

PROJECT ONE

THE SOUTH AUSTRALIAN RESEARCH

<u>Section</u>	<u>Contents</u>	<u>Page Number</u>
1	Foreword	1.1
2	Introduction.	2.1
3	Methodology	3.1
4	Results	4.1
5	Discussion	5.1
	A. General	5.1
	B. The Regression Equations . . .	5.9
	C. The S.A. Project in retrospect	5.11
	D. A working model	5.24
6	Summary	6.1
7	Appendix	7.1
	A. Statistical techniques	
	B. S.A. demographic variables	
	C. S.A. personal votes and donkey votes	
	D. Rocky River - an application of the regression equations	
	E. An integration of attitudinal and statistical research	

1.1

Foreword: This project summarises demographic research into the South Australian elections of 1973 and 1975. The research was used by strategists to help plan the successful S.A. 1977 election campaign.

Project one was included in the present report to show the background and development of demographic research techniques and to put these techniques into the broader context of a complete campaign strategy.

Project one is also useful for the reader in that it integrates theoretical work in a reasonably-practical manner with a campaign that was really quite successful in terms of what it set out to do (see sections 7D and 7E).

This was a sharp contrast to the following state election in S.A. in 1979. This campaign with its "gung-ho" anti-theoretical approach to campaigning used the tactical tools by which a strategy can be implemented (policy initiatives, press statements, campaign speeches and media advertisements) as a substitute for the strategy itself.

The lessons, I believe, are clear. Pragmatic, short-term campaign techniques which operate in a theoretical void do not win elections for Oppositions. However, they can lose elections for Governments.

THE SOUTH AUSTRALIAN RESEARCHIntroduction

In 1977 the South Australian Labor Government was presented with a new set of State electoral boundaries by its recently-formed independent boundaries commission.

The commission had based the boundaries on one state-wide quota with a ten per cent tolerance. This changed the shape of most electorates, creating five extra city seats and abolishing five in the country.

The 1973 State elections had been an exceptionally good result for Labor, while the 1975 poll result had been just good enough for the Government to survive.

Some of the new 1977 electorates were held by Labor on the 1973 results and by non-Labor on the 1975 figures. Some had changed so radically as to be difficult to classify according to the previous votes.

All new candidates and sitting members were anxious to find out how the new electorates had voted in previous elections and how much of this vote had been a function of the personal vote of sitting members and how much could be attributed to a class base of support.

In addition to these practical problems and decisions associated with the quantification of a Labor "base vote" in each of the new seats, the Labor Government's Cabinet Campaign Committee wanted to lay down a medium term strategy for a snap election later in 1977. This strategy would be based on the identification (and political wooing) of the group or groups of voters who had deserted Labor in the State Elections of 1975. The committee wanted to know what sort of persons these voters were, and perhaps even more importantly, how they were distributed across the new electorates. The committee had become increasingly concerned at the variation in swings across electorates and was well aware, that in a (one-quota) single-member constituency system, the electoral strategy which returned over 50 percent of the preferred votes would not necessarily return more than 50 percent of the seats.

Methodology

South Australia was divided into two units: city and country. The city consisted of the 33 new metropolitan seats. The country area was composed of the 14 new country seats.

1971 State Census data was rearranged on the basis of the new electorates. This process consisted simply of the allocation of about 35 very small census collectors' districts (consisting of some 250 homes) to each new electorate. The (35) collectors' districts were then aggregated to form individual electorate summaries.

The 1973 and 1975 State election results were allocated to new electorates, on the basis of polling booth locations (and major transport flows). The 1975 Legislative Council election results (2PP for each seat) were also calculated for new electorates and subtracted from the Lower House 2PP results to obtain Personal Vote scores for 1975 House of Assembly Labor candidates. Four major political variables were then calculated for each new city and country seat:

- * The 1973 ALP Vote, after preferences.
- * The 1975 ALP Vote, after preferences.
- * The 1973-75 Anti-ALP Swing.
- * The 1975 Personal Vote scores for all Labor Candidates in country seats.

Appropriate demographic variables were selected from the 1971 Census data for city and country seats. These variables could be grouped under seven main headings:

- * CLASS
- * HOUSING
- * FAMILY STATUS
- * EDUCATION
- * AGE
- * ETHNICITY
- * RELIGION

1. An explanation of the construction of this variable is provided in the Appendix to Project 1.

3.2

Twenty-one demographic variables were selected for the city seats, and 28 variables were selected for the country seats. The country seats contained extra variables dealing with class, religion and ethnicity. The occupational-class variables used for the country seats also differed slightly in composition to the class variables chosen for the city seats. A break-down of occupational class variables for city and country seats is provided below (Table 5.1), but the country class variables "Blue Collar (Urban)" and "Middle White Collar (Urban)" can be realistically compared with the city class variables "Blue Collar Workers" and "Middle White Collar Workers" respectively.

Further details of these variables and their values for each seat are provided in the Appendix to Project 1.

TABLE 13 OCCUPATION OF EMPLOYED PERSONS

OCCUPATIONAL CLASS - CITY

UPPER WHITE COLLAR (U.W.C.)	1	Architects, Engineers, Surveyors, Professionals	38	Telephone, Telegraph, Telecom. Operators
	2	Chemists, Physicists, Other Physical Scientists	39	Postmasters, Postmen and Messengers
	3	Biologists, Vets, Agronomists, Related Scientists	40	Workers in Transport and Communication
	4	Medical Practitioners	41	Spinners, Weavers, Knitters, Dyers and related
	5	Dentists	42	Tailors, Cutters, Furriers, related workers
	6	Nurses including trainees	43	Leather Cutters, Lasters, Sewers and related
	7	Professional Medical workers	44	Furnacemen, Rollers, Moulders, Metal makers
	8	Teachers	45	Instrument makers, Jewellers and related workers
	9	Clergy and Members of Religious Orders	46	Metal tradesmen, Mechanics etc.
	10	Law Professionals	47	Electricians and related workers
M.W.C.	11	Artists, Entertainers, Writer, Related workers	48	Metal and Electrical Prod - Process workers
	12	Draftsmen and Technicians	49	Carpenters, Wood machinists etc.
	13	Other Professional, Technical and Related workers	50	Painters and Decorators
	14	Administrative and Executive Officials, Government	51	Bricklayers, Plasterers, Construction workers
	15	Employers, Managers, Workers own account	52	Printing Trades workers
	16	Book-keepers and Cashiers	53	Pottery, Glass and Clay workers
	17	Stenographers and Typists	54	Millers, Bakers, Food and Drink workers
	18	Other Clerical workers	55	Chemical, Sugar, Paper, Prod - Process workers
	19	Insurance, Real Estate, Salesmen, Valuers	56	Tobacco Product makers
	20	Commercial Travellers, Manufacturers Agents	57	Rubber, Plastic, Concrete Prod - Process workers
U.W.C.	21	Proprietors, Shopkeepers, Shop Assistants etc.	58	Packers, Wrappers, Labellers
	22	Farmers and Farm Managers	59	Stat. Engine, Excavating, Lifting operators
	23	Farm workers including Farm Foreman	60	Storemen and Freight handlers
	24	Wool Classers	61	Labourers not included elsewhere
	25	Hunters and Trappers	62	Fire Brigade, Police and Protective workers
	26	Fishermen and related workers	63	Housekeepers, Cooks, Maids etc.
	27	Timber Getters	64	Waiters, Bartenders
	28	Miners, Mineral Prospectors and Quarrymen	65	Caretakers, Cleaners - buildings
	29	Well Drillers, Oil, Water and Related workers	66	Barbers, Hairdressers and Beauticians
	30	Mineral Treaters	67	Launderers, Dry Cleaners and Pressers
MIDDLE WHITE COLLAR (M.W.C.)	31	Dock and Engineer Officers - Ship	68	Athletes, Sportsmen and related workers
	32	Deck, Engineer room hands, Ship and Boatmen	69	Photographers and Camera operators
	33	Air Pilots, Navigators, Flight Engineers	70	Undertakers and Crematorium workers
	34	Drivers and Firemen, Rail Transport	71	Service, Sport, Recreation workers
	35	Drivers, Road Transport	72	Members of Armed Services
	36	Guards and Conductors - Railway	73	Inadequately described or not stated
	37	Inspectors, Supervisors, Controllers		
BLUE COLLAR				

BLUE
COLLAR

TABLE 13 OCCUPATION OF EMPLOYED PERSONS

OCCUPATIONAL CLASS - COUNTRY

MIDDLE WHITE COLLAR (URBAN)	1	Architects, Engineers, Surveyors, Professionals	38	Telephone, Telegraph, Telecom. Operators
	2	Chemists, Physicists, Other Physical Scientists	39	Postmasters, Postmen and Messengers
	3	Biologists, Vets, Agronomists, Related Scientists	40	Workers in Transport and Communication
	4	Medical Practitioners	41	Spinners, Weavers, Knitters, Dyers and related
	5	Dentists	42	Tailors, Cutters, Furriers, related workers
	6	Nurses including trainees	43	Leather Cutters, Lasters, Sewers and related
	7	Professional Medical workers	44	Furnacemen, Rollers, Moulders, Metal makers
	8	Teachers	45	Instrument makers, Jewellers and related workers
	9	Clergy and Members of Religious Orders	46	Metal tradesmen, Mechanics etc.
	10	Law Professionals	47	Electricians and related workers
BLUE COLLAR (RURAL)	11	Artists, Entertainers, Writer, Related workers	48	Metal and Electrical Prod - Process workers
	12	Draftsmen and Technicians	49	Carpenters, Wood machinists etc.
	13	Other Professional, Technical and Related workers	50	Painters and Decorators
	14	Administrative and Executive Officials, Government	51	Bricklayers, Plasterers, Construction workers
	15	Employers, Managers, Workers own account	52	Printing Trades workers
	16	Book-keepers and Cashiers	53	Pottery, Glass and Clay workers
	17	Stenographers and Typists	54	Millers, Bakers, Food and Drink workers
	18	Other Clerical workers	55	Chemical, Sugar, Paper, Prod - Process workers
	19	Insurance, Real Estate, Salesmen, Valuers	56	Tobacco Product makers
	20	Commercial Travellers, Manufacturers Agents	57	Rubber, Plastic, Concrete Prod - Process workers
BLUE COLLAR (URBAN)	21	Proprietors, Shopkeepers, Shop Assistants etc.	58	Packers, Wrappers, Labellers
	22	Farmers and Farm Managers	59	Stat. Engine, Excavating, Lifting operators
	23	Farm workers including Farm Foreman	60	Storemen and Freight handlers
	24	Wool Classers	61	Labourers not included elsewhere
	25	Hunters and Trappers	62	Fire Brigade, Police and Protective workers
	26	Fishermen and related workers	63	Housekeepers, Cooks, Maids etc.
	27	Timber Getters	64	Waiters, Bartenders
	28	Miners, Mineral Prospectors and Quarrymen	65	Caretakers, Cleaners - buildings
	29	Well Drillers, Oil, Water and Related workers	66	Barbers, Hairdressers and Beauticians
	30	Mineral Treaters	67	Launderers, Dry Cleaners and Pressers
BLUE COLLAR (URBAN)	31	Dock and Engineer Officers - Ship	68	Athletes, Sportsmen and related workers
	32	Deck, Engineer room hands, Ship and Boatmen	69	Photographers and Camera operators
	33	Air Pilots, Navigators, Flight Engineers	70	Undertakers and Crematorium workers
	34	Drivers and Firemen, Rail Transport	71	Service, Sport, Recreation workers
	35	Drivers, Road Transport	72	Members of Armed Services
	36	Guards and Conductors - Railway	73	Inadequately described or not stated
	37	Inspectors, Supervisors, Controllers		

BLUE
COLLAR
(URBAN)

Pearson r tables were prepared for both city and country seats for all available political and demographic variables. The Pearson r, loosely speaking, measures the strength of relationship between pairs of variables. A score approaching plus or minus one indicates the presence of a very strong positive or negative relationship (respectively) between two variables, while a score approaching zero indicates the absence of a relationship.

It should be stressed that the Pearson r Tables, identify only ecological relationships between political and demographic groups across South Australia. For example, all political observers realise that Housing Trust tenants tend to live in strong, pro-Labor areas. This is reflected in the findings listed in the Pearson r Tables. But it does not necessarily mean that Housing Trust tenants actually vote Labor; and, even if they do, it does not mean that they vote Labor simply because they are tenants of public housing. Indeed, the more sophisticated statistical techniques used in multiple regression analysis (described below), show that Housing Trust tenancy per se, served to reduce the State Labor vote in 1973.

Multiple Regressions for the political variables were then carried out to perform a more critical analysis of South Australian political behaviour and to provide a basis for predictions of future political behaviour. A simplified explanation of the Multiple Regression technique is provided below, together with an explanation of the Multiple Regression Tables.

The following section on Multiple Regression should be read in conjunction with Table 1.6. (page 4.6).

1. There is a title at the top of Table 1.6 which describes the political variable under examination (V3 - 1973 ALP city vote - not adjusted).

2. The column at the far left of the table gives the code number of each demographic variable which "explains" portion of the given political variable. (The code numbers differ between city and country seats.)
3. The demographic variables are described in the column second from the left (e.g. "upper white collar workers").
4. In the third column from the left, the table lists the total variance explained by variables which have been computed to that stage. The term "variance" can be thought of as analagous to "variation (between seats)". For example, in the first table, 85% of the variance is explained by the Upper White Collar Worker Variable. The reader could interpret this loosely as meaning Semaphore recorded a 1973 vote of 76% and Davenport 31%, mainly because Semaphore had only 5.5% of its workforce in the Upper White Collar range, while 31% of the Davenport workforce were in this category.
5. The next column, third from the right, details the "extra amounts of variation" of the political variable explained by the computation of additional demographic variables. For example, in Table 1.6 a relatively-tiny amount of "variation" is explained by the Middle White Collar Workers variable (3.9%), once the first variable, Upper White Collar Workers, has been taken into account.
6. The "Coefficient and Constant" column, second from the right, gives the factor by which each variable must be multiplied to provide for the regression equation (explained below).
7. The sign of the coefficient indicates whether or not the corresponding demographic variable provides a positive or negative contribution to the explanation of the political variable being examined. For example, in Table 1.6 Upper White Collar Workers can be seen to have had a negative impact on the Labor vote (i.e. they voted against us). The constant provided at the base of this column is also used in the prediction equations (see below).

8. The column to the far right of the Regression Tables gives the simple Pearson r between the political variable and demographic variables. The Pearson r usually carries the same sign as the regression coefficient. In Table 1.6, however, the reader can see that there was a positive Pearson r between the city 1973 ALP Vote and Housing Trust tenants (+.47). However the Table 1.6 regression coefficient (-.07) is negative. This can be interpreted as meaning two things: Firstly, that Housing Trust tenants voted Labor in 1973, and secondly, that this pro-ALP vote was only a function of the Trust tenants' Class, Age and Education. In other words, when you allow for the fact that Housing Trust tenants were likely to be Working Class, poorly educated, and of non-pensionable age, their public housing tenancy only served to reduce their support for the ALP. This is an excellent illustration of the uses to which the Multiple Regression technique can be put.

9. The formulae listed below each regression table enable the reader to calculate the value of the political variables (e.g. Vote and Swing) from the demographic variables listed in the Appendix. The Full Formula explains more variance than the Short Formula, and can therefore be considered to be more accurate than the Short Formula.

10. The inaccuracy inherent in the prediction equations is shown in the standard error of estimate figure, to the right of the equation in brackets. In Table 1.6 for example, using the Full Formula, there is a 68% chance that the predicted result will be within 3.7% of the actual 1973 vote in the area under examination. There is a 95% chance the predicted result will be within 7.4% of the actual result.

A more detailed explanation of the theory underlying the Pearson Correlations and the Multiple Regression analysis is produced in the Appendix to Project 1.

Results

The major results from the South Australian project have been summarised in two forms: Pearson r tables (tables 1.2 to 1.5) and Multiple Regression Tables (tables 1.6 to 1.12).

The Pearson r figures, as discussed earlier in the Methodology section of this project, are a relatively superficial and static description of a given political occurrence. They are subject to the ecological fallacy and should be treated with caution, unless the r figures are very high.

Irrespective of this ecological problem, the Pearson r figures are useful in that they serve to focus attention on the more relevant data.

Also, simple ecological relationships are useful in themselves to flesh out a general picture of key target groups for campaign strategists.



