VARIABLE SELECTION

A researcher's primary interest is often:
"to make as good a prediction to a criterion as possible on the basis of several variables."

The aim is usually the selection of the minimum number of variables necessary to account for almost as much of the variance as is accounted for by the total set.

There is no unique statistical procedure for identifying "the best regression equation"

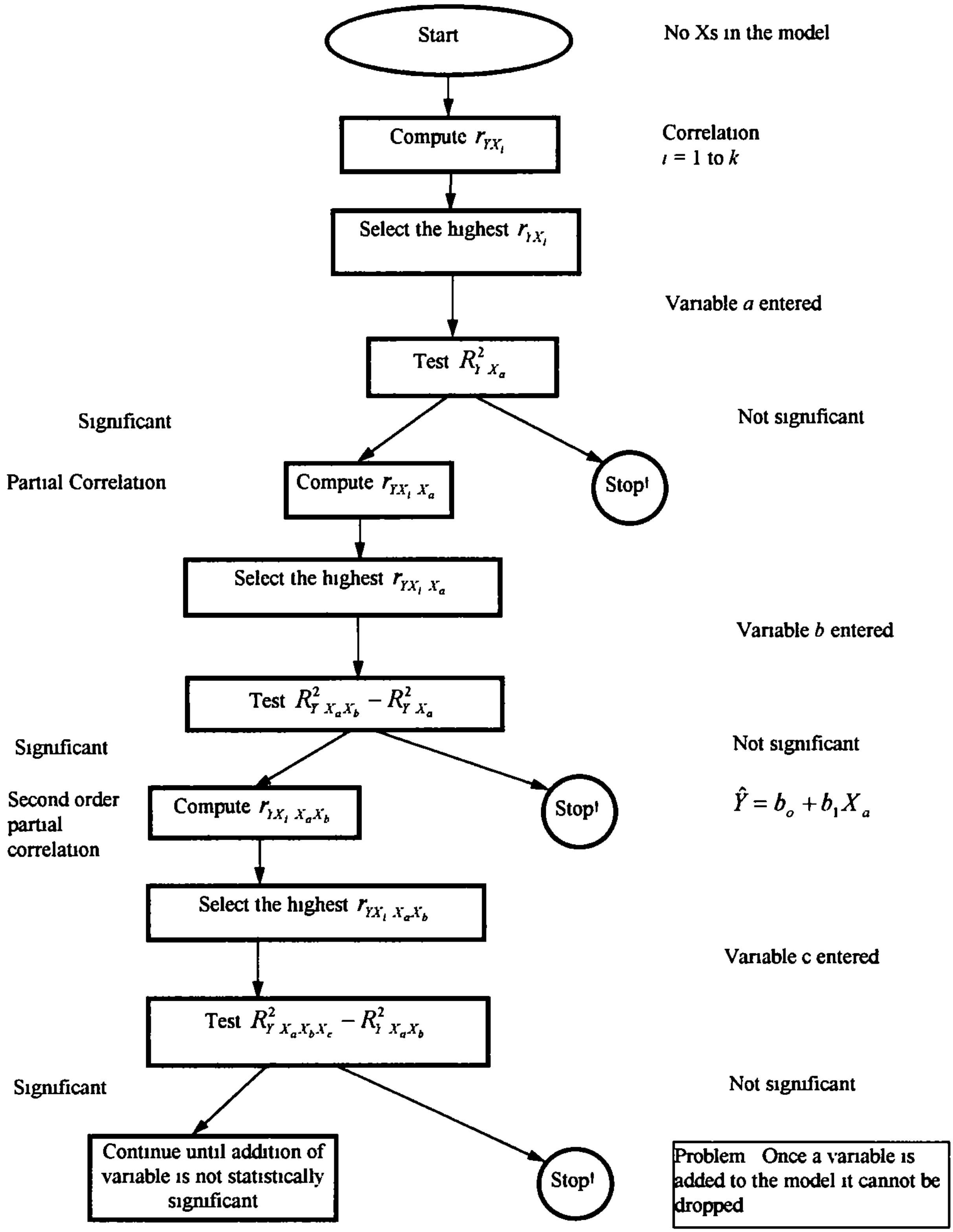
1. Forward Selection Method.

