^2 = \sum_j (x_{ij} - x_{i.})^2/(n_i - 1)$$

* Permanent address: Department of Statistics, Tel-Aviv University, Tel-Aviv, Israel.

† This research was supported in part by National Institutes of Health grant RR-3.

Received March 1973; revised May 1973.

When all population means and variances are equal, this statistic follows the F distribution with $g - 1$ and $N - g$ degrees of freedom.

2) Welch [7]

$$W = \frac{\sum_i w_i(x_{i.} - \bar{x}_{..})^2/(g-1)}{\left[1 + \frac{2(g-2)}{(g^2-1)} \sum_i (1 - w_i/u)^2/(n_i - 1)\right]}$$

where

$$w_i = n_i/s_i^2$$

$$u = \sum_i w_i$$

$$\bar{x}_{..} = \sum w_i x_{i.}/u$$

When all population means are equal (even if the variances are unequal), W is approximately distributed as an F statistic with $g - 1$ and f degrees of freedom where f is implicitly defined as:

$$1/f = (3/(g^2 - 1)) \sum_i (1 - w_i/u)^2/(n_i - 1)$$

3) James [2]

$$J = \sum_i w_i(x_{i.} - \bar{x}_{..})^2/(g-1)$$

If all population means are equal

$$\Pr\left\{J > \chi^2\left[1 + \frac{3\chi^2 + (g+1)}{2(g^2-1)} \cdot \sum_i (1 - w_i/u)^2/(n_i - 1)\right]\right\} = \alpha$$

where χ^2 is the $1 - \alpha$ percentile from the χ^2 distribution on $g - 1$ degrees of freedom.

4) The numerator of the ANOVA F weighs means by n_i while both W and J use n_i/s_i^2. Our last statistic uses n_i as weights but substitutes a denominator which has expectation equal to the numerator when all means are equal. That is,

$$F^* = \frac{\sum_i n_i(x_{i.} - x_{..})^2}{\sum_i (1 - n_i/N)s_i^2}$$

Critical values are obtained from the F distribution $g - 1$ and f degrees of freedom where f is implicitly defined by the Satterthwaite [3] approximation

$$\frac{1}{f} = \sum_i c_i^2/(n_i - 1)$$

and

$$c_i = (1 - n_i/N)s_i^2/[\sum_i (1 - n_i/N)s_i^2].$$

When there are only two groups both F^* and W reduce to what is known as the Welch approximate degrees of freedom approach to the Behrens–Fisher problems which Wang [5] has shown to be adequate for that case.

3. Design of the Monte Carlo Sampling Experiment

Each condition of the experiment is specified by the sample sizes (n_i), population means (μ_i), and population variances (σ_i^2) for the g groups ($i = 1, \cdots, g$). For each condition a set of $N = \sum_i n_i$ observations from a standard normal distribution are generated using the Box–Miller [1] technique and assigned to consecutive groups. These values

