PRESS Used to Study Relationships Between Component Cardiac Weights and Electrocardiographic Patterns

BU-526-M by August, 1974

David M. Allen, James O. Street, A. Mazzoleni

Introduction

This presentation consists of a technique, a problem, and the application of the technique to the problem. The technique is the use of the Prediction Sum of Squares (PRESS) to screen variables.

Press is used in an effort to determine a stable functional relationship between electrical events associated with the heart beat and the physical characteristics sex, length, and age, as well as autopsy variables, including the weight of the ventricles and the presence or absence of myocardial damage. A description of the variables, a list of the data, and computational details are given in appendices.

The Technique

One important aspect of regression analysis is screening a given set of variables. Indeed, variable screening often precedes other uses, helping to identify subsets useful in prediction, estimation, explanation, or control. Variable screening or the more general technique of data augmentation is often necessary because least squares predictors using many variables may be unstable (have a large variance). In order to screen variables, a criterion to evaluate any given subset of variables is required. We define a credible criterion to be 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.

The credible criterion we prefer is the Prediction Sum of Squares (PRESS) defined by

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

where

ℓ	indexes the possible subsets of variables,
Y_i	is the i^{th} observation on the dependent variable,
$\hat{Y}_{\ell(i)}$	is the estimator of $E(Y_i)$ using the ℓ^{th} subset of predictor variables and excluding the i^{th} observation, and
n	is the number of observations.

In words, each observation is "predicted" using the other n - 1 observations. The resulting "errors of prediction" are squared and summed to form PRESS. We like PRESS because it simulates prediction: it does not use an observation to aid in the "prediction" of itself. The computation of PRESS is discussed in Appendix I. Additional discussion of PRESS is given in [1].

We would like to evaluate PRESS for every subset of the predictor variables. We would then choose the subset having the lowest value of PRESS or perhaps a subset with a slightly larger value that is easier to interpret. This procedure is not computationally practical at this time if the number of variables is large. Instead, we use a stepwise procedure that may begin at any subset. For each variable in the subset, the value of PRESS resulting from deleting that variable is determined. For each variable not in the subset the value of PRESS resulting from including that variable is determined. That variable is then included or deleted which yields the smallest value of PRESS. The process is continually repeated until a subset is obtained such that PRESS cannot be decreased by either including or deleting a variable. Since a global minimum is not guaranteed we repeat the process beginning with different initial subsets.

From a practical point of view, the process of variable selection includes choosing the original and transformed sets of variables, applying and re-applying computerized variable screening and ranking procedures, conversing with specialists in the problem context, and, finally, making extra-statistical judgments with respect to a usable, interpretable subset based on a review of the entire problem. Of these activities, the one which most influences the ultimate success or failure of any predictive rule, no matter how subtle the rest of variable selection, is the original choice of possibly useful variables, and this is largely a matter of informed hunch. Also, it may well be important to examine the variables transformed according to reasoning and intuition in addition to the original set, which was itself arbitrary: if we regress areas of circles linearly in their radii, forgetting the work of Archimedes, the fit will be poor.

The Problem

Electrical events associated with the heart beat can be recorded by electrodes applied to the patient's skin. The graphic representation of these electrical events is called an electrocardiogram. In an electrocardiogram, the activation of the ventricles is associated with a series of voltage spikes called the QRS complex. The duration of the QRS complex represents the time taken by the electrical impulses to spread through all the muscle fibers of the ventricles.

It has been stated that enlargement of the human heart is associated with prolongation of the QRS; indeed QRS widening is one of the common criteria for the diagnosis of left ventricular hypertrophy. There is a priori expectation that heavy hearts will have wider QRS complexes and, within limits, this is true. Thus the 15 fold increase in weight of the human heart from infancy to

maturity is accompanied by a QRS widening of about 50%; and a whale heart which weighs 100 times as much as the heart of a human infant has a QRS twice as wide. However, these percentage increases in QRS duration are so modest compared to the accompanying increase in mass as to call into question the assumption that increase in heart weight is an important factor in widening of the QRS duration. This suggests that factors other than weight itself may play a major role in determining the variability of QRS duration found in human hearts.

The objective of the present study is to determine a stable functional relationship between QRS complex and the physical characteristics sex, length, and age, as well as autopsy variables, including the weight of the ventricles and the presence or absence of myocardial damage. The damage can be in the form of an inflammatory reaction (myocarditis) or of dead muscle fibers secondary to occlusion of a coronary artery (myocardial infarction). The infarction may be recent (soft necrotic muscle) or old (firm scar). Fibrosis would be expected to be of great importance for two reasons:

(1) Hearts with fibrosis tend to be heavy [3].