2. Backward Elimination Method.

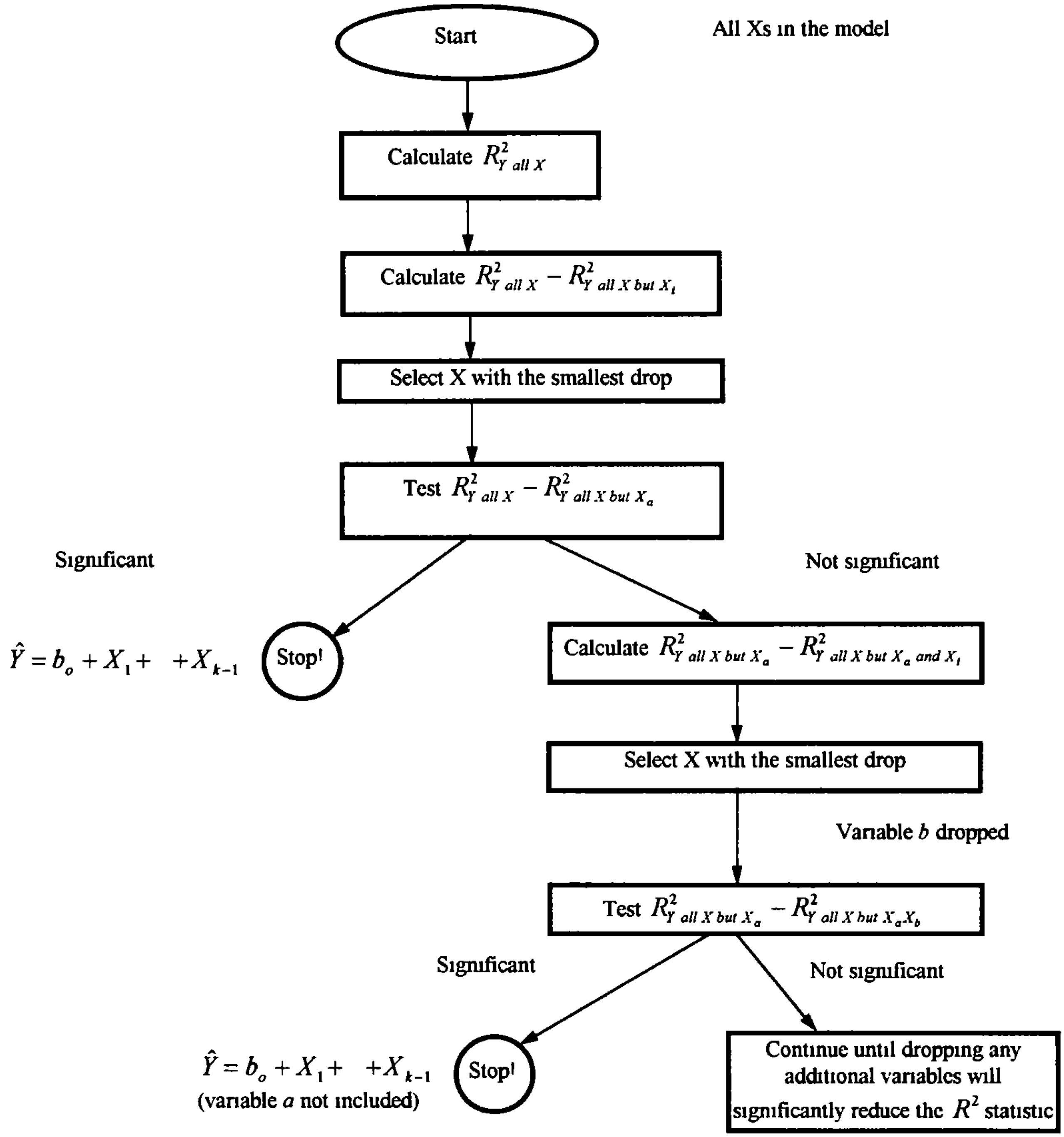
3. Stepwise Selection Method.