Table 1—Size

Number of Groups, g	Sample Sizes n_i, i=1, ..., g	Standard Deviations σ_i, i=1, ..., g	F Nominal Size 10%	F 5%	F 1%	F* Nominal Size 10%	F* 5%	F* 1%	W Nominal Size 10%	W 5%	W 1%	J Nominal Size 10%	J 5%	J 1%
4	4,4,4,4	1,1,1,1	10.2	4.9	.9	7.8	3.4	.5	9.6	4.5	.8	13.3	7.9	2.4
		1,2,2,3	12.0	6.7	1.7	8.9	4.1	.7	10.3	4.7	.8	13.8	8.1	2.7
4	4,8,10,12	1,1,1,1	9.9	5.1	1.1	9.5	4.8	1.0	10.8	5.7	1.6	12.1	6.7	2.1
		1,2,2,3	5.9	3.0	.6	10.3	5.7	1.4	9.8	4.9	.9	10.8	5.6	1.3
		3,2,2,1	21.9	14.4	5.6	11.0	6.2	1.8	11.3	6.5	2.0	12.9	7.7	2.9
4	11,11,11,11	1,1,1,1	10.1	5.1	1.0	9.8	4.9	.9	10.0	5.1	.9	10.6	5.5	1.1
		1,2,2,3	11.4	6.3	1.8	10.7	5.7	1.5	10.1	5.0	1.1	10.6	5.4	1.3
4	11,16,16,21	1,1,1,1	10.3	4.9	1.1	10.3	5.1	1.0	10.2	5.0	1.0	10.5	5.3	1.2
		3,2,2,1	17.3	10.8	3.9	11.1	6.2	1.8	10.5	5.5	1.2	10.9	5.8	1.3
		1,2,2,3	7.3	4.0	1.0	11.5	6.5	1.8	10.6	5.4	1.1	10.9	5.6	1.1
6	4,4,4,4,4,4	1,1,1,1,1,1	9.6	4.9	1.0	7.3	3.4	.4	11.4	6.1	1.4	14.7	9.5	3.8
		1,1,2,2,3,3	13.5	8.3	2.5	9.5	4.9	1.0	13.0	7.0	1.8	16.9	10.9	4.3
6	4,6,6,8,10,12	1,1,1,1,1,1	9.9	4.6	1.1	9.2	4.5	.9	11.2	6.1	1.6	12.6	7.4	2.4
		1,1,2,2,3,3	5.8	3.1	.8	11.5	6.5	1.9	11.2	6.2	1.4	12.4	7.5	2.1
		3,3,2,2,1,1	24.8	17.0	7.0	10.5	5.8	1.5	12.3	6.8	2.2	13.8	8.4	3.1
6	6,6,6,6,6,6	1,1,2,2,3,3	12.2	7.1	2.2	9.9	5.2	1.2	11.1	5.7	1.4	12.8	7.3	2.3
6	11,11,11,11,11,11	1,1,2,2,3,3	12.4	7.3	2.4	11.5	6.5	1.9	10.8	5.7	1.4	11.3	6.2	1.6
6	16,16,16,16,16,16	1,1,2,2,3,3	12.4	7.2	2.2	12.0	6.8	1.9	10.2	5.1	1.1	10.4	5.2	1.2
6	21,21,21,21,21,21	1,1,2,2,3,3	12.1	6.9	2.0	11.6	6.5	1.7	9.5	4.8	1.0	9.5	4.9	1.1
10	20,20,20,20,20,20, 20,20,20,20	1,1,1.5,1.5,2,2, 2.5,2.5,3,3	12.6	7.1	2.3	12.1	6.6	2.1	10.5	5.2	1.2	10.7	5.3	1.2

The percent of rejections at the nominal size α by the 4 test statistics after sampling 10,000 independent replicates of g groups with n_i, $i = 1, \cdots, g$ observations from the Gaussian distribution $N(0, \sigma_i^2)$.

are then multiplied by σ_i and added to μ_i to obtain samples from the specified populations. All 4 statistics are then calculated and compared with their nominal tabulated 90th, 95th, and 99th percentiles. The above was repeated 10,000 times for each condition using a different starting value for the random number generator.

4. Results and Discussion

The results are presented in three tables giving size of each test, power of some tests for selected alternatives, and an indication of the overlap of critical regions.

The empirical size of the four tests are shown in Table 1 at the assumed 10%, 5% and 1% levels of significance. As is well known, the ANOVA F statistic shows marked deviations from its nominal size when the variances of the groups are unequal. In the examples given at the 5% size, its empirical size varies from 17% when the smaller samples have the larger variances to 3% in the opposite situation. The other three statistics show relatively little fluctuation in size. The statistics J and W have similar numerators, differing primarily in their definition of critical values. For small sample sizes, J deviates more widely than from the nominal size, rejecting the null hypothesis too often. The modification of F, F^*, varies in size slightly more than W. For groups with more than 10 subjects, the difference between the nominal and empirical sizes of both W and F^* is small.

Power results for 4 groups of 11, 16, 16 and 21 observations are presented in Table 2. The ANOVA F is shown only when the variances are equal. The statistic J is omitted since W gives a better approximation to the critical values. When the variances are equal, both W and F^* are close in power to the ANOVA F. When extreme means have small variances, W is more powerful than F^*. This is readily explained by their difference in weighting means, i.e., W by n_i/s_i^2 and F^* by n_i . Therefore, extreme means with small variances tend to increase W more than F^* and conversely for extreme means with large variances.

Table 3 presents a comparison of the 5% critical regions of W and F^* to F when the variances are equal. The per cent disagreements indicate disparities in their critical regions since one test is accepting the data as consistent with equal population means and the other is not. The critical regions defined by F and F^* are more coincident since they agree more often with each other than does W. Since the ANOVA F defines this region by a likelihood argument, this means that W deviates from this principle.