The Multiple Regression Tables provide much more significant examples of possible causality. They also enable a more dynamic method of analysis to be employed.

The regression procedure is based on a series of steps, represented by successive lines in the tables. Before proceeding to each new step, the regression program employed calculates a completely new set of partial correlations which allow for the variance explained by the preceding step or steps. In this way the ecological problems encountered by the use of Pearson r s only are minimised and the additional steps are only used if they provide ^{extra}/explanatory power.

As explained in the Methodology, this extra explanatory power is quantified in the "Extra Variance Explained" column of the Regression Tables.

NOTE: The reader should note that the swings between 1973 and 1975 were against Labor in all seats. In the present project these swings are given a positive sign. Therefore a positive correlation between any demographic variable and the 1973-75 anti-Labor swing indicates that support for Labor fell amongst this demographic group between 1973 and 1975. A negative correlation indicates a rise in support among any demographic group. In subsequent projects the sign was used as a measure of the direction of swing and this caution can be disregarded.

THE 1973 STATE LABOR VOTE : ALIGNED AND
NON-ALIGNED GROUPS IN THE CITY AND THE COUNTRY

1973 LABOR VOTE		
TREND	CITY	COUNTRY
 LABOR STRENGTH	+ .98 1975 Labor Voters + .93 Blue Collar Workers + .56 Overseas-born + .47 Housing Trust tenants + .37 Catholics	+ .96 1975 Labor Voters + .87 Blue Collar Workers (Urban) + .83 Catholics + .81 Housing Trust tenants + .78 18-24 year olds + .73 Middle White Collar Workers (Urban) + .71 Overseas-born + .69 German-born + .68 U.K.-born + .64 Short-term residents + .61 35-44 year olds + .59 Yugoslav-born + .59 "Other European"-born + .57 25-34 year olds + .53 Italian-born
	+ .33 U.K.-born	+ .50 Schoolchildren + .50 Church of England + .50 Agnostics
 LABOR WEAKNESS	- .33 55-64 year olds	- .49 Lutherans
	- .57 65+ year olds - .75 Matriculants - .89 Middle White Collar Workers - .92 Upper White Collar Workers	- .53 Methodists - .58 45-54 year olds - .61 Blue Collar Workers (Rural) - .62 55-64 year olds - .63 Personal voters - .66 65+ year olds - .84 Agricultural Workforce

NOT
SIGNIFICANT
TO .05

Note:

TABLE 1.2

City Seats

Country Seats

For r of .34, sign. = .05

For r of .52, sign. = .0

For r of .44, sign. = .01



For r of .60, sign. = .0

Some Pearson r correlations not significant at the

.05 level have been included for both city and

country seats.

THE 1975 STATE LABOR VOTE: ALIGNED AND
NON-ALIGNED GROUPS IN THE CITY AND THE COUNTRY

1975 LABOR VOTE		
TREND	CITY	COUNTRY
 LABOR STRENGTH	+ .98 1973 Labor Voters + .96 Blue Collar Workers + .56 Housing Trust tenants + .55 Overseas-born + .40 Catholics	+ .96 1973 Labor Voters + .93 Blue Collar Workers (Urban) + .82 Housing Trust Tenants + .79 Catholics + .78 Middle White Collar Workers (Urban) + .75 German-born + .73 Overseas-born + .71 18-24 year olds + .71 U.K.-born + .65 Short-term residents + .60 Yugoslav-born + .59 "Other-European"-born + .56 35-44 year olds + .53 25-34 year olds + .53 Italian-born
LABOR WEAKNESS 	- .51 65+ year olds - .75 Matriculants - .91 Middle White Collar Workers - .93 Upper White Collar Workers	- .58 65+ year olds - .60 45-54 year olds - .67 Personal Voters - .81 Blue Collar Workers (Rural) - .89 Agricultural Workforce

Note:

TABLE 1.3

City Seats

For r of .34, sign. = .05

For r of .44, sign. = .01



Country Seats

For r of .52, sign. = .

For r of .60, sign. = .

THE 1973 STATE LABOR VOTER WHO VOTED NON-LABOR

IN 1975 : THE VOLATILE VOTER

THE VOLATILE VOTER		
TREND	CITY	COUNTRY
 INCREASING VOLATILITY	+ .74 25-34 year olds + .58 Short-term residents + .56 Pre-school children + .51 British-born + .41 35-44 year olds + .40 Church of England	+ .76 Presbyterians
		+ .46 Schoolchildren + .39 18-24 year olds + .32 1973 Labor Voters + .31 35-44 year olds + .31 Church of England + .30 Catholics
INCREASING STABILITY (OR PRO- LABOR SWING) 	- .46 45-54 year olds - .49 65+ year olds - .50 18-24 year olds - .57 Tertiary students - .58 55-64 year olds	- .39 Lutherans - .40 Tertiary Students - .40 65+ year olds - .41 Methodists
		- .53 55-64 year olds

↑
NOT
SIGNIFICANT
TO .05
↓

Note:

TABLE 1.4

City SeatsCountry Seats

For r of .34, sign. = .05 For r of .52, sign. = .05
 For r of .44, sign. = .01 For r of .60, sign. = .01
 Some Pearson r correlations not significant at the
 .05 level have been included for the country seats.

THE 1975 COUNTRY VOTER WHO SPLIT HIS TICKET
BETWEEN THE UPPER AND LOWER HOUSES :
THE PERSONAL VOTER





THE PERSONAL VOTER	
TREND	COUNTRY
 PERSONAL VOTER	+.72 Blue Collar Workers (rural)
	+.46 Agricultural Workforce - +.40 Lutherans +.30 Greek-born 
CLASS VOTER 	-.33 UK-born -.34 18-24 year olds -.44 Housing Trust tenants -.44 Italian-born -.46 Blue-collar workers (urban) 
	-.54 Catholics -.63 1973 Labor Voters -.63 Middle White Collar Workers (urban) -.66 1975 Labor Voters

TABLE 1.5

Note : For r over .52, significance = .05
 For r over .60, significance = .01
 Some Pearson r correlations not significant
 at the .05 level have been included for
 country seats.

MULTIPLE REGRESSION

V3 1973 ALP CITY VOTE (NOT ADJUSTED)

VARIABLE NUMBER	VARIABLE	VARIANCE EXPLAINED %	EXTRA VARIANCE EXPLAINED %	COEFFICIENT AND CONSTANT	PEARSON R
V6	Upper White Collar Workers	85.0	85.0	-.93	-.92
V5	Middle White Collar Workers	88.9	3.9	-.48	-.88
V14	65+ Year Olds	90.6	1.6	-.40	-.57
V22	Matriculants	92.1	1.5	-.52	-.76
V8	Housing Trust tenants	92.6	0.5	-.07	+.47
V15	Overseas-born	92.8	0.1	+.14	+.56
				96.7	

ULL FORMULA: (92.8% VARIANCE)

$$V3 = -.93V6 - .48V5 - .40V14 - .52V22 - .07V8 + .14V15 + 96.7$$

(68% + 3.7%)
(95% ± 7.4%)

HORT FORMULA: (92.1% VARIANCE)

$$V3 = -1.02V6 - .57V5 - .45V14 - .35V22 + 101.5$$

(68% + 3.7%)
(95% ± 7.5%)

TABLE 1.6

MULTIPLE REGRESSION

V1 1975 ALP CITY VOTE (NOT ADJUSTED)

VARIABLE NUMBER	VARIABLE	VARIANCE EXPLAINED %	EXTRA VARIANCE EXPLAINED %	COEFFICIENT AND CONSTANT	PEARSON R
V4	Blue Collar Workers	89.5	89.5	+ .20	+ .95
V6	Upper White Collar Workers	91.1	1.6	-.77	-.93
V22	Matriculants	91.9	0.8	-.50	-.75
V14	65+ Year Olds	92.7	0.8	-.05	-.52
V5	Middle White Collar Workers	92.9	0.2	-.20	-.91
V8	Housing Trust tenants	93.1	0.2	+.07	+.56
V16	U.K.-born	93.2	0.1	+.06	+.29
V15	Overseas-born	93.2	0.1	+.09	+.55
				+64.3	

VLT FORMULA: (93.2% VARIANCE)

$$V1 = .20V4 - .77V6 - .50V22 - .05V14 - .20V5 + .07V8 + .06V16 + .09V15 + 64.3$$

HORT FORMULA: (91.1% VARIANCE)

$$(68\% + 3.6\%) \pm 7.2\% (95\%)$$

$$(68\% + 3.7\%) \pm 7.3\% (95\%)$$

$$V1 = .62V4 - .80V6 + 29$$

MULTIPLE REGRESSION

V7 - THE CITY 1973-75 ANTI-LABOR SWING : THE VOLATILE VOTER

VARIABLE NUMBER	VARIABLE	VARIANCE EXPLAINED %	EXTRA VARIANCE EXPLAINED %	COEFFICIENT AND CONSTANT	PEARSON R
V10	25-34 Year Olds	54.5	54.5	+ .22	+ .74
V8	Housing Trust tenants	61.3	6.8	-.04	-.29
V12	45-54 Year Olds	66.4	4.9	+ .43	-.45
V26	Short term residents	69.6	3.2	+ .15	+ .57
V6	Upper White Collar Workers	72	2.4	+ .03	-.02
V5	Middle White Collar Workers	72.7	0.7	+ .06	+ .10
V20	Catholics	74	1.2	+ .27	-.32
V17	Italians	77.7	3.7	-.34	-.30
V24	Schoolchildren	78.3	0.5	-.13	+ .17
V14	65+ Year Olds	79.4	1.1	-.17	-.50
				-15.5	

ULT. FORMULA: (79.4% VARIANCE)

$$V7 = +.22V10 - .04V8 + .43V12 + .15V26 + .03V6 + .06V5 + .27V20 - .34V17 - .13V24$$

SHORT FORMULA: (72% VARIANCE)

$$(68\% + 1.2\%) \quad (95\% \pm 2.4\%)$$

$$V7 = +.26V10 - .07V8 + .51V12 + .12V26 - .08V6 - 12.8$$

$$(68\% + 1.3\%) \quad (95\% \pm 2.5\%)$$

TABLE 1.8

MULTIPLE REGRESSION

V2 - THE 1973 ALP COUNTRY VOTE (NOT ADJUSTED)

VARIABLE NUMBER	VARIABLE	VARIANCE EXPLAINED %	EXTRA VARIANCE EXPLAINED %	COEFFICIENT AND CONSTANT	PEARSON R
V4	Blue Collar Workers (Urban)	76.4	76.4	+.88	+.87
V29	Lutheran	87.9	11.5	-.32	-.49
V11	Schoolchildren	92.5	4.6	+4.68	+.50
V5	Agricultural Workforce	94.8	2.2	-.53	-.84
V33	Short-term residents	96.0	1.2	-.43	+.64
V6	Blue Collar Workers (Rural)	98.3	2.2	+1.22	-.62
V8	Personal voters	99.8	1.5	-1.84	-.63
V13	Housing Trust tenants	99.9	0.1	-0.13	+.81
V22	Greek-born	99.9	0.04	-0.44	+.09
V12	Tertiary Students	100	0.001	-.31	-.25
				-70.3	

FULL FORMULA: (100% VARIANCE)

$$V2 = +.88V4 - .32V29 + 4.68V11 - 0.53V5 - 0.43V33 + 1.22V6 - 1.84V8 - 0.13V13 - 0.44V22 - .31V12 - 70.3$$

SHORT FORMULA: (94.8% VARIANCE)

$$(68\% \pm 0.06) \\ (95\% \pm 0.11)$$

$$V2 = +.32V4 - .0.41V29 + 3.13V11 - 0.52V5 - 22.6$$

$$(68\% \pm 5\%) \\ (95\% \pm 10\%)$$

MULTIPLE REGRESSION

V1 - 1975 ALP COUNTRY VOTE (NOT ADJUSTED)

VARIABLE NUMBER	VARIABLE	VARIANCE EXPLAINED %	EXTRA VARIANCE EXPLAINED %	COEFFICIENT AND CONSTANT	PEARSON R
V4	Blue Collar Workers (Urban)	86.6	86.6	+1.25	+.93
V8	Personal Vote	93.9	7.3	-2.49	-.67
V22	Greek-born	96.9	3.0	+1.98	+.08
V29	Lutherans	98.2	1.2	-.14	-.40
V13	Housing Trust tenants	98.9	0.7	+.34	+.82
V33	Short-term residents	99.3	0.4	-.28	+.65
V6	Blue Collar Workers (Rural)	99.6	0.3	+.50	-.72
V31	Presbyterians	99.8	0.2	-.12	+.23
V24	German-born	99.9	0.1	-5.26	+.75
V5	Agricultural Workforce	99.98	0.02	+.05	-.89
				-3.62	

ULL FORMULA: (99.9% VARIANCE)

$$V1 = +1.25V4 - 2.49V8 + 1.98V22 - .14V29 + .34V13 - .28V33 + .50V6 - .12V31 - 5.26V24 + .05V5 - 3.62$$

HORT FORMULA: (98.2% VARIANCE)

$$V1 = .99V4 - 2.61V8 + 2.54V22 - .21V29 + 1.8$$

(68% + 0.4%)
(95% ± 0.8%)

(68% + 2.4%)
(95% ± 4.9%)

MULTIPLE REGRESSION

V3 - THE COUNTRY 1973-75 ANTI-LABOR SWING : THE VOLATILE COUNTRY VOTER

VARIABLE NUMBER	VARIABLE	VARIANCE EXPLAINED %	EXTRA VARIANCE EXPLAINED %	COEFFICIENT AND CONSTANT	PEARSON R
V31	Presbyterians	57.9	57.9	+1.03	+.76
V7	Middle White Collar (Urban)	69.1	11.2	-6.64	-.04
V28	Church of England	75.1	6.0	+.91	+.31
V32	Agnostics	79.6	4.5	-1.61	+.06
V4	Blue Collar Workers (Urban)	86.4	6.8	+.88	-.02
V33	Short-term residents	90.5	4.1	-.16	+.09
V23	Italian-born	95.0	4.5	-4.83	+.13
V6	Blue Collar Workers (Rural)	95.9	0.9	+.16	+.22
V8	Personal Voters	97.6	1.7	-1.09	+.02
V17	45-54 Year Olds	99.5	1.9	+2.07	-.03
				-16.5	

FULL FORMULA: (99.5% VARIANCE)

$$V3 = +1.03V31 -6.64V7 +.91V28 -1.61V32 +.88V4 -.16V33 -4.83V23 +.16V6 -1.09V8 +2.07V17 -16.5$$

SHORT FORMULA: (95.0% VARIANCE)

(68% + 0.7%)
(95% ± 1.3%)

$$V3 = +.96V31 -4.41V7 +.98V28 -1.11V32 +.57V4 -.34V33 -2.52V23 +15.6$$

(68% + 1.5%)
(95% ± 3.0%)

MULTIPLE REGRESSION

V8 - THE 1975 COUNTRY PERSONAL VOTER

VARIABLE NUMBER	VARIABLE	VARIANCE EXPLAINED %	EXTRA VARIANCE EXPLAINED %	COEFFICIENT AND CONSTANT	PEARSON R
	(Variable deleted - explained by subsequent steps)	47.1	47.1		
V29	Lutherans	58.6	11.4	+ .32	+ .40
V12	Tertiary Students	65.4	6.8	-5.32	-.08
V9	Matriculants	69.3	4.0	+1.41	+ .09
V24	German Born	79.7	10.3	-11.75	-.26
V22	Greek Born	85.1	5.4	-.13	+ .30
V14	18-24 Year Olds	88.4	3.3	-4.24	-.34
V23	Italian Born	90.5	2.0	+ .32	-.44
V7	Middle White Collar (Urban)	92.8	2.4	+3.10	-.63
V19	65+ Year Olds	97.3	4.5	-2.00	+ .16
V13	Housing Trust Tenants	99.8	2.4	+ .06	-.44

+80.2

JLL FORMULA: (99.8% VARIANCE)

$$V8 = +.32V29 - 5.32V12 + 1.41V9 - 11.75V24 - .13V22 - 4.24V14 + .32V23 + 3.10V7 - 2.00V19 + .06V13 + 80.2$$

PORT FORMULA: (85.1% VARIANCE)

(68% \pm 0.2%)
(95% \pm 0.4%)

$$V8 = +.17V29 - 2.04V12 + 0.83V9 - 4.92V24 + 0.57V22 - .28$$

(68% \pm 1.0%)
(95% \pm 1.9%)

DISCUSSION

Tables 1.2 and 1.3:

Stability and the strength of the class-vote relationship emerge as the two factors dominating the 1973 and the 1975 ALP vote. It can clearly be seen that, in terms of descriptive power, the class-composition of a given area is almost as good an indicator of its likely future vote, as is its most recent vote:

Blue Collar Workers: City 1975 Vote \Rightarrow r of +.96
 City 1973 Vote : City 1975 Vote \Rightarrow r of +.98

Blue Collar (Urban): Country 1975 Vote \Rightarrow r of +.93
 Country 1973 Vote : Country 1975 Vote \Rightarrow r of +.96

It is rather difficult to make comparisons between the city and the country voters, but it appears that the country ALP vote is based partly on place of residence as well as class. The middle-class persons you would expect to see living in the larger country towns appear to be much more likely to support the Labor Party than middle-class persons living in the Adelaide metropolitan area.