4. All Possible Regressions.

FORWARD METHOD

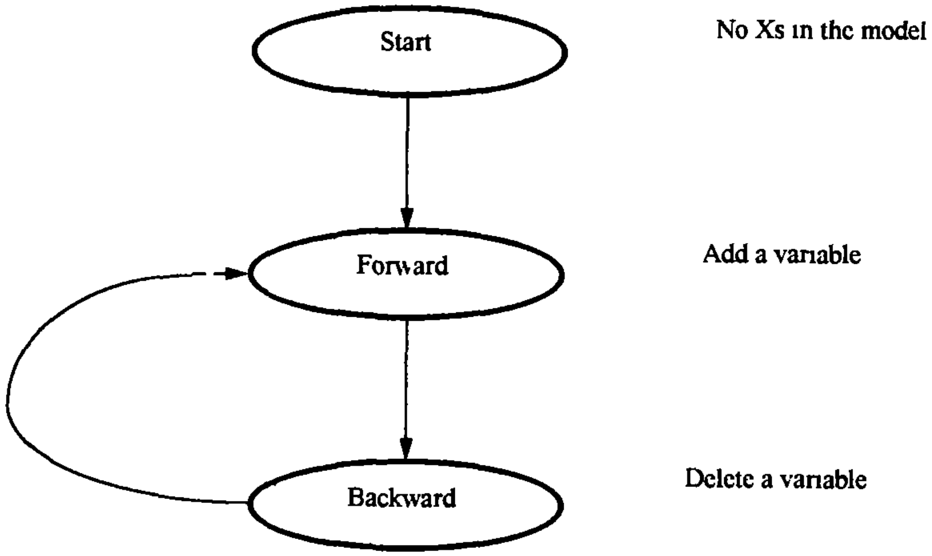


BACKWARD METHOD



Problem Once a variable has been dropped, it is no longer considered later

STEPWISE METHOD



Continue until the addition of the new variable does not increase R^2 and the deletion of additional variables reduces the R^2 statistic significantly

ALL POSSIBLE REGRESSIONS

Adequate R^2

The R^2 statistic is not significantly different from the Full model R^2

$$R^2 = 1 - \left(1 - R^2_{FM}\right)\left(1 + \left[k \bullet F^*_{k,N-k-1} / (N - k - 1)\right]\right)$$

Mallow's C_P (1973)

$$C_P = \frac{SSE_{(k)}}{MSE_{FM}} - N + 2(k + 1)$$

Values of $C_P \approx k + 1$ are desirable

(Marginal note: SSE refers to residual)

MSE or R^2_{adj}

$$R^2_{adj} = 1 - \frac{MSE}{MST}$$

Usually the larger the k, the smaller the MSE. However, the MSE can increase if the researcher adds too many correlated variables.

Choose the model with the smallest MSE or largest R^2_{adj} .

The following table summarizes the analysis of all possible regressions.

# variables	Regressors	$SSE_{(k)}$	R^2	$_{adj}R^2$	$MSE_{(k)}$	C_p
none	0	2715.76	0	0	226.31	442.92
1	X1	1265.68	.534	.492	115.06	202.55
1	X2	906.34	.666	.636	82.39	142.49
1	X3	1939.40	.286	.221	176.31	315.16
1	X4	883.86	.675	.645	80.35	138.73
2	X1X2	57.90	.979	.974	5.79	2.68
2	X1X3	1227.07	.548	.458	122.70	198.10
2	X1X4	74.76	.972	.966	7.47	5.50
2	X2X3	415.44	.847	.816	41.54	62.44
2	X2X4	868.88	.680	.616	86.89	138.23
2	X3X4	175.73	.935	.922	17.57	22.37
3	X1X2X3	48.11	.982	.976	5.35	3.04
3	X1X2X4	47.97	.982	.976	5.33	3.02
3	X1X3X4	50.84	.981	.975	5.65	3.50
3	X2X3X4	73.81	.973	.964	8.20	7.34
4	X1X2X3X4	47.86	.982	.973	5.98	5.00

An adequate R^2 would equal:

$$R^2_0 = 1 - (1 - R^2_{Y|X1X2X3X4})\,(\,1 + [KF^*_{K,N-K-1,\alpha}/(N-K-1)])$$

$$= 1 - (1 - .982)(1 + (4)(3.84)/8))$$
$$= 1 - (.018)(2.92)$$
$$= 1 - .05256$$
$$= .9474$$