5. Conclusions

Considering the small loss in power in using F^* or W rather than F even when the population variances are equal, the protection against F's distortion in size recommend the use of F^* or W. The asymptotic approximations of the critical values of F^* and W are valid when each group has at least 10 observations and are not unreasonable for W down to 5 observations. The choice between W and F^* depends upon whether extreme means are thought to have extreme variance (W) or not (F^*). Considering the different weighting methods, one would also use F^* if some sample variances appear unusually low.

Table 2—Power

Variances	Means	F		F*		W	
		Nominal size		Nominal size		Nominal size	
		5%	1%	5%	1%	5%	1%
1, 1, 1, 1	0, 0, 0, 0	4.9	1.1	5.1	1.0	5.0	1.0
	1, 0, 0, 0	68.6	43.2	67.6	42.1	65.0	38.3
	0, 0, 0, .7	55.3	31.1	54.4	30.1	52.3	26.3
	5, 0, 0, .5	33.6	14.3	33.3	13.7	31.8	12.3
3, 2, 2, 1	0, 0, 0, 0			6.2	1.8	5.5	1.2
	1.5, 0, 0, 0			33.2	16.2	22.2	8.2
	0, 0, 0, 1			22.7	7.6	47.8	22.0
	1.3, 0, 0, 1.3			42.4	19.1	68.2	40.0
1, 2, 2, 3	0, 0, 0, 0			6.5	1.8	5.4	1.1
	1.3, 0, 0, 0			29.1	9.5	65.5	39.4
	0, 0, 0, 1			27.3	13.0	17.5	5.5
	1, 0, 0, 1			29.8	13.4	40.0	17.8

The percent of rejections at the 5% and 1% nominal size after sampling 10,000 independent replicates of 4 groups with 11, 16, 16, and 21 observations from Gaussian distributions.

TABLE 3—Critical Regions

Means	Rejected by F only	Rejected by both F and W	Rejected by W only	Rejected by F only	Rejected by both F and F*	Rejected by F* only	Rejected by F* only	Rejected by both F* and W	Rejected by W only
0, 0, 0, 0	1.34	3.57	1.43	.42	4.49	.59	1.39	3.69	1.31
1, 0, 0, 0	7.48	61.07	3.91	1.95	66.60	1.03	5.88	61.75	3.23
0, 0, 0, .7	5.16	50.15	2.15	1.87	53.44	1.00	5.31	49.13	3.17
.5, 0, 0, .5	4.40	29.22	2.56	1.48	32.14	1.20	4.23	29.11	2.67

The percent of rejections at the nominal 5% size after sampling 10,000 independent replicates of 4 groups with 11, 16, 16, and 21 observations all from Gaussian distributions with unit variances.

6. SUGGESTIONS FOR FUTURE WORK

This paper shows that further research in this area is needed. One might attempt to:

1) improve the degrees of freedom approximation for W and F^* so that can be used with smaller sample sizes;
2) consider a Winsorizing technique on the sample variances to decrease the chance of a false rejection caused by an unusually small s_i^2 compared to σ_i^2;
3) develop a statistic between W and F^* which is reasonably powerful no matter which means have small variances.

REFERENCES

[1] BOX, G. E. P. and MULLER, M. A. (1958). A Note on the Generation of Random Normal Deviates. *Ann. Math. Statist. 29*, 610–613.

[2] JAMES, G. S. (1951). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika 38*, 19–43.

[3] SATTERTHWAITE, F. E. (1941). Synthesis of variance. *Psychometrika 6*, 309–316.

[4] SCHEFFE, H. (1959). *The Analysis of Variance.* John Wiley & Sons, Inc., New York, 351–358.

[5] WANG, Y. Y. (1971). Probabilities of the Type I Errors of the Welch Tests for the Behrens-Fisher Problem. *J. Amer. Statist. Assoc. 66*, 605–608.

[6] WELCH, B. L. (1947). The generalization of 'Students' problem when several different population variances are involved. *Biometrika 34*, 28–35.

[7] WELCH, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika 38*, 330–336.

Exact Tests for Zero Drift Based on First Passage Times in Brownian Motion

V. Seshadri
McGill University
Montreal, Canada

and

J. J. Shuster
University of Florida
Gainesville, Florida

In this paper, we present a method of obtaining the exact critical region, based on a UMP test statistic of Nádas (1973), for testing for zero drift in a Brownian Motion process, where the data come from an Inverse Gaussian (first-passage) distribution. The alternative hypothesis is that the drift is positive.