(2) Fibrosis does not conduct electricity as the myocardium does and therefore would be expected to affect the QRS duration [2,5].

The Application of the Technique to the Problem

The cases observed comprise 185 consecutive adult autopsies performed at the Beth Israel Hospital in Boston, Mass., during the period 1953 - 1955. Criteria for inclusion in the study were:

(1) An electrocardiogram was recorded during the six months prior to death. In all but 31 cases an electrocardiogram was recorded during the month preceding death.

(2) The electrodardiogram displayed neither right bundle branch block nor left bundle branch block according to the criteria of Wolff.[6]

The duration of the QRS complex was determined according to the criteria of the New York Heart Association. The technique of Reiner et al. [4] was used for examination of the heart and weighing of its component parts.

We performed stepwise variable screening on the 107 cases having measurements of all variables using the PRESS criterion. Appendix II contains a description and Appendix III a listing of this data. With no variables initialized into the model, the algorithm led to the subset, in order of entry, LENGTH, P-7, CLV, AGE, CLV/CRV, P-11, P-3, where the original eighteen variables plus intercept were available for inclusion. The value of PRESS for this set of seven variables was 3.10×10^5 (residual sum of squares = 2.67×10^5). Since a stepwise rule is locally prudent but globally blind, other computer runs were made using assorted subset initializations as well as those variable transformations suggested by physical reasoning and by examination of previous output. Ironically, the lowest value of PRESS obtained, namely 2.85×10^5 (residual sum of squares = 2.67×10^5), used an eight variable model resulting from a mistaken initialization. Even so, the squared multiple correlation coefficient for this model was only .47, indicating a weak relationship between QRS and the independent variables. The prediction equation for this model is

$$\begin{aligned} QRS = {} & -28.5\ SEX + 9.42(P\text{-}7) + 15.9(P\text{-}11) + 0.443\ LENGTH^{4/5} \\ & -.604(CLV/CRV)^2 + 36.5(P\text{-}3)(P\text{-}7) + 0.531(AGE \times LENGTH) \\ & + 8.08(LV + IVS)^{5/12} \end{aligned}$$

The variables which repeatedly entered the model from differing initializations were considered to be of potential importance in further research. These include AGE, CLV/CRV, P-3, P-7, and P-11.

APPENDIX I. Computational Details

This appendix shows that the PRESS residual is a simple function of the ordinary least squares residual. This is very important for computational purposes.

For simplicity the subscript ℓ is not used. Let X be an n x p matrix of predictor variables and $\underset{\sim}{x}_i'$ be its i^{th} row. Let $\underset{\sim}{Y}$ denote the vector of observations. Equivalent expressions for the $\hat{Y}_{(i)}$ are

$$x_i'(X'X - x_i x_i')^{-1}(X'Y - x_i Y_i) =$$

$$x_i'[(X'X)^{-1} + (X'X)^{-1}x_i(1 - x_i'(X'X)^{-1}x_i')^{-1}x_i'(X'X)^{-1}](X'Y - x_i Y_i) =$$

$$\hat{Y}_i - Q_i Y_i + Q_i(1 - Q_i)^{-1}\hat{Y}_i - Q_i(1 - Q_i)^{-1}Q_i Y_i =$$

$$(1 - Q_i)^{-1}\hat{Y}_i - Q_i(1 - Q_i)^{-1}Y_i$$

where $Q_i = x_i'(X'X)^{-1}x_i$ and $\hat{Y}_i = x_i'(X'X)^{-1}X'Y$.

Thus the residual $Y_i - \hat{Y}_{(i)}$ is $(1 - Q_i)^{-1}(Y_i - \hat{Y}_i)$.