On the other hand, working-class persons living in sparsely-settled rural areas appear to be much more likely to support the non-Labor parties than working-class persons living in the metropolitan area.

For example, the following persons who would tend to live in country towns are significantly more likely to support the Labor Party than people with similar jobs living in Adelaide:

- * Clerks
- * Catholics
- * Housing Trust tenants
- * Younger Persons (18-24 year olds in particular)
- * Overseas-born
- * Persons with above-average education.

And the following persons, who would tend to live in rural areas, are much less likely to support the Labor Party than people with similar jobs living in metropolitan Adelaide:

- * Mining and Quarry Workers
- * Farm Labourers, Fishermen
- * Older Persons
- * Persons with below-average education.

There seems to have been no significant change in the class base of support for Labor and non-Labor in either the city or the country between 1973 and 1975.

* * *

Table 1.4

The 1973-75 South Australian volatile voter presents a much more curious picture than the more stable class voter.

In the metropolitan area, the volatile voters were typically younger, married couples, of any class or education, who had just moved into their first home, with a very young family.

This stereotype represents an easily-identified target group which can be readily catered for with policy emphasis on such issues as interest rates, pre-school child care and the provision of services to outer, developing suburban areas.

In the country, the volatile voter is somewhat more difficult to identify from the Pearson r table. The only thing that can be said with confidence about swinging voters in the South Australian country areas is that they are probably Presbyterians, and they are not likely to be older than about 55 years.

Other than that, it is fair to infer from the Pearson r table that the country swinging voter shares some age and family status characteristics with the city swinging voter.

The country swinging voter will tend to be younger (18-24, or 35-44), and have a young family at school. He/she will tend to be Church of England or Catholic, and not Lutheran or Methodist. Country Tertiary Students also were a good deal less unsympathetic towards the Labor Party in 1975 (they swung less against us) than non-tertiary students of a similar age (18-24 year olds).

Stronger Labor areas swung more heavily against the Labor Party than the weaker Labor areas, indicating a merging of the Labor Voter/Swinging Voter stereotype in the country.

* * *

Table 1.5

The country personal voter can best be identified by class labels. He/she will tend to be found in areas dominated by Blue Collar Workers (Rural) and Agricultural Workers.

Personal voters will also tend to live in areas dominated by non-Labor voters, Middle White Collar Workers (Urban) and Catholics.

Regression Table 1.6

The 1973 ALP city vote - The ALP's 1973 performance in the city area was dominated by what could be called "repressive" factors. These factors (Upper and Middle-White Collar Workers, older persons, the better educated, Housing Trust tenants) served to hold the ALP vote down, at the same time as it was "buoyed up" by Overseas-born persons.

It is probably fair to conclude from this that the ALP city vote was near its upper limits in 1973. The first four demographic groups at the top of the regression table therefore represent the 1973 anti-Labor demographic coalition: the affluent, the elderly, the better educated, joined by - surprisingly enough - public housing renters.

The overseas-born represent the strongest base of demographic support for the Left in 1973.

Regression Table 1.7

The 1975 ALP city vote - Support for the Labor Party in Adelaide in 1975 came from Blue Collar Workers, Housing Trust tenants, U.K.-born and overseas-born persons. Groups hostile to the ALP in 1975 were Upper White Collar and Middle White Collar Workers, Matriculants and Aged Pensioners.

Housing Trust tenants can therefore be seen as a hostile group in 1973 and a pro-Labor group in 1975. Why was this so? Evidence presented below (and reconfirmed by subsequent national analyses) indicates that this is a key volatile group in Australian politics.

Regression Table 1.8

The 1973-75 Anti-Labor city swing - the Volatile Voter. The following groups swung against the Labor Party in 1975:

- | | | |
|----------------------------|---|--------------------------------|
| INCREASING
SIGNIFICANCE | ↑ | 1. 25-34 year olds |
| | | 2. 45-54 year olds |
| | | 3. Short-term residents |
| | | 4. Upper White Collar Workers |
| | | 5. Catholics |
| | | 6. Middle White Collar Workers |

The groups listed below either swung in favour of the Labor Party in 1975, or swung against the Party to a smaller degree:

- | | | |
|----------------------------|---|------------------------------------|
| INCREASING
SIGNIFICANCE | ↑ | 1. Housing Trust tenants |
| | | 2. Italians |
| | | 3. 65+ year olds |
| | | 4. (The parents of) Schoolchildren |

It is not possible to say from the above evidence if the above two lists describe: demographic groups which voted for or against the State ALP in 1975 for specific, issue-orientated reasons; potential long-term sources of electoral stability and volatility; or a combination of these two factors. This situation is discussed after consideration of the Regression Tables.

Regression Table 1.9

The 1973 ALP country vote - The 1973 ALP country vote was boosted by Blue Collar Workers (Urban), the parents of school-children, and Blue Collar Workers (Rural).

The country Labor vote was depressed by Lutherans, Agricultural Workers (mainly farmers), short-term residents, personal voters, Housing Trust tenants, Greeks and Tertiary students.

It is interesting to note that the country ALP vote - like the city ALP vote - is dominated by class factors. It is also interesting to note the increased significance of religion in country areas: the non-Labor country vote seems dependent to a large degree on the influence of the (conservative) Lutheran church and the communications' stranglehold it holds on some smaller country towns.

Regression Table 1.10

The 1975 ALP country vote - In the country in 1975 the Labor Party was supported by Blue Collar Workers (Urban), Greeks, Housing Trust tenants, Blue Collar Workers (Rural), and Agricultural Workers.

The Liberal Party was supported by Personal Voters, Lutherans, short-term residents, Presbyterians and Germans.

Several significant realignments therefore, can be seen to have taken place in the country between 1973 and 1975: firstly, the Labor Party was supported by three 1973 Liberal demographic blocs in the form of Greeks, Housing Trust tenants, and Agricultural Workers. I am quite frankly staggered at the thought that the ALP

received support - however minor - from a farmer-dominated group in 1975. The "cockies" therefore cannot be blamed for Labor's disastrous performance in the country in 1975. Who, then, was responsible?

Regression Table 1.11

The 1973-75 anti-Labor country swing - The country anti-Labor swing in 1975 was clearly dominated by religious factors - almost 70% of the variation in swing between seats could be attributed to religion - or lack of it. Presbyterians, and Church of England groups swung towards the Right, while Agnostics swung towards the Left.

Other groups to swing against the Labor Party were Blue Collar Workers (Urban), 45-54 year olds, and Blue Collar Workers (Rural).

Groups who swung towards the Labor Party were Middle White Collar Workers (Urban), Italians, Short-term residents, and personal voters.

In summary, it could perhaps be argued that the State Branch of the Labor Party in 1975 lost support in country areas among its traditional working-class supporters over issues that were presumably based on religious/moral judgments.

The regression prediction equation accompanying Table 6.5 also clearly illustrates the electoral volatility of the South-Eastern region of the State. With Presbyterianism the major indicator of volatility, Mount Gambier (19.8%) and Victoria (12.8%) are the two major strongholds of the Presbyterian faith in the South Australian country areas.

Regression Table 1.12

The 1975 country Personal Voter - The 1975 country personal voters were:

INCREASING
SIGNIFICANCE



1. Lutherans
2. Matriculants
3. Middle White Collar Workers (Urban)
4. Housing Trust Tenants
5. Italians

The 1975 country non-personal, or class-voters, were:

INCREASING
SIGNIFICANCE



1. Germans
2. Tertiary Students
3. Greeks
4. 65+ year olds
5. 18-24 year olds

The results contained in Table 1.12 confirmed that almost 100% of the variation in the personal vote between seats in 1975 could be attributed to variation in demographic characteristics between seats.

It seems that the (negative) personal votes obtained by (most Labor) country candidates in 1975 were largely pre-determined by demographic factors.

Several interesting points can also be made about the nature of the personal voter:

Germans, Greeks, Tertiary Students, the very old or the very young are quite inflexible class-voters and are not responsive to appeals to alter their "natural" class vote through personal loyalties.

Labor's country candidates should concentrate their personal vote drive on: Lutherans, the better-educated, town-based middle-white collar workers, Housing Trust tenants, and Italians.

These persons apparently felt content to express their class loyalty via their upper-house vote, and their personal loyalty through their lower house vote.

*

*

*

It is now a relatively straight forward task to construct plausible explanations for South Australian electoral behaviour between 1973-75, given the above evidence. One explanation could run as follows:

PRO-LIBERAL ISSUES

25-34 year olds and short-term residents (young couples in the first few years of mortgage repayments) could have blamed the State Government for high interest rates, and correspondingly-high mortgage payments.

Upper-White Collar Workers, Middle White Collar Workers and 45-54 year olds may have blamed the State Government for erosion of their family savings through inflation.

Catholics may have been more concerned than most persons about the anti-Labor moral issues promoted by the Opposition in 1975.

PRO-LABOR STABILITY

Italians and 65+ year olds could be regarded as relatively stable voters, perhaps more isolated than most members of the community from information networks and contemporary media bias.

PRO-LABOR ISSUES

The parents of schoolchildren may have swung towards the Labor Party because of Federal Labor's education policies. I am unsure of the motivation behind a pro-Labor swing in Housing Trust tenants. Irrespective of the debate over why the so-called volatile groups behaved as they did in 1975, two points can be made with some certainty:

1. Demographic groups which voted for the Labor Party in 1973 and against the Labor Party in 1975 will have more potential to swing back to the Labor Party at the next State election than other demographic groups in the community.
2. These volatile groups should therefore form the prime target for Labor's campaign strategy across the Adelaide metropolitan area and within the 33 new metropolitan electorates.

The Regression Equations

The evidence described above for the Regression Tables is presented in more concise numerical form in the Regression Prediction Equations. The prediction equations enable us to calculate the following:

1. The ALP vote, on 1973 demographic alignments ("on 1973 figures"), for any area in S.Aust from the smallest 100-home collectors' district to the entire metropolitan area.
2. The ALP vote, on 1975 demographic alignments, for any area in South Australia.
3. The 1973-75 swing (and ~~potential~~ 1975-78 swing) for any area in South Australia.
4. Long-term changes to the ALP vote or electoral volatility of any area in South Australia.

For example, recent housing developments in the western suburbs are suspected to have had some adverse impact on the ALP vote in Semaphore, Henley Beach, Hanson and Morphett. This factor, often impossible to gauge using polling booth data, can be measured in these five seats (prior to the scheduled 1978 election) using 1971 and 1976 census data.

Similarly, we can calculate the likely impact of any sampled pro-Labor swing between now and the next State elections.

For example, the five metropolitan divisions sampled by ANOP recently for the ALP registered an average pro-Labor swing of some 8 percent. But the seats sampled were about 1.3 times as volatile as the metropolitan average in 1975 (using the formula for V7 in Table 1.8). Therefore, the true metropolitan pro-Labor swing detected by the sample may not have been 8 percent but 8 percent divided by 1.3 - about 6 percent. See the Appendix to Project One for a more detailed interpretation of this survey.

This six percent can in turn be applied to all new metropolitan seats - according to their 1975 volatility - to produce a predicted ALP 1978 vote for each seat. For an average swing of 6 percent the individual seat swings could range from about 4 percent in Bragg, to about 12% in Newland and Mawson.

Also, we can produce an estimate of the volatility of each collectors' district within our new marginal seats, to direct candidates to areas where campaign efforts will tend to bring the greatest rewards. Obviously, it does not make much sense to have candidates doorknocking pockets of older areas in new marginal seats when they could be concentrating on more volatile suburbs.

The South Australian Project in retrospect and its contribution to future national research.

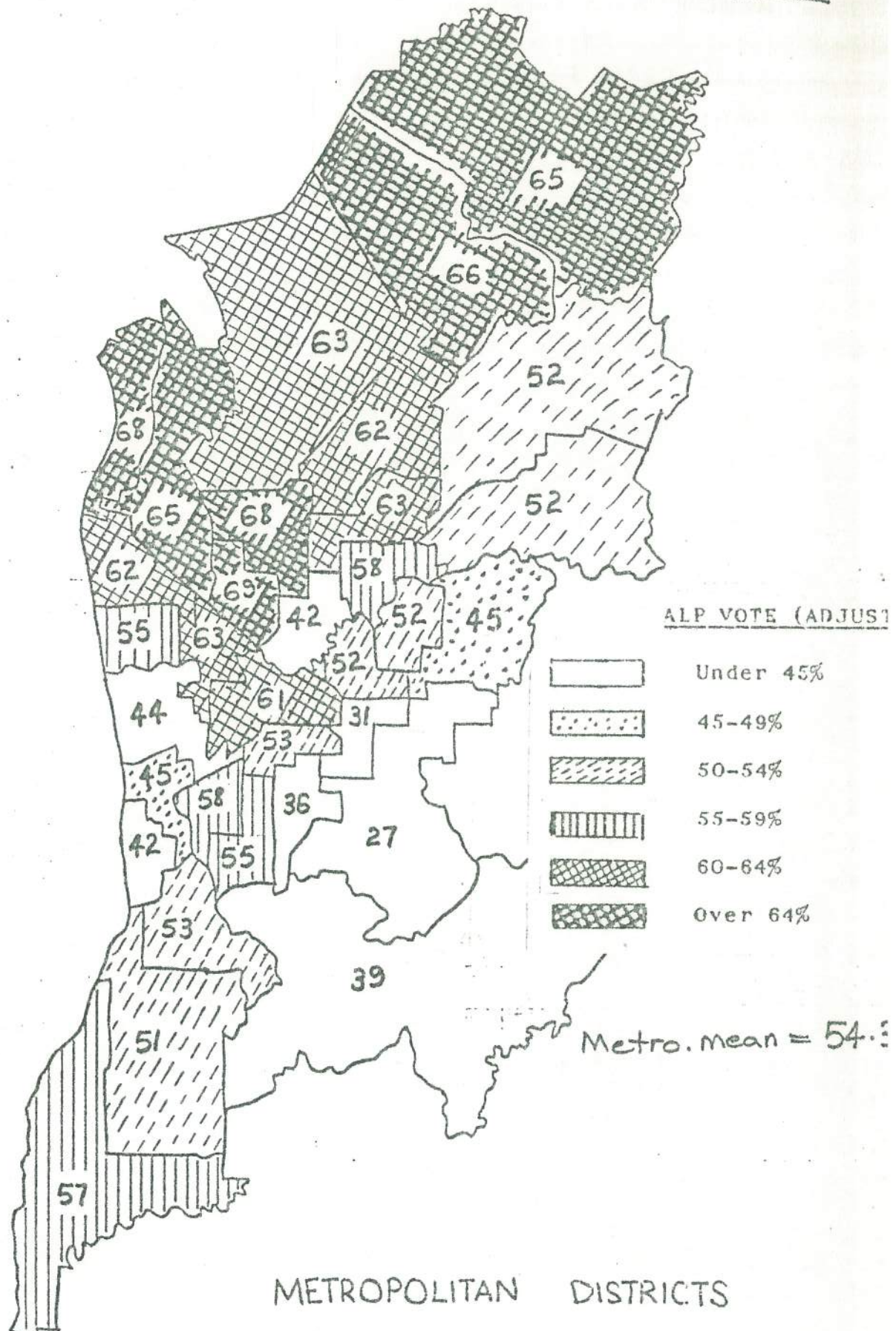
This section of the report deals retrospectively with the major implications of the S.A. Project and its implications for the national projects which followed. The major contributions of this project are discussed in turn below:

1. The relationship between occupational class (as determined by the census) and the level of support for Labor (as measured by the Two Party Preferred vote).