KEY WORDS

Brownian Motion
Inverse Gaussian
Drift

1. INTRODUCTION AND SUMMARY

Let W_t be a separable Brownian Motion process with $W_0 = 0$, drift δ, and diffusion constant β^2. It is well known (see for example Tweedie (1957), Shuster and Miura (1972) or Cox and Miller (1965, p. 210)) that the first passage time, X, to the state $A > 0$ has the inverse Gaussian distribution, whose *pdf* in the notation of Tweedie (1957) is:

$$f(x; \mu, \lambda) = (\lambda/(2\pi x^3))^{1/2} \exp[-\tfrac{1}{2}\lambda\mu^{-2}x^{-1}(x-\mu)^2] \quad x > 0. \tag{1}$$

The parameters have the following interpretation:

$$\mu = \text{mean first passage time} = A/\delta, \tag{2}$$

$$\lambda = A^2/\beta^2. \tag{3}$$

Note that λ does not involve the drift δ. If $X_1, \cdots, X_n$ is a random sample from $f(x; \mu, \lambda)$, Tweedie (1957) proved:

$$\bar{X} = \sum X_i/n \text{ has } pdf \quad f(x; \mu, n\lambda), \tag{4}$$

$\bar{X}$ is independent of

$$Y = \sum_{i=1}^{n} (X_i^{-1} - \bar{X}^{-1}), \tag{5}$$

and

$$\lambda Y \sim \chi_{n-1}^2. \tag{6}$$

Shuster (1968) proved:

$$n\lambda\mu^{-2}\bar{X}^{-1}(\bar{X}-\mu)^2 \sim \chi_1^2. \tag{7}$$

Hence:

$$F = n(n-1)\mu^{-2}\bar{X}^{-1}Y^{-1}(\bar{X}-\mu)^2 \sim F_{1,n-1}. \tag{8}$$

Received March 1973; revised May 1973.

(8) can be used to find exact confidence intervals for μ:

$$P[\mu^2(1 - [n(n-1)]^{-1}\bar{X}YG) - 2\bar{X}\mu + \bar{X}^2 < 0] = 1 - \alpha, \tag{9}$$

where

$$G = F_{1,n-1,\alpha}. \tag{10}$$

Note that $\mu = +\infty$ iff the drift is zero (or negative). The confidence interval is of the form (c, ∞) iff:

$$[n(n-1)]^{-1}\bar{X}Y > G^{-1}. \tag{11}$$

The test statistic (LHS of (11)) is easily seen to be a linear function of that of equation (5) of Nádas (1973). The acceptance region is given by (11). The critical region is therefore of the same type as Nádas (1973), (which is UMP), except that we shall show: (exactly)

$$P[[n(n-1)]^{-1}\bar{X}Y \leq G^{-1} \mid H_0] = \alpha. \tag{12}$$

The above follows, by noting that if the drift is zero, then $X_1, \cdots, X_n$ have *pdf*: (see Karlin, (1966, p. 279))

$$g(x; a) = a(2\pi)^{-1/2}x^{-3/2} \exp[-\tfrac{1}{2}a^2x^{-1}] \quad x > 0, \tag{13}$$

where

$$a = A/\beta. \tag{14}$$

$\sum X_i$ is the first passage time of a separable Brownian Motion process, $W(t)$, (with $W(0) = 0$, drift $\delta = 0$, and diffusion β^2), to the state nA. Hence $\bar{X}$ has *pdf*:

$$g(x; n^{1/2}a). \tag{15}$$

(15) and (13) yield:

$$na^2/\bar{X} \sim \chi_1^2. \tag{16}$$

By straightforward analysis, the conditional *pdf* of $X_2, \cdots, X_n$ given $\bar{X}$ can be shown to be: $(x_i > 0, x_2 + \cdots + x_n < \bar{x}:)$

$$n^{1/2}(2\pi)^{-(n-1)/2}a^{n-1}[\bar{x}/(x_1x_2 \cdots x_n)]^{3/2} \exp[-\tfrac{1}{2}a^2Y], \quad (17)$$

where Y is given by (5). (The conditional *pdf* is zero elsewhere.)

Since the support of the conditional *pdf* (17) does not depend on a, and integrates to unity for all a, the conditional moment generating function of a^2Y given $\bar{X}$ is easily seen to be:

$$(1 - 2t)^{-(n-1)/2} \qquad t < \tfrac{1}{2}, \quad (18)$$

and hence: $\bar{X}$ and Y are independent and

$$a^2Y \sim \chi_{n-1}^2. \quad (19)$$

Equation (12), under the hypothesis of zero drift is therefore immediate from (16), (18), and (19). (Recall that G is defined in equation (10)).