APPENDIX II. Data Description and Format

The card format is: (4X, F1.0, F2.0, 8F3.0, 7F1.0, 2X, F1.0, 6X, F3.0, F1.0)

Variable		Card Columns	Description
(1)	SEX	5	0 if male, 1 if female
(2)	AGE	6-7	(years)
(3)	LENGTH	8-10	Length of body (centimeters)
(4)	FAT	11-13	Thickness of abdominal fat (centimeters)
(5)	LV + IVS	14-16	Weight of left ventricle (grams)
(6)	RV	17-19	Weight of right ventricle (grams)
(7)	LV/RV	20-22	Ratio of variable (5) to variable (6)
(8)	CLV	23-25	Weight of left ventricle corrected for sex: 1.37* LV if female, LV if male
(9)	CRV	26-28	Weight of right ventricle corrected for sex: 1.29* RV if female, RV if male
(10)	CLV/CRV	29-31	Ratio of variable (8) to variable (9)
(11)	P-2	32	1 = normal myocardium 0 = abnormal myocardium
(12)	P-3	33	1 = myocarditis (no dead fibres) 0 = otherwise
(13)	P-4	34	1 = myocarditis (some dead fibres) 0 = otherwise
(14)	P-5	35	1 = exactly one small scar 0 = otherwise
(15)	P-6	36	1 = more than one small scar 0 = otherwise
(16)	P-7	37	1 = one or more large scars 0 = otherwise

(17)	P-8	38	1 = small acute infarction 0 = otherwise
(18)	P-11	41	1 = large acute infarction 0 = otherwise
(19)	QRS	48-50	Duration of QRS complex (milliseconds)
(20)	BBB	51	0 = No bundle branch block 1 = Left bundle branch block 2 = Right bundle branch block

NOTES:

1. Card columns 1-4 contain the subject identification labels, numerically within years: A=1953, B=1954, C=1955.

2. Ignore information in card columns 39-40, 42-47.

APPENDIX III: DATA LIST

```
 A780541620.51730434.01730434.01000000000  -0150950
 A800561762.51860553.41860553.40001000000  -0151000
 A831751523.40940303.11290393.30000001000  -0450700
 A881511601.22150454.82950585.10000000010  +0150900
 A890641701.01380433.21380433.21000000000  -0450700
 A930651551.00980352.80980352.81000000000  +0750600
 A961641504.02660594.53640764.80000100000  -0150850
A1010781755.03300655.13300655.10000100101  -0151200
A1060451703.53170654.93170654.91000000000  +0150900
A1170741555.02600525.02600525.00000010000  -0450850
A1280671641.01500473.21500473.21000000000  +0151000
A1290701481.01610503.21610503.20000001100  +0450600
A1320621563.52570426.12570426.10000011000  +0150950
A1421611603.02470604.13380774.40000010101  +0151500
A1441651685.01150353.31580453.51000000000  -0150800
A1451771454.51430403.61960523.81000000000  -0451100
A1581661552.51790463.92450594.21000000000  -0150600
A1670881701.01220343.61220343.61000000000  +0150900
A1700841642.52470425.92470425.90000010000  +0750950
A1711571573.02610624.23580804.50000011100  -0151050
A1731851501.01420334.31950434.50000100000  -0150950
A1741761501.01640553.02250713.20001000000044-0151400
A1751671551.71560612.62140792.70101000000  -0451050
 B70691601.52540843.02540843.00000100000  -0450950
B161251611.01000442.31370572.41000000000  +1050700
B171791605.01490423.52040543.81000000000  -0150800
B200651620.51520662.31520662.31000000000  -0450900
B260631603.52750574.82750574.80000011000  +0151000
B360831641.51850345.41850345.40000000101  +0450900
B370711682.51910404.81910404.81000000000  -0450900
B381701584.01680463.72300593.90000010101  +1051200
B451731603.01460334.42000434.70000001000  -0150800
B460711624.02250464.92250464.91000000000  +0150700
B471471680.41401221.11921571.20001000000  +1050800
B501671603.51170393.01600503.20100000000  -0151000
B521651454.02020533.82770684.10000000111  +0150850
B541681532.01250423.01710543.21000000000  +0450800
B550801657.02230514.42230514.40000000101  -0751400
B561311603.01960424.72690545.01000000000  +0450600
B570731602.51930563.51930563.51000000000  -0151000
```