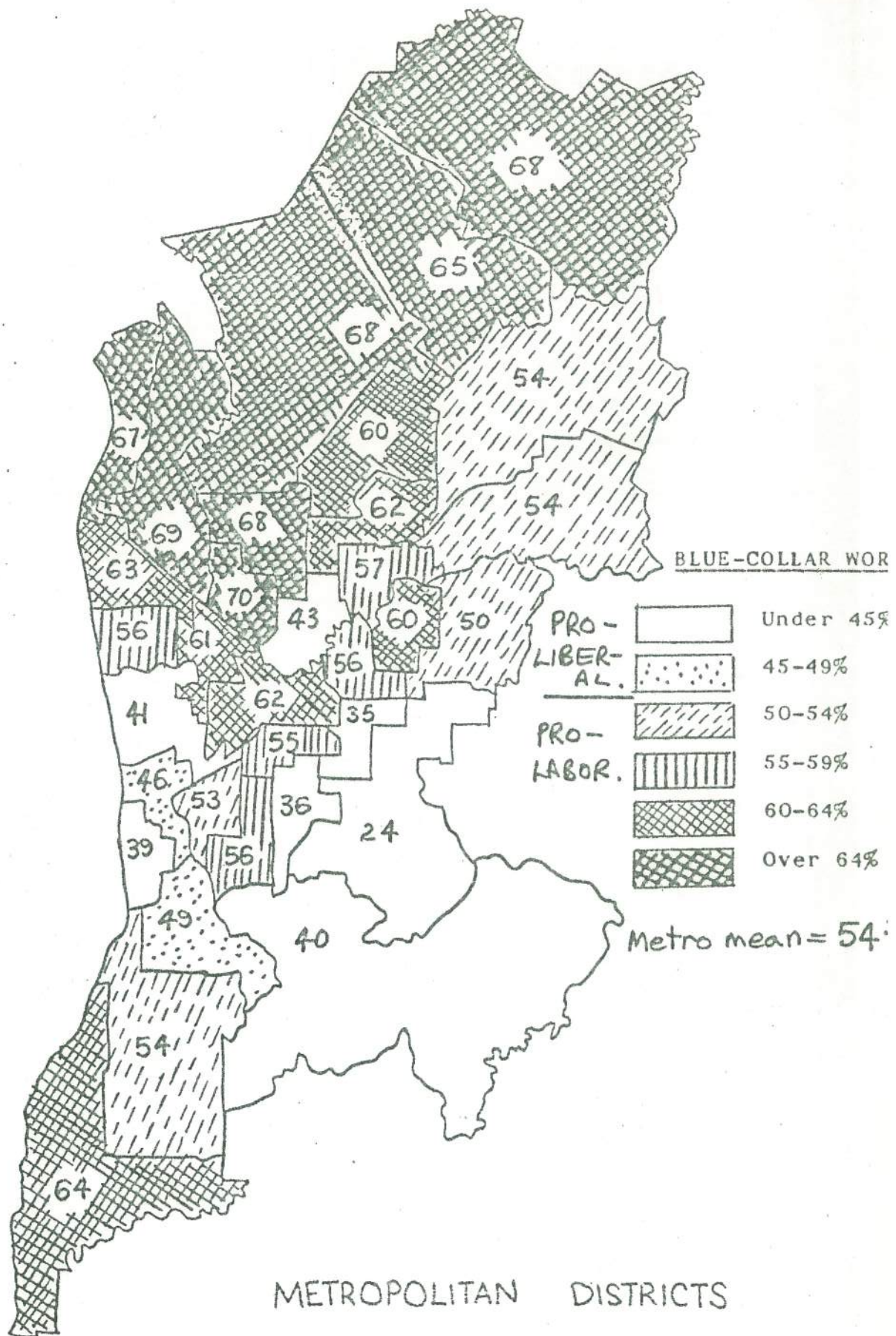
The relationship between Class and vote in South Australia in the early seventies was extraordinarily powerful. Given the efforts of some academic "experts" in the late seventies to divorce class from debate about electoral behaviour, the S.A. evidence is particularly striking. Correlations of .96 (Blue Collar Workers and the 1975 city vote) and .93 (Blue Collar Workers and the 1973 city vote) are not obtained very often in Social Science.

In both 1973 and 1975 Occupational Class factors explained about 90 percent of the variance in the Labor vote in city seats.

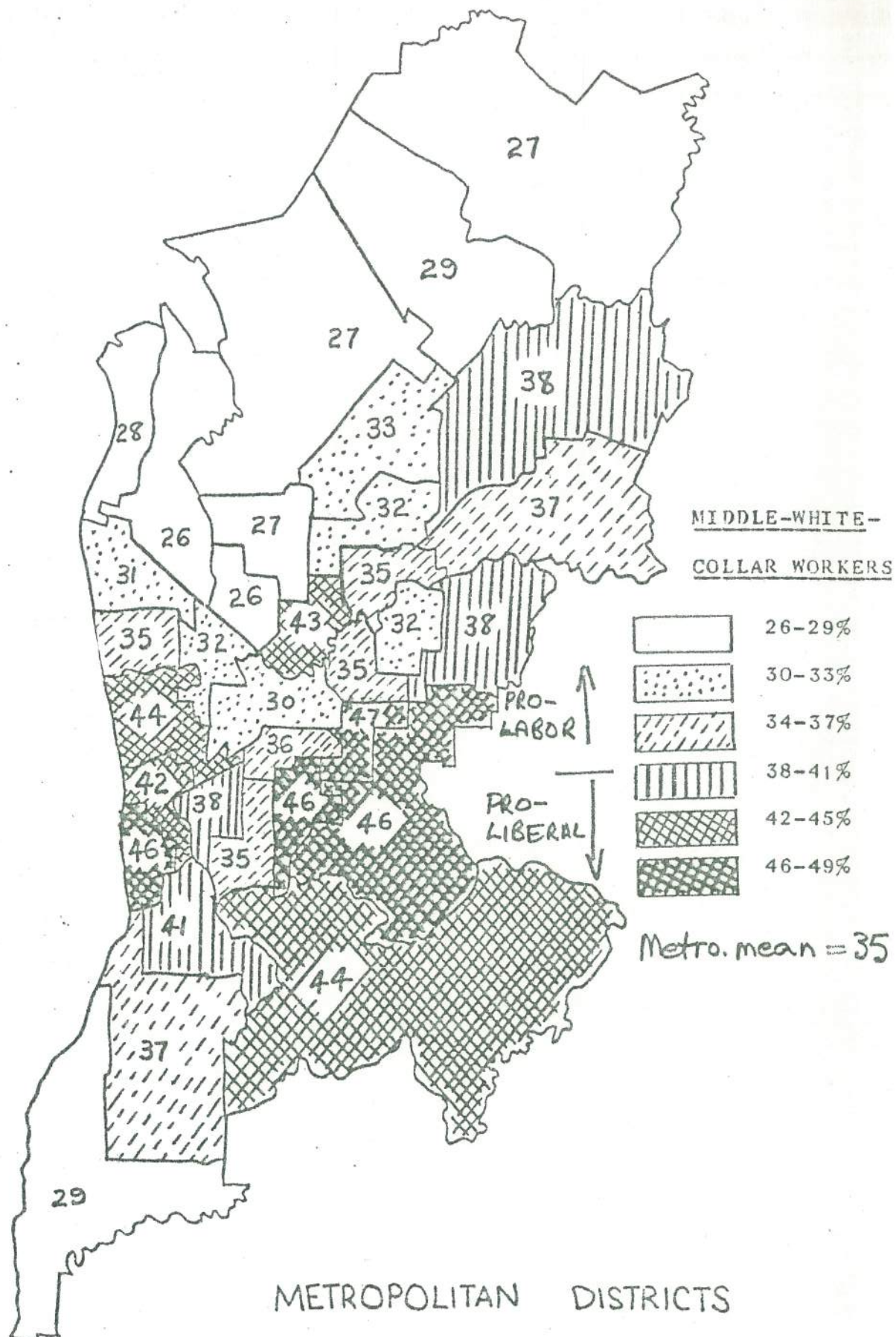
Visual evidence of this exceptionally-strong relationship is also provided in maps I prepared at the time showing the 1975 ALP 2PP votes for city seats and the 1971 Census figures for the same seats showing the distribution of Blue Collar Workers, Middle White Collar Workers and Upper White Collar Workers (Maps 1, 2, 3 and 4).



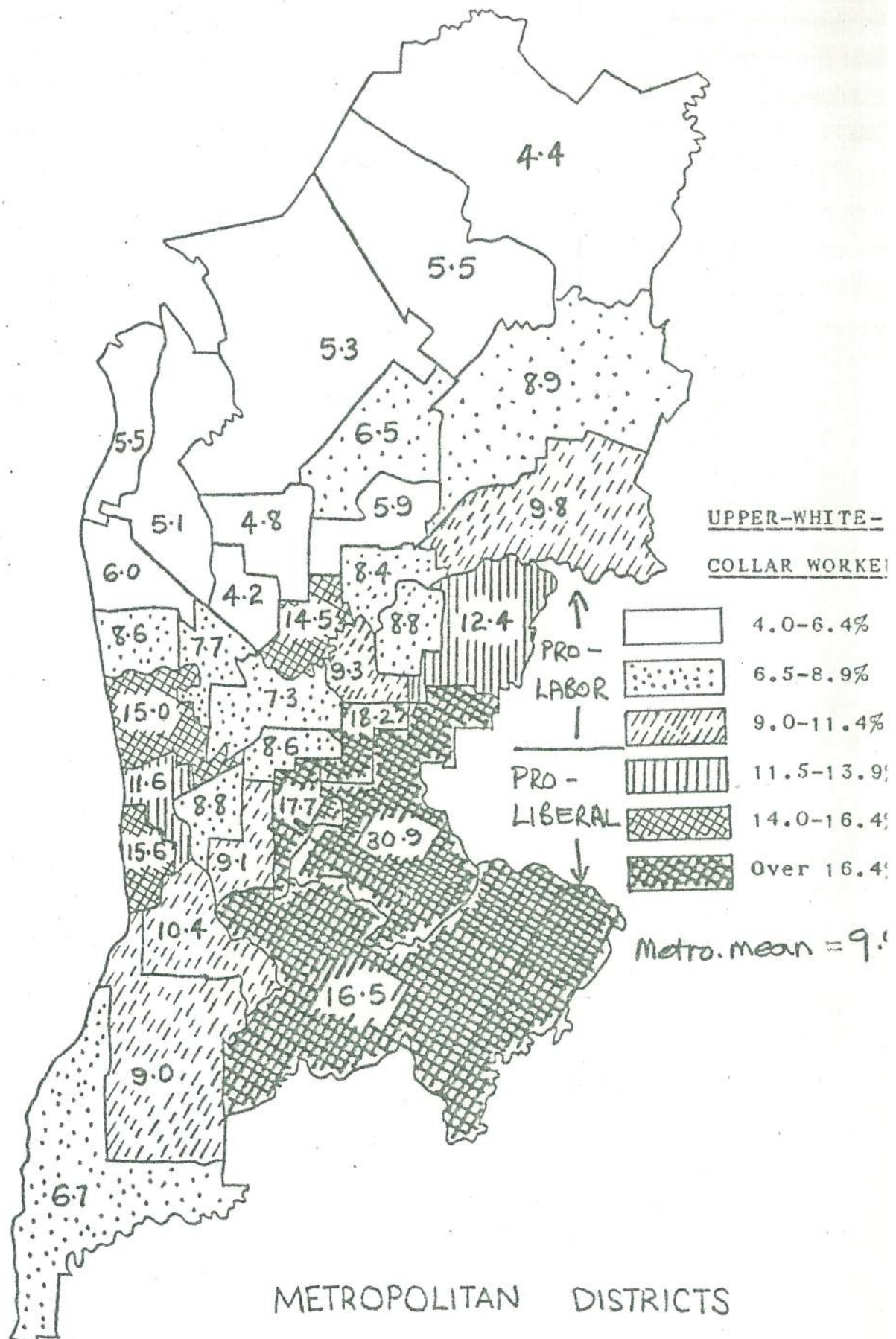
Map shows: 1975 ALP two-party-preferred vote adjusted for the loss of personal vote and donkey vote (where applicable)



Map Shows: Employed persons in Census Occupation Codes 22-73,
as a percentage of the total work force.



Map Shows: Employed persons in Census Occupation Codes 6-8, 11-12, 16-21, as a percentage of the total work f



Map Shows: Employed persons in Census Occupation Codes 1-5, 13-15, as a percentage of the total work force.

The maps clearly show the 1975 Labor voters concentrated in the central and north-western industrial suburbs of Adelaide, with the "dress circle" suburbs to the south-east, and the more marginal seats centred on a south-west to north-east diagonal.

The Blue Collar Workers variable in 1971 Census followed this pattern extremely closely, with an almost one-to-one relationship between the occupational class of the workforce (both sexes included) and the 2PP Labor vote.

Both the Upper White Collar Workers and the Middle White Collar Workers were distributed in a reciprocal fashion. The strong link between the Middle White Collar Workers and the Upper White Collar workers can also be seen.

No political observer, no matter how biased towards a class-free analysis of (South Australian) electoral behaviour, could seriously argue ^{against} this empirical evidence provided by the Pearson r figures, the variance figures in the Regression equations, and the visual evidence of the electorate maps.

Any future national analysis therefore had to place heavy emphasis on the role of occupational class and the nature of the workforce.

2. The strength of 1971 census data as a predictor of 1975 electoral support, compared to the strength of the 1973 vote as a predictor of 1975 electoral support.

Despite the fact that the 1971 census data was four years old at the time of the 1975 election, the Pearson r was .96, compared to a Pearson r of .98 between the 1973 election result and the 1975 vote (for city seats).

Obviously, then, it became theoretically possible to base future analyses of election results on data that was up to four or five years "out of date". This conclusion was based at the time on the empirical evidence cited above, which indicated that demographic turnover in electorates had little impact on the class composition of those electorates. The sons and daughters of older voters tended towards the same sorts of employment and voting patterns as their parents, and new settlers in any area (in existing homes or newly-built homes) tended to have similar jobs and voting allegiances as the people they either replaced, or in whose suburb they had built their new home. A Phrase used by biologists and political geographers provides a useful analogy: "The cells come and go, but the organism remains the same".

3. The definition and measurement of electoral instability and its relationship to demographic variables, particularly age and occupational class.

The anti-Labor South Australian state swing of 1973-75 reflected the general decline in popularity of the then Federal Labor Government. Just prior to the 1975 state election - in the midst of the "Loans Affair revelations" (to use the media term) state market research indicated that support for Federal Labor had plummeted by 20 percent in some state seats. Of this 20 percent, all but about five percent were still prepared to maintain support for the relatively-popular State Labor Government.




As media speculation continued about the extent of the "rub-off" of the Federal Labor's unpopularity, the revelations continued apace, with each day seeing another compromising telegram or alleged "deal" with the ubiquitous Tirath Khemlani. The State Liberals quite naturally capitalised on the Federal Government's unpopularity by basing their state campaign on federal issues and building up visits by Federal Liberals to the State to "assist" the state campaign.

About ten days before the election day the then Premier Don Dunstan appeared in special campaign advertisements, effectively disowning his Federal colleagues ("My Government is being smeared - and it hurts!" was one headline from a newspaper advertisement).

The Labor Prime Minister's response was to reconvene Federal Parliament for a special one-day sitting on the last day before the State media blackout to debate the Loans Affair. His announcement to this effect followed quickly on from Dunstan's actions and meant that the last week of the campaign was devoted to media speculation firstly about the contents of the debate and then about the results of the debate. State issues were completely overwhelmed.

I feel that for the above reasons, the South Australian State election in 1975 was more of a Federal by-election than a state election, and the 1973-75 swing therefore measured the general weakness of pro-Labor identification across the South Australian electorate, rather than any lack of popularity of the State Labor Government over specific state issues. In short, the results of the State 1973-75 swing had significant implications for Federal Labor.

To allow for a more detailed look then at the 1973-75 State swing, I first list all the correlations for the city seats and the swing.