2. Evaluation of Exact Entries in Table 1 of Nádas (1973)

In this section we shall express the rejection region in the notation of Nádas (1973, equations (4), (5)):

$$MH^{-1} < nc(n, \alpha) = (n - 1)G^{-1} + 1, \quad (20)$$

where

$$M = \bar{X} \quad \text{and} \quad H = n[\textstyle\sum X_i^{-1}]^{-1}, \quad (21)$$

and G is given in (10). Hence:

$$c(n, \alpha) = [(n - 1)G^{-1} + 1]/n. \quad (22)$$

Table 1—*Values of* $c(n, \alpha)$

n \ α	.01	.05	.10
2	.500	.503	.526
4	.272	.324	.386
6	.217	.293	.372
8	.197	.282	.368
10	.185	.276	.368
15	.172	.270	.368
20	166	.267	.368
40	.158	.263	.368
60	.155	.263	.368
∞	.151	.263	.368

It should be pointed out that Nádas' table, obtained by Monte-Carlo Methods, is quite accurate. However, for $n = 60$, the apparent non-monotonicity in Nádas' table is an artifact of his Monte-Carlo procedure.

3. Numerical Example

In Nádas' (1973) application, he observed for $n = 10$:

$$\bar{X} = 1.352, \quad (23)$$

$$H^{-1} = \frac{1}{n}\sum X_i^{-1} = .948 \quad (24)$$

Hence

$$Y = \sum (X_i^{-1} - \bar{X}^{-1}) = 2.083. \quad (25)$$

To test for zero drift, we note that the value of the test statistic,

$$[n(n - 1)]^{-1}\bar{X}Y = .0313, \quad (26)$$

while for type I error, $\alpha = .05$, from (10):

$$G^{-1} = (F_{1,9,.05})^{-1} = .195. \quad (27)$$

Since the observed value of the test statistic falls in the critical region (i.e., is less than .195) the hypothesis of zero drift is rejected with 95% confidence.

A 95% confidence interval for the mean first passage time (or mean failure time) can be obtained by solving the quadratic inequality inside the probability statement of (9). Upon substitution of the observed values, we obtain: (with 95% confidence)

$$.840\mu^2 - 2.704\mu + 1.828 < 0. \quad (28)$$

Solving (28) yields:

$.97 < \mu < 2.25$ is a 95% confidence interval for μ.

4. Conclusion

In this paper, we have given an exact method to test for zero drift in a Brownian Motion Process against the alternative of positive drift. The data are n independent failure-times, each of which is distributed as the first passage of Brownian Motion. Since the exact rejection region, given in (12) is of a very simple form, the test can easily be performed for any n and α.

5. Acknowledgement

The authors wish to acknowledge the support of the National Research Council of Canada.

References

[1] Cox, D. R. and Miller, H. D. (1965). The Theory of Stochastic Processes. Methuen and Co. Ltd., London.

[2] Karlin, S. (1966). A First Course in Stochastic Processes. Academic Press, New York.

[3] Nádas, A. (1973). "Best Tests for Zero Drift Based on First Passage Times in Brownian Motion", *Technometrics 15*, 125–32.

[4] Shuster, J. J. (1968). "A Note on the Inverse Gaussian Distribution Function", *J. Amer. Statist. Assoc. 63*, 1514–16.

[5] Shuster, J. J. and Miura, C. (1972). "Two-way Analysis of Reciprocals", *Biometrika, 59*, 478–81.

[6] Tweedie, M. C. K. (1957). "Statistical Properties of the Inverse Gaussian Distribution, I", *Ann. Math. Statist. 28*, 362–77.

A Restricted Least Squares Problem*

M. S. Waterman

Idaho State University
Pocatello, Idaho

In this note least squares problems with certain convex restraints are solved without the use of linear programming. The solution involves solving the normal equations for a number of unrestricted problems.