APPENDIX III: DATA LIST

```
 B630631753.02140464.72140464.70000000010  +0150900
 B671481585.01360482.81860623.00000010000  +0451000
 B740891602.52090385.52090385.51000000000  +0150900
 B751671594.02070474.42840614.71000000000  -0451100
 B760701701.33150605.23150605.20000010101  +0751300
 B790431763.03800924.13800924.10100000000  +0151000
 B820791582.51830374.91830374.90000011000  -0150900
 B831681534.01700453.82330584.01000000000  -0150800
 B861601534.51330304.41820394.70100000000  -0150900
 B880691682.01780384.71780384.70001000000  +0450800
 B891681581.52030553.72780713.90000010101  +0151000
 B960741720.72320425.52320425.50100010000  -0151400
B1041581607.01890394.82590505.20010000000  -0151200
B1110671682.02630654.02630654.00010010000  +1351100
B1160741621.82440554.42440554.40000010000  +0150950
B1200581611.81500374.01500374.00000000101  -0451300
B1210531633.02270405.72270405.71000000000  -0150800
B1220591552.51600622.61600622.60000101110  +0751000
B1240671553.52630564.72630564.70000011000  +0451400
B1301541633.61540493.12100633.31000000000  -0150900
B1310691652.22560495.22560495.20000010101  -0151000
B1340361602.03530635.63530635.61000000000  -0151100
B1370721701.02040405.12040405.11000000000  +0451000
B1410621622.02400425.72400425.70000010111  +0150900
B1460651551.02510703.62510703.60000010000  +0150900
B1481671533.01090313.61490393.81000000000  +0150800
B1511621633.11140383.01550483.20010001000  -0450900
B1541691500.21160264.51590344.70100000000  +1050700
B1640721754.02710535.12710535.10000011000  -0151000
B1720751671.02070434.82070434.80000100000  -1050700
B1731671491.71580453.52160583.71000000000  +0450800
B1760671601.01360423.31360423.30000010000  +0150900
B1770591601.12290663.52290663.50000010010  -0451300
B1810541782.53060525.93060525.90100000000  -0150900
B1821761523.12470445.63380575.91000000000  +0151000
B1850591593.11710453.81710453.80000000101  -0450750
B1860511602.03010565.43010565.40000100000  -0151100
  C40721703.01720533.31720533.31000000000  -0150900
  C81491632.51290891.41771151.50010000000  +1350700
 C131781401.21770384.62420494.90001001000  +0450900
```

APPENDIX III: DATA LIST

```
 C180661752.01560493.21560493.20000010010  +0750900
 C240711460.11280255.21280255.20000001000  +0450750
 C280701612.01340334.11340334.10000100000  +0450800
 C311681623.51230343.61690443.81000000000  +0750800
 C410811563.01550513.01550513.01000000000  -0450750
 C441431550.51630414.02230524.31000000000  +0150700
 C461671703.01500433.52050553.71000000000  +0450900
 C610441651.21130402.81130402.81000000000  +0750800
 C621711452.01200323.81640414.01000000000  +0750850
 C711661602.72100683.12880883.30100010000  +1351400
 C751671642.81610404.02210524.20000100101 44-0751400
 C821721452.01620394.22220504.40000010101  -0151000
 C830741522.02350852.82350852.80000010000 44-0151250
 C850691802.02210554.02210554.00000010000  -0451100
 C980731502.02560455.72560455.70000010000  +0150950
C1051631431.51480463.32030593.41000000000  +0450800
C1130531621.02260554.12260554.11000000000  +0151000
C1150601681.81560443.51560443.51000000000  -0450900
C1171221571.41780394.62440504.91000000000  +0150600
C1200651732.12690913.02690913.00000011000  +1051600
C1211771581.51210294.21660374.50000010000  +0750800
C1230671630.82330713.32330713.30000010000  +0750900
C1300561722.41850603.11850603.11000000000  +0150600
C1330591552.02580803.22580803.20000010000  +0751100
C1351631605.51680553.12300713.21000000000  -0150950
C1420561806.52250942.42250942.40010000000  +0150900
C1470511702.04510895.14510895.10000001000  +0451100
```

References

[1] Allen, D. M. (1974). The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction. Technometrics 16, 125-127.

[2] First, S. R., Bayley, R. H., and Bedford, D. R. (1950). Peri-Infarction Block: Electrocardiographic Abnormality Occasionally Resembling Bundle Branch Block and Local Ventricular Block of Other Types. Circulation 2, 31.

[3] Mazzoleni, A., Reiner, L., Rodriguez, F. L., and Freudenthal, R. R. (1964). The Weight of the Human Heart. III. Ischemic Heart Disease. Archives of Pathology 77, 205.

[4] Reiner, L., Mazzoleni, A., Rodriguez, F. L., and Freudenthal, R. R. (1959). The Weight of the Human Heart. I. "Normal" Cases. Archives of Pathology 68, 58.

[5] Rosenbaum, M. B., Elizari, M. V., and Lazzari, J. O. (1970). The Hemiblocks. Tampa Tracings.

[6] Wolff, L. (1956). Electrocardiography, Fundamentals and Clinical Application. W. B. Saunders Co., Philadelphia-London.

10/74