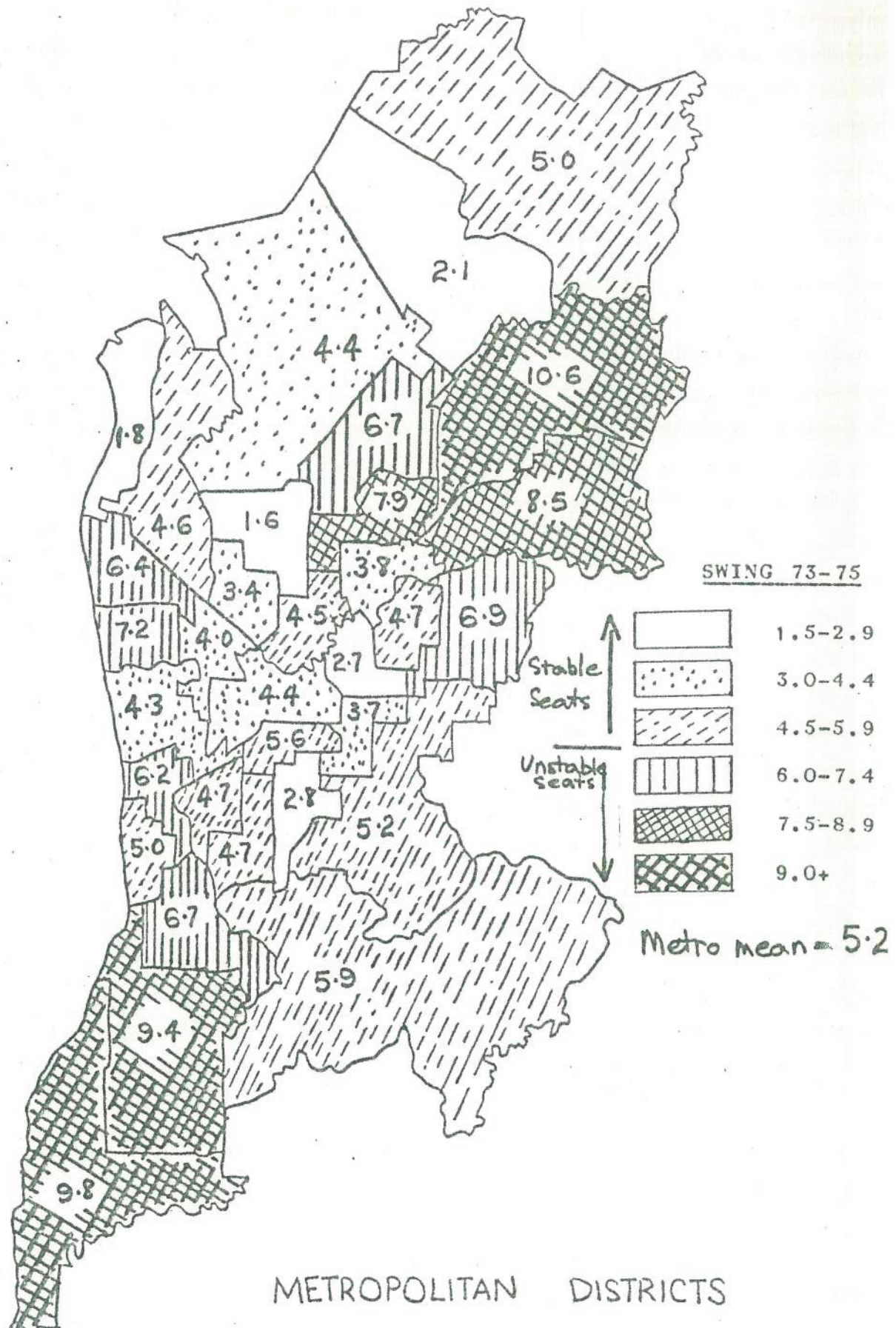
Rank	Variable			
1	25-34 year age group	.74		INCREASING INSTABILITY AND ANTI-LABOR SWING
2	Short-term residence	.58		
3	Pre-school children	.56		
4	British-Born	.51		
5	35-44 year age group	.41		
6	Church of England	.40		
<hr/>				
7	Overseas-Born	.29		NOT SIGNIFICANT TO .05
8	Agnostics	.28		
9	Matriculants	.17		
10	School Students	.17		
11	Middle white collar workers	.12		
12	Upper white collar workers	0.0		
13	Blue collar workers	-0.07		
14	Greek-born	-.28		
15	Italian-born	-.30		
16	Housing Trust tenants	-.31		
17	Catholics	-.32		
<hr/>				
18	45-54 year age group	-.46		INCREASING STABILITY AND PRO-LABOR SWING
19	65+ age group	-.49		
20	18-24 year age group	-.50		
21	Full-time students	-.57		
22	55-64 year age group	-.58		

The most outstanding fact to be noted from the listing of the swing correlates is that every one of the six age groupings 18 and over is significantly correlated with swing, either positively or negatively. It should also be noted that the three class variables provide the three most insignificant correlations.

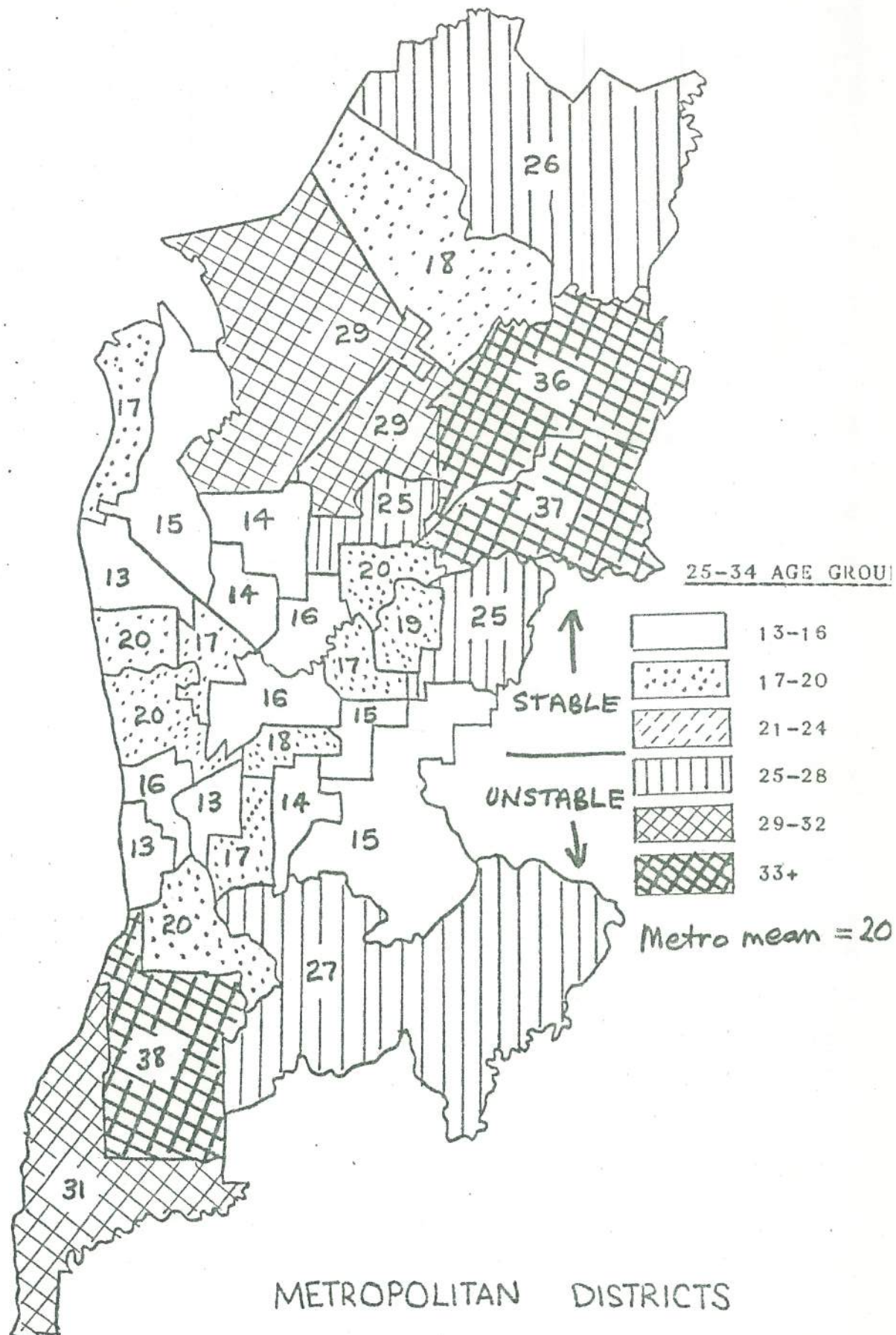
The importance of age, rather than occupational class or other demographic groups, as a predictor of swing is confirmed by the 1973-75 Regression Table 1.8.

Here we can see that age variables explained 77 percent of the total explained variance, class variables explained four percent, and other factors (public housing, mobility, religion, ethnicity etc) explained 19 percent.

This trend can also be highlighted visually by the simple comparison of political and demographic maps for metropolitan Adelaide. (See Maps 5 and 6 below)



Map Shows: Anti-Labor swing 1973-1975, based on two-party preferred vote, adjusted for loss or gain of personal vote and donkey vote (where applicable)



Map Shows: Persons aged 25-34 as a percentage of total persons aged 18 and over.

Map 5 shows the anti-Labor swing between 1973 and 1975 on a 2PP basis. It should be noted that in the South Australian project I listed the swing as being towards the non-Labor parties. A positive swing figure therefore represents a swing against Labor. (The reverse was done in subsequent projects).

Map 5 clearly identifies the focus of major swings as the north-east and south-west suburbs.

Map 6 shows with equal clarity that the 35-34 year olds (the best predictor of swing from the Pearson r tables and the Regression Tables) are also located in the same area.

At the time of writing the South Australian project I came to a number of conclusions based on the above results and other known aspects of electoral behaviour. The original text and diagrams are reproduced below:

I have not altered the content (or the somewhat more precocious style of writing) because this original model was not amended significantly by subsequent National analysis.

A Working Model - Background: Political scientists have been trying for decades to satisfactorily explain the concept of swing between elections. Most get bogged down in the preliminaries of trying to explain why a seat with a 70% Labor vote usually experiences a similar swing during a given period to a 51% Labor seat, or to a 30% Labor seat.

The logic usually employed is as follows :

"In given election campaign, with a given anti-Labor (say) mood across the country, persons who normally vote Labor are going to change their minds and vote Liberal. Let us assume that 5% of Labor supporters in all electorates change their mind and vote Liberal. That would give us an anti-Labor swing of 3.5% in the 70% Labor seat, 2.55% in the 51% Labor seat, and only 1.5% in the 30% Labor seat.

"But the swing was actually 2.5% across all seats .. so .. ".

Usually what follows is a series of intellectual acrobatics and flights of fantasy that succeed only in embarrassing the discerning reader. In a delightful piece of understatement on this subject two political scientists David Butler and Donald Stokes admit that "... we feel obliged to a quite exceptional degree to make clear the slenderness of the empirical foundations of our findings" (Political Change in Britain, "The Sources of Uniform Swing", page 377).

Another source of error and confusion concerning the swinging voter appears to have arisen from the observations of political journalists and, to a lesser extent, market researchers.

The "swinging voter" has been described by journalists as intelligent, perceptive, well educated and middle class. This stereotype of the swinging voter has had a disturbing impact on market research which tends to find out things people like to think are true, rather than what is actually true. Most people like to be thought of as intelligent, rational, well educated, and middle class. It is therefore flattering to the respondent's ego to be considered a swinging voter. Unfortunately the evidence indicates that Middle-Class persons vote Liberal, and that the better-educated members of the community (those educated to matriculation standard) also vote Liberal. Better educated, middle-class persons, as a group, have no apparent links with electoral instability.

The Model - Explanation: Figure 1 (page 527) shows the strength of the correlation between swing and age groups, across the six specified age distributions. Figure 2 (page 528) goes one step further and outlines a visual interpretation of the model. The figures shown on the vertical axis of Figure 2 are given only as an aid to a clearer understanding of the model. They are not based on factual observation.

Contrary to what one would expect, the 18-24 year olds appear to be quite stable. A portion of this apparent stability would no doubt be due to the fact that this age group tends to live in older, more stable areas of the city. However, by no means all of the 18-24 year olds' apparent stability can be explained by residential patterns.

The model appears to fit two universally-acknowledged trends in political behaviour which I have not as yet seen integrated into a single model. The two trends are :

- 1) Very young voters are likely to vote in the same manner as their parents and this tendency diminishes with age.

QUOTE: "Partisanship (the sharing of parents' party preference) over the individual's lifetime has some of the quality of a photographic reproduction that deteriorates with time: it is a fairly sharp copy of the parents' original at the beginning of political awareness, but over the years it becomes somewhat blurred, although remaining easily recognisable" (Butler and Stokes, page 68).

- 2) Habit is a strong stabilising influence on electoral volatility, which takes some time to establish itself, but, unlike parental influence, it increases in strength with age.

QUOTE: "With the aging of the voter, the relatively plastic attitudes of youth tend to harden and the acquired habits of the early voting years begin to become more deeply fixed" (Butler and Stokes, page 78).