KEY WORDS

Least Squares
Regression Analysis

Let $X = (x_{ij})$ be a fixed $m \times n$ matrix of reals and $Y = (y_i)$ be a fixed n-dimensional column vector. The usual least squares problem is to minimize

$$(1) \qquad L(\lambda) = (Y - X^T\lambda)^T(Y - X^T\lambda),$$

where λ is a m-dimensional column vector and A^T denotes the transpose of the matrix A. In this note we consider minimizing $L(\lambda)$ where λ is subject to the restriction $\lambda_i \geq 0$ for $i = 1, 2, \cdots, m$. In the literature various types of inequality restrictions have been considered in least squares and regression problems [3], and the usual solution seems to involve linear programming. Below we solve our stated problem by considering 2^m unrestricted problems.

The real valued function $L(\lambda)$ can be considered defined on $([0, \infty))^m$. Now

$$(2) \qquad L(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ij}\lambda_j\right)^2,$$

so that, with λ_j fixed for all $j \neq k$, we have either $\lim_{\lambda_k \to \infty} L(\lambda) = +\infty$ or $L(\lambda)$ constant. The latter situation occurs only when $x_{ik} = 0$ for all $i = 1, 2, \cdots, m$. Therefore, for the purpose of minimizing $L(\lambda)$, we could restrict λ to a closed and bounded subset of R^m and thus there is a $\lambda \in ([0, \infty))^m$ which minimizes $L(\lambda)$.

If we solve the system

$$(3) \qquad \frac{\partial L(\lambda)}{\partial \lambda_i} = 0, \qquad i = 1, 2, \cdots, m,$$

we find the solution to the unrestricted problem (by solving $XX^T\lambda = XY$, the usual normal equation). If this λ satisfies $\lambda_i \geq 0$, $i = 1, 2, \cdots, m$, then no further work need be done. Otherwise, as (3) is a necessary condition for a minimum to occur in the interior of a set, the solution of the restricted problem must be a boundary point of $([0, \infty))^m$ (see [1, p. 149]). Therefore, at least one $\lambda_i = 0$. Let the index set I range over all 2^m subsets of $\{1, 2, \cdots, m\}$. For each such set I, we consider

$$(4) \qquad L_I(\lambda_I) = \sum_{i=1}^{n} (y_i - \sum_{j \in I} x_{ij}\lambda_j)^2$$

and obtain the (unrestricted) solution. If all λ_k ($k \in I$) satisfy $\lambda_k > 0$, we compute $L_I(\lambda_I)$. Finding a minimum $L_I(\lambda_I)$ solves the problem for we then set $\lambda_l = 0$, $l \notin I$.

What we have in fact done is perform unrestricted least squares calculations for the matrix obtained from X by deleting all lth rows where $l \notin I$. There are $\binom{m}{k}$ least squares calculations in which X has k rows deleted. This procedure is easy to program using existing regression programs. Some excellent treatments of the computation of all possible regressions can be found in [2] and [4].

For a simple example, we let $X = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 3 & 0 \end{bmatrix}$ and $Y = \begin{bmatrix} 2 \\ -3 \\ 6 \end{bmatrix}$. The table below exhibits those λ with all $\lambda_i \geq 0$ and the corresponding $L(\lambda)$.

λ^T	$L(\lambda)$
(11/3, 1/3, 0)	50/3
(4, 0, 0)	17
(0, 9/5, 0)	164/5
(0, 0, 0)	49

Therefore, the minimum value of $L(\lambda)$ is 50/3 and is assumed for $\lambda^T = (11/3, 1/3, 0)$.

The above procedure can be modified to yield

* This work was supported in part by NSF grant GP-28312A1.

Received Sept. 1972; revised May 1973.

solutions to problems where for each i we pick (exactly) one of the conditions (a) $\lambda_i \varepsilon R$, (b) $\lambda_i \geq a_i$, or (c) $\lambda_i \leq a_i$. The solution follows in the same manner.

ACKNOWLEDGEMENT

The author is indebted to J. R. McCown, Jr. of Idaho State University for detecting an arithmetic error in an earlier version of the paper and for helpful discussions.

REFERENCES

[1] APOSTOL, TOM M. (1957). *Mathematical Analysis.* Reading, Mass. and London: Addison-Wesley Publishing Company, Inc.

[2] FURNIVAL, G. M. (1971). All possible regressions with less computation. *Technometrics 13*, 304–408.

[3] JUDGE, G. G. and T. TAKAYAMA (1966). Inequality restrictions in regression analysis. *J. Amer. Statist. Assoc. 61*, 166–181.

[4] SCHATZOFF, M., R. TSAO and S. FEINBERG (1968). Efficient calculation of all possible regressions. *Technometrics 10*, 769–779.