FIGURE 1: STRENGTH OF
CORRELATION BETWEEN
SWING AND AGE
ACROSS AGE GROUPS

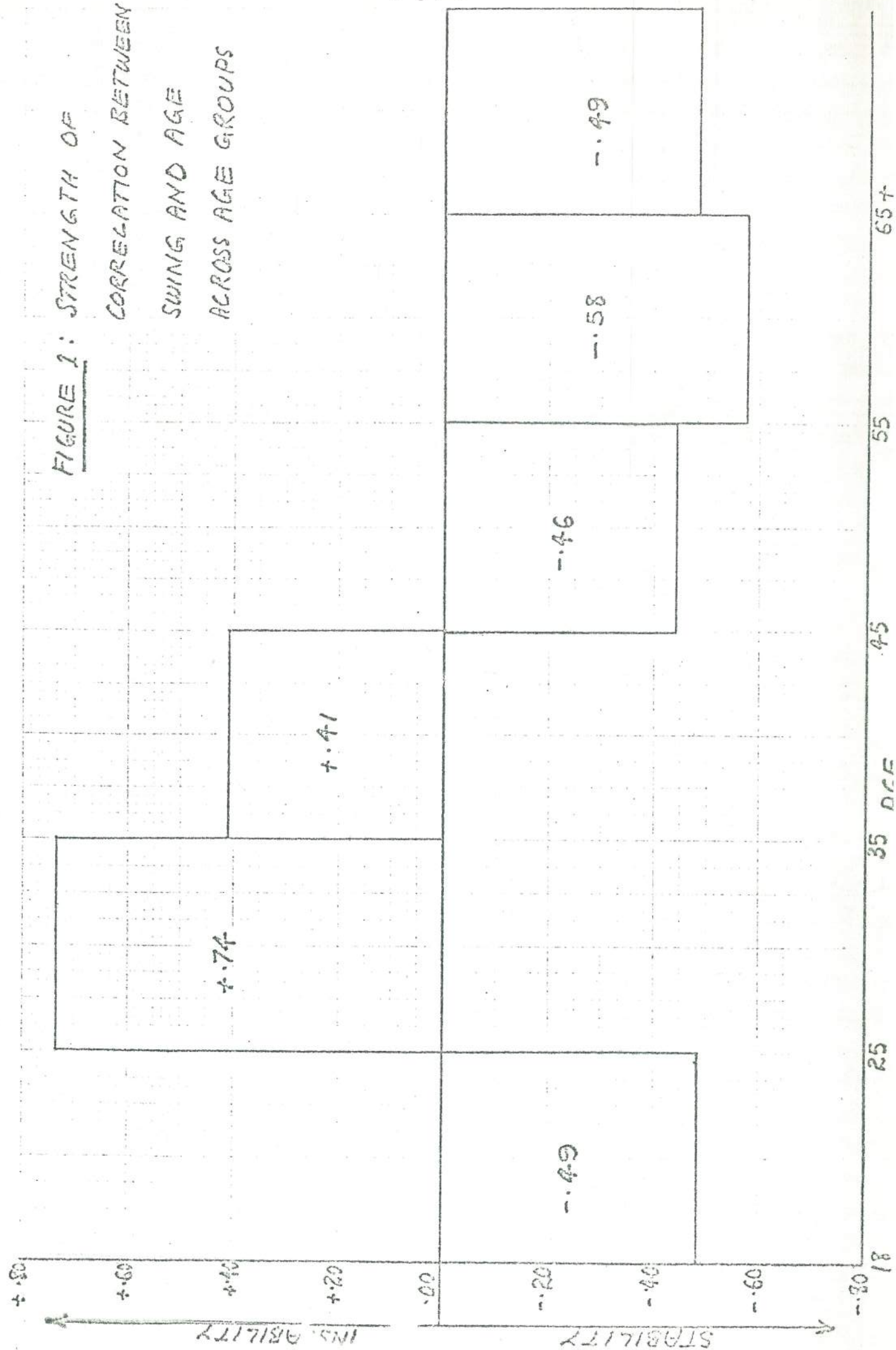
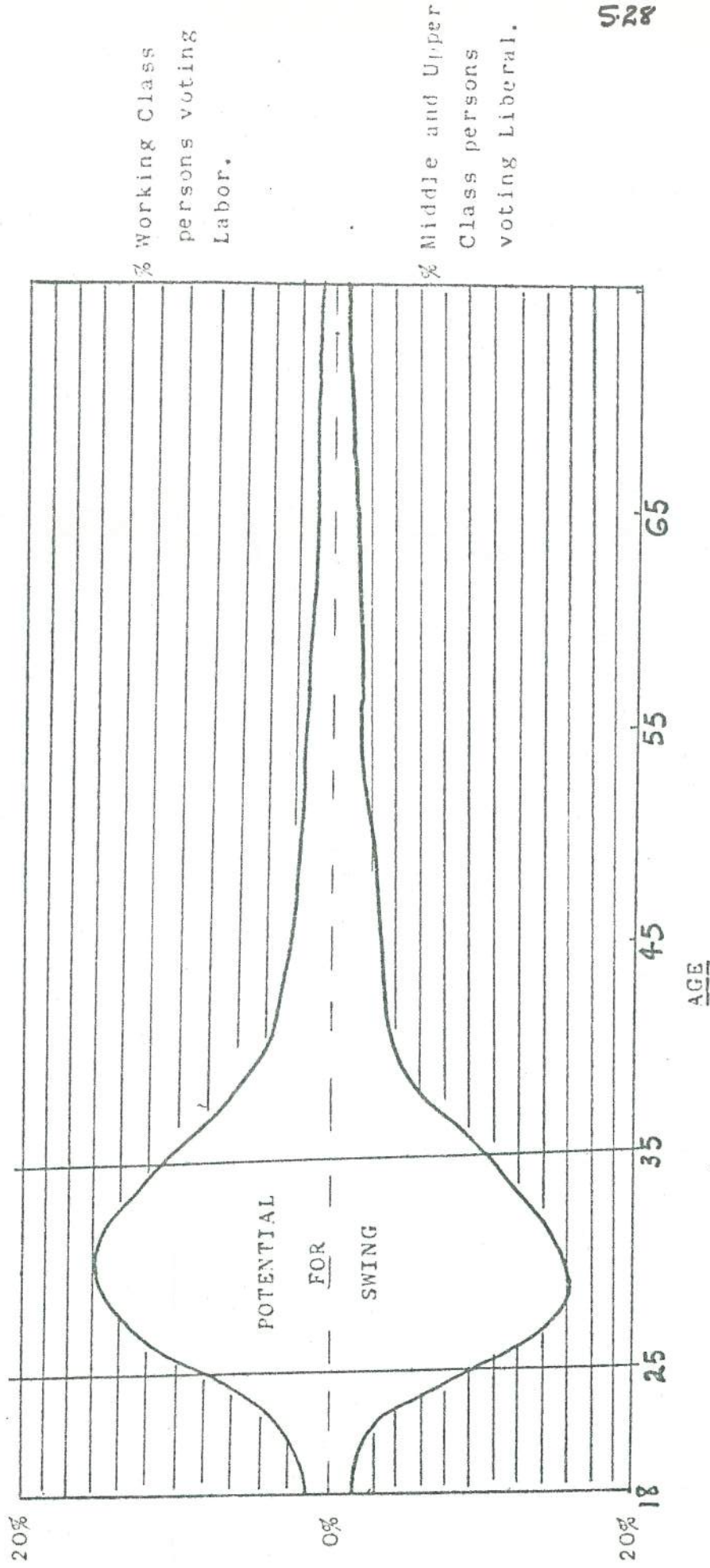


FIGURE 2*

STRENGTH OF CLASS-VOTE RELATIONSHIP ACROSS AGE GROUPS



* Note: This is a visual interpretation of the swinging voter model, based partly on Figure 1, but also based to a large degree on assumptions outlined in the text. It has been drawn to assist the reader, rather than state known facts.

So it appears that on the available evidence, the 25-34 year old has experienced a "blurring" of the parent-vote relationship, and that this blurring has not yet had time to become polarised by habit. At this stage in life the voter appears to be the most vulnerable to perceived prevailing moods for electoral change.

So, while class may be the best predictor of vote, age seems to be the best predictor of instability of this class-vote relationship.

Summary: The South Australian Research made a number of major contributions to future research.

Firstly, the methodology worked. The statistical marriage of political data (the Two Party Preferred Vote) and Demographic data (from the Census) through Correlation and Multiple Regression techniques produced results which "made sense". They interpreted given events and used this interpretation to make useful predictions about future behaviour.

The project highlighted the usefulness of class as a predictor of the ALP vote; it validated the use of Census data up to four years old as a predictor of the ALP vote; it reaffirmed the idea of a Two-Party Preferred Swing as a useful measure of electoral instability, but perhaps most importantly, it introduced the concept of a second major demographic variable - age- as a predictor of swing.

My "horizontal turnip" figure linking vote, swing, class and age (Figure 2) was crude, but it highlighted the essential finding of the SA Project: if the long-run average vote in a given seat is determined by class factors, and the variation in this long-run vote (its volatility) is determined by factors other than class (especially age), then whether this given seat is won or lost by Labor depends not simply on the marginal nature of the seat's previous vote, but also on the distribution of the key volatile groups (especially age groups) in that seat. Put more simply, a 60 percent Liberal seat which is twice as volatile as a 56 percent Liberal seat will, *ceteris paribus*, need a smaller average swing to be won by Labor. Volatile voters are like any other demographic group in the community; they are distributed across electorates in a standard-normal-curve fashion. The range of swing is predetermined by this distribution and it is this range of swing which, in a close contest, determines which party wins Government.

APPENDIX TO PROJECT ONE

I have included as an appendix to project one five documents. They are:

- A. A summary of the theory of the statistical techniques of analysis used in project one. This summary comprises extracts from a book entitled Statistical Package for the Social Sciences and deals with Pearson Correlations, Multiple Regression Analysis and Partial Correlation sub-programs.
- B. The demographic and political variables used in the S.A. analysis for all S.A. electorates.
- C. The methodology by which the personal vote scale was established for all S.A. electorates. The donkey vote formula was somewhat conjectural and in any event not a large component of the final scale. The number one score in Norwood was obtained by the then S.A. Premier Don Dunstan and the number forty-seven score of minus 23.2 percent was obtained by the unfortunate candidate for Pirie who was endorsed by a central pre-selection system still used (in a marginally diluted form) by the S.A. Branch. The endorsed candidate enjoyed somewhat less than the unanimous support of the local Port Pirie community and the "safe Labor" seat of Pirie as a result was won by an independent Labor candidate who had lost the pre-selection and was later re-admitted to the A.L.P. as the sitting member.
- D. The New Rocky River electorate. This short paper proved the usefulness of the regression equations to analyse long-term rises and falls in Labor support due to demographic trends in specific areas. It also highlighted the accuracy of the Regression Equations as predictors of future votes in certain circumstances where exogenous factors are minimised. The 1977 vote in Rocky River was in fact 38.8 percent, very close to the pre-election prediction of 38.0 percent. At the time I made this prediction (in late 1976 or early 1977) the conventional wisdom in some Labor circles was that the Labor candidate for Rocky River had a "good chance" of victory.

2

E. This document prepared for the S.A. Cabinet Campaign Committee provided an analysis of the then recent A.N.O.P. survey within the theoretical framework of my demographic analysis. It has never been my contention that my statistical analyses can completely take the place of attitudinal analyses; rather I argue in this paper that attitudinal analyses of electoral behaviour are subject to potentially large errors due to the infidelity of respondents and the difficulty of selecting a sample which is representative not just of the previous vote and the major geographical regions, but which is also representative of long-run swinging voters.

The document was completed well before the 1977 elections and I have prepared the following retrospective comments on the 1977 elections to underscore the validity and the accuracy of the analysis and its usefulness as a tool of long-term campaign strategy.

Page 1. I argued that the A.N.O.P. sample contained a disproportionate concentration of swinging voters and, with a general swing back to Labor, it therefore provided an overly optimistic estimate of Labor's state-wide support. This was in fact shown to be the case in 1977.

Page 2. The right-hand column of Table 2 showed my predictions of the 1977 vote based on the 1976 A.N.O.P. survey (adjusted downwards because of the volatility factor outlined above). My predictions, the A.N.O.P. 1976 results, and the actual 1977 results, are listed below:

	1976 A.N.O.P. 2 P.P.	1976 J.B. 2 P.P.	1977 Actual 2 P.P.
City	60.1	58.7	58.2
Country	42.9	40.5	40.9
State	55.9	53.5	53.2

Page 3. Paragraph one talks about the interaction of variability

3

of swing with previous vote. In fact in 1979 Labor did lose the "safe" seat of Todd and yet retained the "marginal" seat of Hartley.

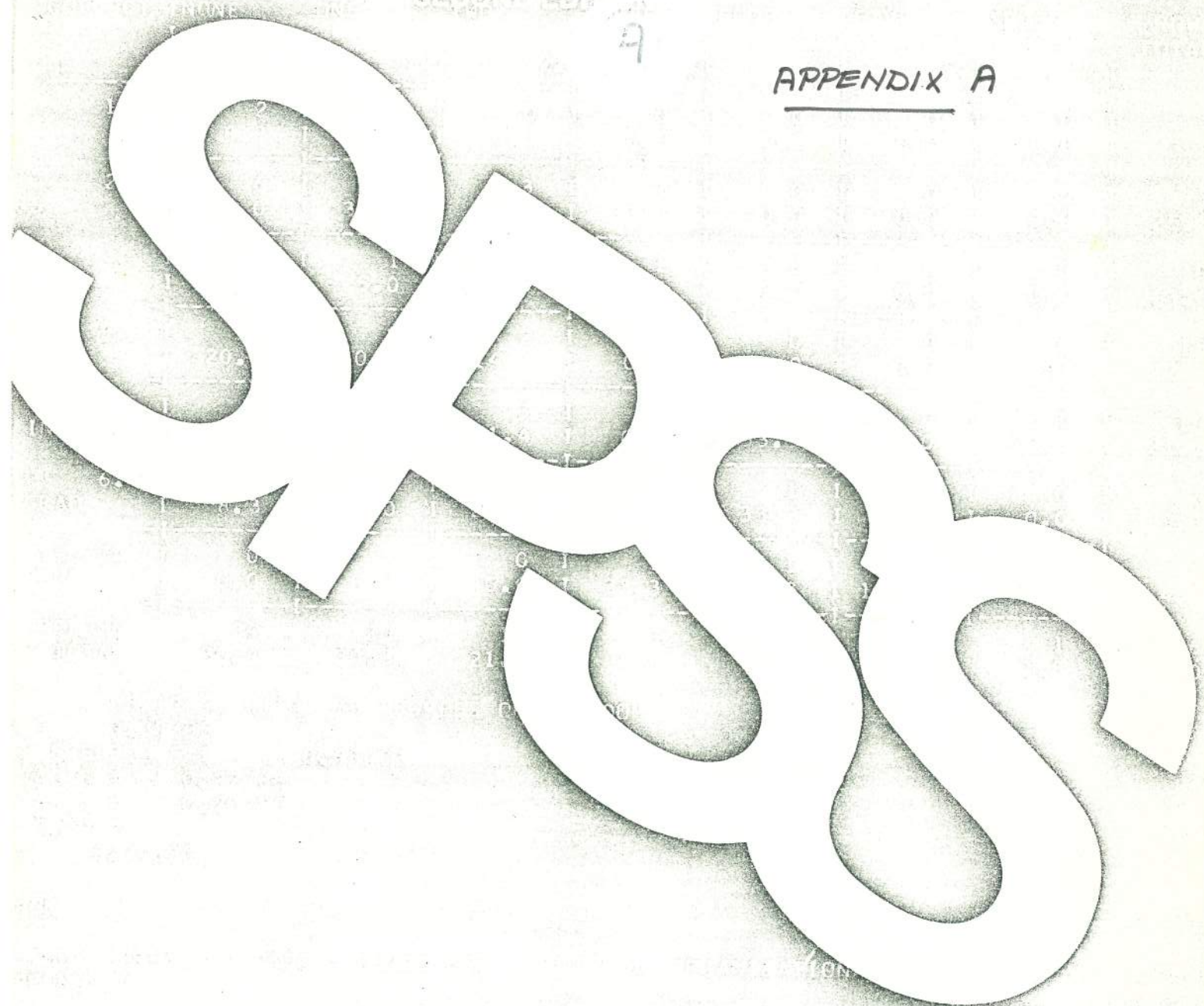
Table 4 and paragraph three relate to the predictions of seats to be won in 1977. In accordance with the 95 percent confidence qualification, Labor in fact won Morphett and lost Coles.

Page 4, paragraph three explains the only other error in my forecast: Labor failed to win Mt. Gambier in 1977. I believe the reason for this was the error involved in attitudinal sampling which does not measure the personal influence of sitting members, especially in small State Country seats.

The remaining pages of the document deal with the interaction of the A.N.O.P. attitudinal survey and my own S.A. model to set guidelines for the 1977 strategy. This is the sort of work which should be done for the National elections scheduled for 1983. By and large, the strategy employed by the Cabinet Campaign Committee (Premier Don Dunstan, Deputy Premier Des Corcoran, Mines and Energy Minister Hugh Hudson, Local Government Minister and former State A.L.P. Secretary Geoff Virgo, State A.L.P. Secretary Howard O'Neill and the Premier's Executive Assistant Rob Dempsey) followed the guidelines in document 4 and other research papers, and the swings were obtained in the required areas, together with a healthy State vote.

Perhaps it doesn't need to be said, but the 1979 S.A. campaign appeared to be the very antithesis of the above-mentioned 1977 campaign and it had correspondingly disastrous results.

APPENDIX A



NORMAN D. DIX
C. RABALA RUI
JEANNE J. JONES
KARL E. JENSEN

BIVARIATE CORRELATION ANALYSIS: PEARSON CORRELATION, RANK-ORDER CORRELATION, AND SCATTER DIAGRAMS

The SPSS system furnishes three subprograms for bivariate correlation analysis: PEARSON CORR, NONPAR CORR, and SCATTERGRAM. PEARSON CORR computes Pearson product-moment correlation coefficients for pairs of interval-level variables.¹ Spearman and Kendall rank-order correlations, appropriate for ordinal-level variables, are calculated by the NONPAR CORR subprogram. Both subprograms provide the user with significance tests and have the capability of producing correlation matrices (on an output medium of the user's choice) for input into other programs. The SCATTERGRAM subprogram prints two-variable scattergrams of data points. It will also compute a simple linear regression. Each of the programs contains several options for handling missing data. As usual, all data-selection and data-modification procedures available in SPSS may be employed while using these subprograms.

18.1 INTRODUCTION TO CORRELATION ANALYSIS

Bivariate correlation provides a single number which summarizes the relationship between two variables.² These correlation coefficients indicate the degree to which variation (or change) in one variable is related to variation (change) in another. A correlation coefficient not only summarizes the strength of association between a pair of variables, but also provides an easy means for comparing the strength of relationship between one pair of variables and a

¹Several social science methodologists argue that the Pearson correlation coefficients (and other statistics originally designed for interval-level variables) may be used even if the data satisfy only the assumptions of ordinal-level measurement (Labovitz, 1970, 1972; Tufte, 1969). Since such a usage is not standard procedure, users should pursue it cautiously and only after awareness of the implications of such a decision.

²Many introductory statistics texts offer detailed discussions of the statistics computed by the programs discussed here. In particular, see Blalock (1972) and Mueller, Schuessler, and Costner (1970).

different pair. Of course, this is done at the sacrifice of the detail which one has in a crosstabulation, scattergram, or list of values for each case.

In the CROSSTABS subprogram, data are reported in contingency table form and a number of measures of association can be computed (see Chap. 16). These measures of association are also correlation coefficients in that they summarize the strength of the bivariate relationship. Most of them, however, were specifically designed to supplement crosstabulations, especially those based on nominal- and ordinal-level variables with relatively few categories (see Sec. 1.2.1 for a discussion of level of measurement). The correlations presented in this chapter are appropriate for variables measured at the interval or ratio level and for ordinal-level variables with many categories.

Spearman's rho and *Kendall's tau* are the two nonparametric correlations computed by the NONPAR CORR subprogram. *Nonparametric* means that no assumptions are made about the distribution of cases on the variables. Indeed, these statistics require nothing more than an ordinal level of measurement and a large number of categories or ranks on each of the variables. They are basically designed to determine whether two rankings of the same cases are similar. For instance, we might ask two experts on international relations to rank-order 50 industrialized countries according to their evaluation of the overall military strength of those countries. The two sets of rankings would probably be very similar but some differences might exist. Rho or tau would give us a measure of how similar (or dissimilar) they actually are.

Note that a ranking presumes there will not be a large number of cases with identical scores; therefore, a crosstabulation would not be useful, and a visual comparison of the two sets of ranks would be rather confusing when there is a large number of cases. A correlation coefficient then becomes very useful, because it is a summary measure. This is not to say that the details should be ignored, but rather that a summary of the strength of relationship conveys a great deal of information and is often sufficient for many research needs.

Interval- and ratio-level variables are usually unsuited for crosstabulations since they are frequently composed of a large number of distinct categories. When this is the case, scattergrams and the Pearson product-moment correlation (r) can give us a picture of the relationship. A *scattergram* is a graph of data points based on two variables, where one variable defines the horizontal axis and the other defines the vertical axis. The values of the variables for any given case serve as the coordinates of the point representing that case. Figure 18.1 (a) is a hypothetical example.

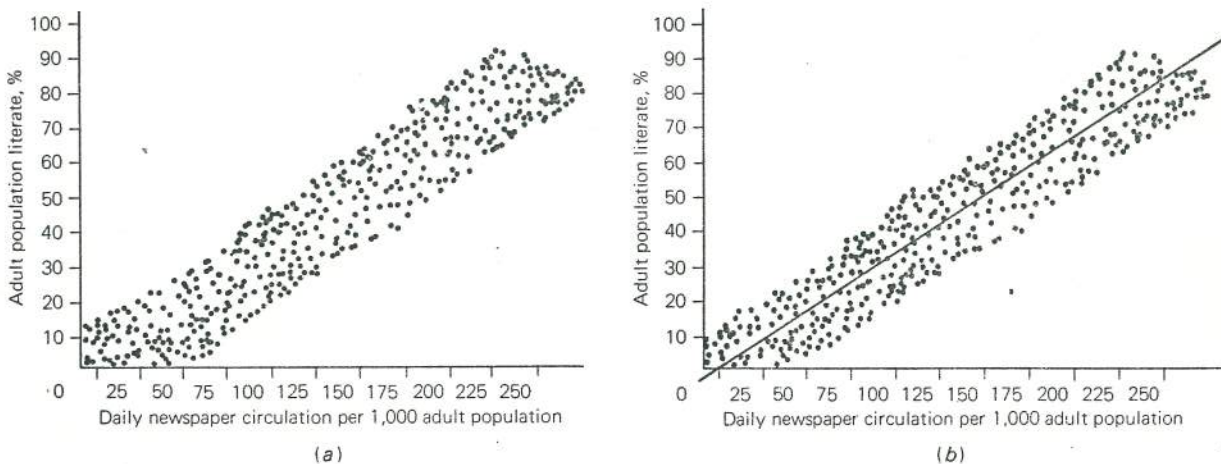


FIGURE 18.1 Scattergram illustrating a strong positive linear relationship.

As with crosstabulations, scattergrams often suffer from excessive detail. One way to reduce the detail is to draw a straight or curved line through the scattergram in such a manner that it approximates the pattern of points. This is quite easy when the pattern is clear and consistent. Thus, in Fig. 18.1 the rate of adult literacy seems to be highly positively related to newspaper circulation, because the points cluster in a narrow band forming a pattern that could be well summarized by a straight line drawn through the scatter of data points, as has been done

in Fig. 18.1 (b). In contrast, there does not appear to be any systematic relationship among the data represented in Fig. 18.2(a) since the points do not show any distinct pattern. A mild degree of clustering along a downward sloping straight line seems present in Fig. 18.2(b), indicating a moderate degree of negative association. The pattern in Fig. 18.2(c) is very distinct, indicating a fairly high degree of association based on a curvilinear relationship. If a line with known mathematical properties can be found to represent the general pattern of the data, then the formula for that line can serve as a summary of the form of the relationship between the two variables. In addition, the closer the data points fall to the line that best summarizes the relationship, the stronger the correlation between the two variables.

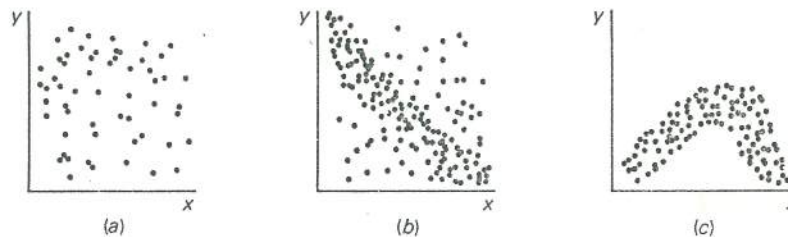


FIGURE 18.2 Scattergrams illustrating different types of relationships.

The most common statistical procedure for fitting a line to a scattergram based on interval level variables is called *least-squares regression*. This method is based on the belief that the *best-fitting* line is the one in which the vertical distances of all the points from the line are minimized. The line itself is called the *regression line*. That is, if some straight or curved line were drawn through the scattergram, any point which did not fall exactly on the regression line would be incompletely accounted for. The amount of "error," then, is the vertical distance from the point to the line. Actually, the distances are squared and then added together. This summation of the squared error distances is a measure of the total error involved when the regression line is used as the prediction of the location of the data points. A line which minimizes this sum of squared distances will serve as a better predictor than any other line. If variable Y is plotted along the vertical axis and variable X along the horizontal axis, we would call the resulting line the regression of Y on X since it is the vertical distances that are being minimized. (If we were to compute the regression of X on Y , we would be minimizing the horizontal distances; our result would usually be a different line.)

The most common type of regression is *linear regression*, in which the objective is to locate the best-fitting *straight* line. Linear regression is most commonly used because it gives a simple summary of the relationship, although not necessarily the "best," and since most variables of interest to social scientists are assumed to be related in a straightline manner. The general formula for a straight line is

$$Y = a + bX$$

where a is called the *intercept* and is the value of Y at the point where the line crosses the Y (vertical) axis (X is zero there), and b is the *slope* of the line (it denotes how much Y changes for a one unit change in X). When the values of a and b are determined by the least-squares regression method, b is called the *regression coefficient*. The SCATTERGRAM subprogram not only prints a plot of the data points, but it also computes the linear regression coefficient, the intercept, and other associated statistics (see Sec. 18.4 for details).

Sometimes a bivariate relationship, such as that shown in Fig. 18.2(c), is more aptly described by a curve. Regression methods for fitting a curve are called *curvilinear* or *polynomial regression*. The criterion of least-squares distances still applies, but the formula derived is

$$Y = a + b_1X + b_2X^2 + b_3X^3 + \cdots + b_nX^n$$

Here, the largest exponent (n) defines the *degree* of the polynomial and is determined by what the researcher feels would be necessary to adequately describe the relationships between the two

variables. In order to perform a polynomial regression the REGRESSION subprogram (Chap. 20) must be used, since each power of X is really treated as though it were a separate variable, i.e., it is a multivariate problem. A scattergram can be very helpful though in displaying the relationships, and from the display the researcher can decide whether a polynomial regression is warranted.

In most social science research it is highly unusual to find a regression line, especially a straight one, which perfectly fits the data. Whether this is because the true relationship does not quite fit the curve being drawn or because of errors or imprecisions in collecting the data, a measure of the "goodness of fit" of the regression line is called for. The *Pearson product-moment correlation coefficient*, symbolized by r , serves this purpose for linear regression. When there is a perfect fit (no error), r takes on the value of $+1.0$ or -1.0 , where the sign is the same as the sign of the regression coefficient. A negative r does not mean a bad fit, rather it denotes an inverse relationship — as X becomes larger, Y tends to become smaller, as occurs in Fig. 18.2(b). A positive correlation means that X and Y tend to increase (or decrease) together, as depicted in Fig. 18.1. When the linear regression line is a poor fit to the data, r will be close to zero. Indeed, the value of zero denotes the absence of a linear relationship, as seems to be the case for Fig. 18.2(a) and (c).

Pearson's r , which is computed both by SCATTERGRAM and PEARSON CORR, serves a dual purpose. Besides its role as an indicator of the goodness of fit of the linear regression, it is a measure of association indicating the strength of the linear relationship between the two variables. The regression coefficient b does not serve this purpose; it merely denotes the slope of the line. When we want to know the strength and direction of a linear relationship, we consult r . If the value of r is close to zero, we can assume there is little or no linear relationship between the two variables. If the value of r approaches $+1.0$ or -1.0 , we can assume there is a strong linear relationship.

If we square the Pearson's r we get another statistic, denoted by r^2 . Actually, r^2 is a more easily interpreted measure of association when our concern is with strength of relationship rather than direction of relationship. (It ranges from a minimum of 0 to a maximum of 1.0.) Its usefulness derives from the fact that r^2 is a measure of the proportion of variance in one variable "explained" by the other.

Variance is a measure of the variability, or lack of homogeneity, in a variable (see Sec. 14.1 for further details). When the cases cluster close to the mean, variance will be small; as the cases become more spread out, variance increases. The objective of correlation analysis is to determine the extent to which variation in one variable is linked to variation in the other (referred to as *concomitant variation*).

Concomitant variation of one variable with another explains variance in the following sense. If we want to predict the value of some variable Y (for example, the percentage of the adult population which is literate) for a given country without having any prior knowledge of the country's characteristics, our best guess would be the average (mean) literacy figure for all the countries. The variance of Y gives an indication as to how far off our prediction is likely to be, since it is based on the sum of squared distances of the cases from the mean of the variable. Now, if we find some characteristic X (for instance, daily newspaper circulation) of these countries which happens to be linearly correlated with Y , our ability to predict the level of literacy will be improved. The prediction strategy is to compute the regression line and to predict that the value of Y (literacy) is the point on the regression line corresponding to the country's position on X (newspaper circulation). Thus, if we knew that a particular country had a newspaper circulation of 100 per thousand adults, a regression line to fit the data depicted in Fig. 18.1 would predict a literacy rate of about 20 percent. Yet, clearly, the several countries with this level of newspaper circulation have actual literacy rates ranging from about 10 to 30 percent. If there is a high correlation, as measured by r , most of the data points will fall very close to the line, and the differences (errors) between our predictions and the true values will be much smaller on the average than the discrepancy which would occur by always predicting the mean value of Y .

The size of the error is measured by the vertical distance from the actual data point to the regression line. These distances are squared and then summed together over all the cases and divided by the number of cases minus two ($N - 2$). We thus have a statistic called *residual variance*,

that is, the amount of original (total) variance which cannot be explained by using the regression line as a prediction device. Residual variance will never be greater than total variance, and the proportion that it is less the proportion of variance explained (r^2). That is,

$$r^2 = \frac{\text{total variance} - \text{residual variance}}{\text{total variance}}$$

Since r and r^2 are symmetric measures of association, it does not matter which variable is considered to be predicting the other. Both r and r^2 measure the strength of the *linear* relationship.

Often, we are not even interested in prediction or the regression line itself. Rather, we wish only to know the strength of the relationship or to obtain the correlation coefficient for other statistical purposes. The PEARSON CORR subprogram is very convenient for such situations since it can easily compute a large number of correlation coefficients without taking the time to display a scattergram or compute a regression equation.

18.2 SUBPROGRAM PEARSON CORR: PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENTS

Subprogram PEARSON CORR computes Pearson product-moment correlations for pairs of variables. (These are *zero-order* correlations because no controls for the influence of other variables are made. Higher-order *partial* correlations are produced by the PARTIAL CORR subprogram.) The Pearson correlation coefficient r is used to measure the strength of relationship between two interval-level variables. In this case, the strength of relationship indicates both the goodness of fit of a linear regression line to the data and, when r is squared, the proportion of variance in one variable explained by the other (see the discussion in Sec. 18.1).

Mathematically, r is defined as the ratio of covariation to square root of the product of the variation in X and the variation in Y , where X and Y symbolize the two variables. This corresponds to the formula

$$r = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\left\{ \left[\sum_{i=1}^N (X_i - \bar{X})^2 \right] \left[\sum_{i=1}^N (Y_i - \bar{Y})^2 \right] \right\}^{1/2}}$$

where X_i = i th observation of variable X

Y_i = i th observation of variable Y

N = number of observations

\bar{X} = $\sum_{i=1}^N X_i / N$ = mean of variable X

\bar{Y} = $\sum_{i=1}^N Y_i / N$ = mean of variable Y

This formula can be restated by dividing the numerator and denominator by $N - 1$ to show that the correlation coefficient can also be defined as the covariance in X and Y divided by the product of their standard deviations. The covariance in X and Y is defined as

$$\frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

The actual formula used by SPSS for computing Pearson correlation coefficients is

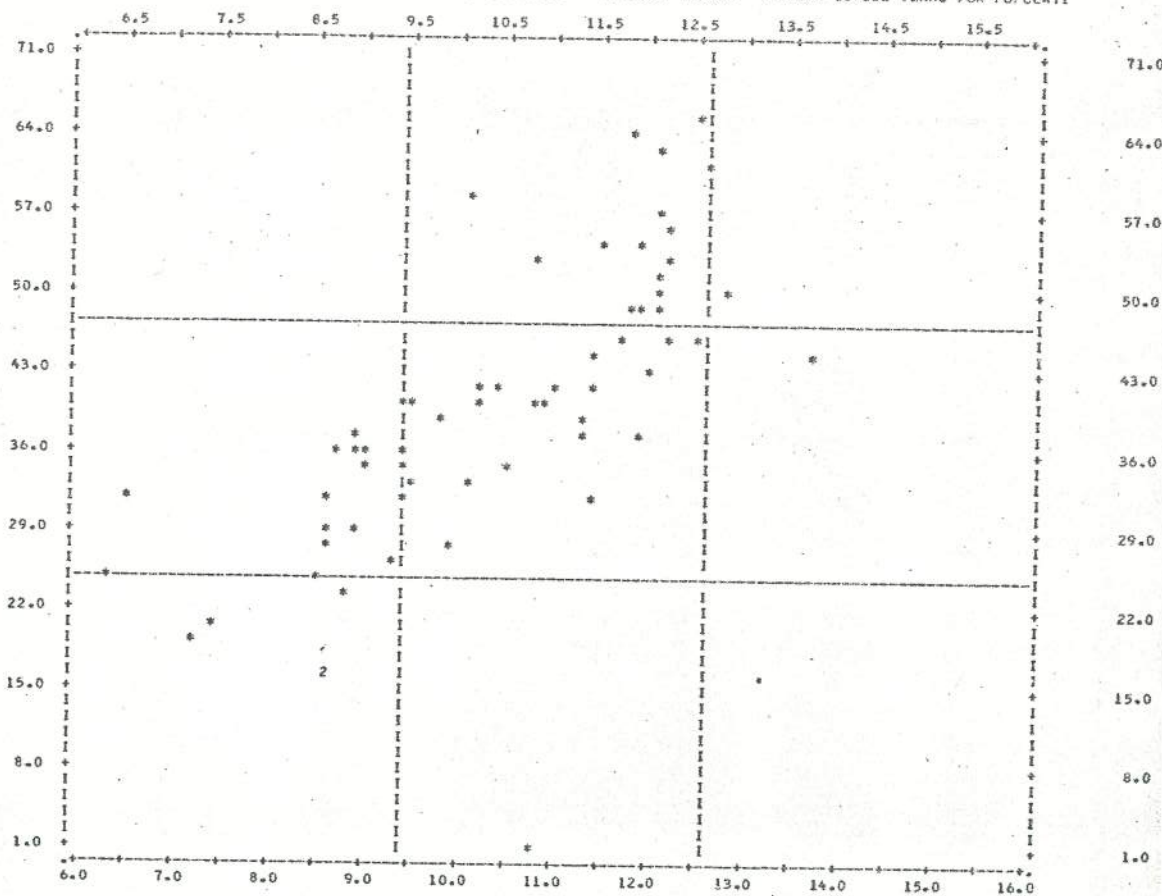
$$r = \frac{\sum_{i=1}^N X_i Y_i - \left(\sum_{i=1}^N X_i \right) \left(\sum_{i=1}^N Y_i \right) / N}{\left\{ \left[\sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i \right)^2 / N \right] \left[\sum_{i=1}^N Y_i^2 - \left(\sum_{i=1}^N Y_i \right)^2 / N \right] \right\}^{1/2}}$$

STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES SPSSH - VERSION 5.01

02/17/74

PAGE 3

FILE COMSTUDY (CREATION DATE = 02/17/74) STUDY OF AMERICAN SKALL COMMUNITIES
 SCATTERGRAM OF (DOWN) WHTCOLAR PERCENT CIVILIAN LABOR IN WHITE (ACROSS) MEDSCH MEDIAN SCHOOL YEARS FOR POPULATI



STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES SPSSH - VERSION 5.01

02/17/74

PAGE 4

STATISTICS..

CORRELATION (R)-	0.70395	R SQUARED -	0.49555	SIGNIFICANCE -	0.00001
STD ERR OF EST -	9.07510	INTERCEPT (A) -	-15.63027	SLOPE (B) -	5.36871
PLOTTED VALUES -	63	EXCLUDED VALUES-	0	MISSING VALUES -	1

***** IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

FIGURE 18.8 Output from subprogram SCATTERGRAM.

**MULTIPLE REGRESSION ANALYSIS:
SUBPROGRAM REGRESSION**

Jae-On Kim
Frank J. Kohout
University of Iowa

The SPSS multiple regression subprogram combines standard multiple regression and stepwise procedures in a manner which provides considerable control over the inclusion of independent variables in the regression equation. The variable transformation features of the SPSS package allow the regression subprogram to be used for a variety of multivariate analyses, such as polynomial regressions, dummy regressions, and analysis of variance and covariance. In addition, the subprogram allows the user to examine the residuals and predicted values for later analyses. Output of normalized (standardized) regression coefficients in addition to the ordinary (unstandardized) regression coefficients also allows easy calculation of path coefficients.

Input to the REGRESSION subprogram may consist of either raw data cases (from a raw-input-data file or SPSS system file) or a correlation matrix. If the input consists of raw data cases any of the SPSS variable-transformation features may be used, and several options are available for handling missing data. The alternative possibility of using a correlation matrix for input, with or without means and standard deviations, allows the researcher to perform extended analyses without calculating the correlation matrix more than once. This feature provides considerable savings of time when large files and/or large matrices are involved. It also permits the user to input correlation matrices obtained from other sources, such as the SPSS subprogram PEARSON CORR and correlation matrices from other programs in which the coefficients are corrected for attenuation. Computational techniques developed for this program permit the handling of a large number of independent and dependent variables in a fast and accurate manner.

In general, multiple regression requires that variables are measured on interval or ratio scale and the relationships among the variables are linear and additive. These restrictions are not absolute, however. As will be shown in Chap. 21, nominal variables can be incorporated into

regression through the use of "dummies," nonlinear and nonadditive relationships can be handled through transformation of variables or through the introduction of product-terms. In view of the complexity of the subject, the topic will be presented in two separate chapters. In this chapter, Sec. 20.1 is devoted to the basic statistical concepts involved in multiple regression procedures. Users who are familiar with multiple regression may wish to go directly to Sec. 20.2. Section 20.2 contains a description and explanation of the procedure necessary to activate and use subprogram REGRESSION and its various statistics and options. In Chap. 21, we present a brief discussion of special topics in the general linear approach, such as nonlinear regression, uses of categorical variables in multiple regression, analysis of variance and covariance, and path analysis. By the use of actual examples of data-transformation facilities of SPSS we will demonstrate how these analyses may be accomplished with subprogram REGRESSION.

Although we have attempted to include enough statistical material for the intelligent use of the REGRESSION subprogram, the approach is informal. Users who want a more rigorous approach should consult the references cited at the end of Chaps. 20 and 21.

20.1 INTRODUCTION TO MULTIPLE REGRESSION

Multiple regression is a *general* statistical technique through which one can analyze the relationship between a dependent or criterion variable and a set of independent or predictor variables. Multiple regression may be viewed either as a *descriptive* tool by which the linear dependence of one variable on others is summarized and decomposed, or as an *inferential* tool by which the relationships in the population are evaluated from the examination of sample data. Although these two aspects of the statistical technique are closely related, it is convenient to treat each separately, at least on a conceptual level. Since the method (as a descriptive tool or inferential tool) can be used for a variety of related purposes, we will illustrate only a few of its most common applications.

The most important uses of the technique as a descriptive tool are: (1) to find the best linear prediction equation and evaluate its prediction accuracy; (2) to control for other confounding factors in order to evaluate the contribution of a specific variable or set of variables; and (3) to find structural relations and provide explanations for seemingly complex multivariate relationships, such as is done in path analysis.

Suppose, for example, that a researcher is interested in predicting Political Tolerance (the dependent variable) from Education, Occupation, and Income (the independent variables), all of which have been measured at least on interval scales for a sample of respondents. Through multiple regression techniques the researcher could obtain a prediction equation that indicates how scores on the independent variables could be weighted and summed to obtain the best possible prediction of Political Tolerance for the sample. The researcher would also obtain statistics that indicate how accurate the prediction equation is and how much of the variation in Political Tolerance is accounted for by the joint linear influences of Education, Occupation, and Income. The researcher may also wish, in this connection, to "simplify" the prediction equation by deleting independent variables that do not add substantially to prediction accuracy, once certain other independent variables are included. For instance, if the contribution of Income to explaining variation in Political Tolerance is trivial when used in combination with Education and Occupation, the researcher may decide to delete Income from the predictors. The main focus of the analysis is, however, the evaluation and measurement of *overall* dependence of a variable on a set of other variables.

Instead of focusing on prediction of the dependent variable and its overall dependence on a set of independent variables, the researcher may concentrate on the examination of the relationship between the dependent variable and a particular independent variable. For example, the researcher may wish to examine the influence of Education on Tolerance. However, a simple regression of Tolerance on Education will not provide an appropriate answer because the level of Education is confounded with Occupation and Income, that is, the more educated one is, the more likely one is to have a higher status occupation and higher income. Occupation and

income levels may themselves affect tolerance. Therefore, the researcher would want to examine the impact of Education while controlling for variation in Occupation and Income, and would use multiple regression to get a variety of "partial coefficients." Emphasis in this case is on the examination of particular relationships within a multivariate context.

Another application of multiple regression as a descriptive tool is the use of multiple regression technique in conjunction with causal theory. The emphasis in such an application is neither on the overall dependence of one variable on another nor the relationship between any particular pair of variables. Rather, multiple regression is used to describe the entire structure of linkages between independent and dependent variables and to assess the logical consequences of a structural model that is posited a priori from some causal theory. The best known such application is *path analysis*. For illustration, the researcher might have constructed the causal theory represented by the diagram in Fig. 20.1. The causal theory specifies an "ordering" among the variables that reflects a presumed structure of cause-effect linkages. Multiple regression techniques are then used to determine the magnitude of direct and indirect influence that each variable has on other variables that follow it in the presumed causal order. Each arrow in Fig. 20.1 represents a presumed causal linkage or path of causal influence. Through regression techniques, the strength of each separate path is estimated. This estimation actually involves several regression equations since Occupation is a dependent variable for the Education-Occupation relationship, Income is a dependent variable for the Education-Occupation-Income relationship, and Political Tolerance is a dependent variable with regard to the remaining three variables. The foregoing examples by no means exhaust the possible variations in multiple regression as a descriptive tool. More will be considered in the following exposition and in Chap. 21.

For every use of regression as a descriptive tool, there is usually a corresponding question of statistical inference—whether one can generalize the results of the sample observation to the universe. The problems of statistical inference can be conveniently grouped into two general categories: estimation and hypothesis testing. The purpose of estimation is to find the most likely population parameters from the examination of sample observations. For example, the researcher may estimate the regression coefficients in the population from his sample data and may establish some confidence intervals. The main focus here is in delineating a particular value or values for the population. The researcher may, on the other hand, focus on evaluating various hypotheses about the population. That is, instead of asking what value a population parameter is likely to have, one may simply test the null hypothesis that its value is zero against the alternative hypothesis that its value is greater or less than zero. Some of the most often used null hypotheses in multiple regression are:

- 1 There is no linear relationship between a dependent variable and a set of independent variables, for example, that Tolerance is not related to any one of the socioeconomic status variables.
- 2 A particular independent variable has no linear effect on the dependent variable once the effects of other independent variables are adjusted for. For example, there is no relationship between Tolerance and Income; the observed relationship between the two is merely due to the sampling fluctuation.
- 3 The relationship between the dependent variable and particular independent variable is non-linear, and that the effects of two or more variables are not additive.

Tests for the last two hypotheses are discussed in Chap. 21.

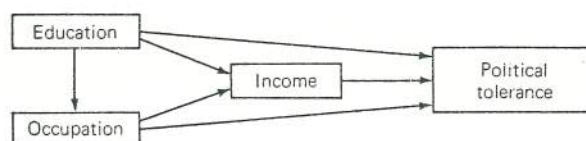


FIGURE 20.1

These examples again do not exhaust the possible hypotheses and estimates that may be involved in a multiple regression problem. More will be considered in the following exposition. As a preliminary to multiple regression, we first provide a brief review of a simple bivariate linear regression. Principles and concepts introduced in the bivariate context are then applied to the multivariate situation.

20.1.1 SIMPLE BIVARIATE REGRESSION

20.1.1.1 Meaning of Regression Coefficients

In simple regression analysis, values of the dependent variable are predicted from a linear function of the form

$$Y' = A + BX \quad (1)$$

where Y' is the estimated value of the dependent variable Y , B is a constant by which all values of the independent variable X are multiplied, and A is a constant which is added to each case.

The difference between the actual and the estimated value of Y for each case is called the *residual*, i.e., the error in prediction, and may be represented by the expression

$$\text{Residuals} = Y - Y'$$

The regression strategy involves the selection of A and B in such a way that the sum of the squared residuals is smaller than any possible alternative values. Expressed in another way,

$$\Sigma(Y - Y')^2 = SS_{\text{res}} = \text{minimum}$$

It can be shown that the optimum values for B and A are obtained from the following formulas:

$$B = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2} = \frac{SP_{xy}}{SS_x} \quad (2)$$

$$A = \bar{Y} - B\bar{X} \quad (3)$$

where SP_{xy} is our symbolic notation for the sum of cross products of X and Y and SS_x denotes the sum of squares of X .

The B and A coefficients are asymmetrical since X has been taken as the predictor of Y . Their values will not, in general, equal those obtained when Y is used as a predictor of X .

As Fig. 20.2 indicates, the constant A (referred to as the *Y intercept*) is the point at which the regression line crosses the Y axis and represents the predicted value of Y when $X = 0$. The constant B , usually referred to as the (nonstandardized) regression coefficient, is the slope of the regression line and indicates the expected *change* in Y with a *change* of one unit in X .¹ The predicted Y' values fall along the regression line, and the vertical distances ($Y - Y'$) of the points from the line represent residuals (or errors in prediction). Since the sum of squared residuals is minimized, the regression line is called the *least-squares line* or the *line of best fit*. In other words, there is no other line which is "closer" to the points, i.e., for no other line is $\Sigma(Y - Y')^2$ smaller.

20.1.1.2 Partitioning of Sum of Squares

The total sum of squares in Y (which is the variability of the dependent variable Y) can be partitioned into components that are (1) explained or accounted for by the regression line,

¹To be more precise, it indicates the expected *difference* on Y between two groups that happen to be different on X by one unit.

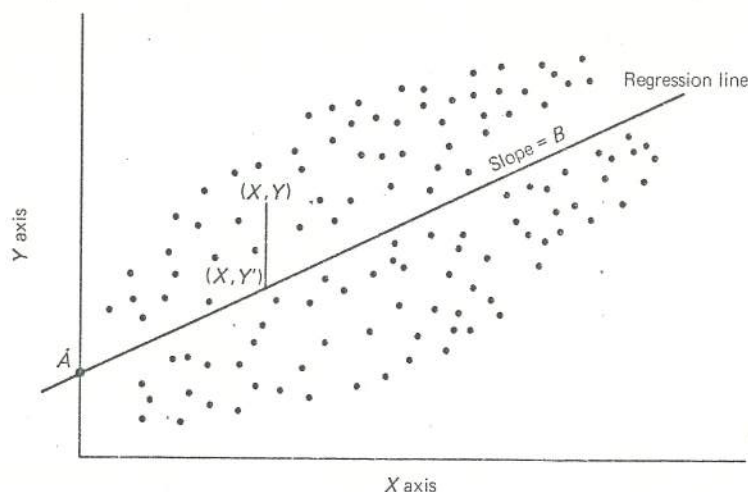


FIGURE 20.2

denoted by SS_{reg} , and (2) unexplained (the sum of squared residuals), $SS_{\text{res}} = \sum (Y - Y')^2$. Since the least-squares solution guarantees that the residuals are independent of the predictor Y' , we may write the partition as¹

$$\begin{aligned} SS_y &= SS_{\text{reg}} + SS_{\text{res}} \\ \sum (Y - \bar{Y})^2 &= \sum (Y' - \bar{Y})^2 + \sum (Y - Y')^2 \end{aligned} \quad (4)$$

Given this partitioning, a natural measure of prediction accuracy and the strength of linear association is the ratio of explained variation in the dependent variable Y to the total variation in Y .

$$\begin{aligned} r_{xy}^2 &= \frac{SS_{\text{reg}}}{SS_y} \\ &= \frac{SS_y - SS_{\text{res}}}{SS_y} \end{aligned} \quad (5)$$

This ratio is sometimes referred to as the *coefficient of determination*. The square root of this ratio is the Pearson product-moment correlation between variables X and Y . (For a discussion of the correlation coefficient see Chap. 18.) While the correlation coefficient always has the same sign as the regression coefficient, these two coefficients will not be equal except in the special case where the variances of X and Y are equal, as for example when both X and Y are standardized variables.

¹The regression sum of squares has many equivalent forms, the examination of which is quite instructive. Recall that B is the regression coefficient from the equation $Y' = A + BX$, $SP_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$, and $SS_x = \sum (X_i - \bar{X})^2$

$$\begin{aligned} SS_{\text{reg}} &= \sum (Y' - \bar{Y})^2 \\ &= B(SP_{xy}) \\ &= B^2(SS_x) \\ &= r^2(SS_y) \end{aligned} \quad (6)$$

Note that B is given by (SP_{xy}/SS_x) .

The residual sum of squares may be represented as the difference between the regression sum of squares and the total sum of squares.

$$\begin{aligned} SS_{\text{res}} &= SS_y - SS_{\text{reg}} \\ &= (1 - r^2)SS_y \end{aligned} \quad (7)$$

20.1.1.3 Standardized Regression Coefficient

When both X and Y are standardized to have unit variance (i.e., the standard deviations of both X and $Y = 1$), the regression coefficients B_{yx} and B_{xy} will not only be equivalent to each other but will be equivalent to the simple correlation

$$B_{yx} = B_{xy} = r_{xy} \quad (8)$$

Bold face B 's are used in this latter expression to indicate that they have been computed on standardized X and Y values rather than the original unstandardized values. Standardized regression coefficients are also referred to as *beta weights*. The relationship between beta weights and unstandardized regression coefficients is shown in the identity:

$$B_{yx} = B_{yx} \left(\frac{S_x}{S_y} \right) \quad (9)$$

where S_x is the standard deviation of X and S_y is the standard deviation of Y .

While B_{yx} (the beta weight) does not enable one to estimate Y values in the original raw value units, the standardized regression coefficient is more convenient to use in a number of contexts. Working with beta weights enables one to simplify the linear regression equation, since the constant A (the Y intercept) is always equal to zero and therefore can be omitted. Furthermore, when there are two or more independent variables measured on different units (such as income in dollars and education in years), standardized coefficients may provide the only sensible way to compare the relative effect on the dependent variable of each independent variable. Moreover, a standardized regression coefficient is quite readily transformed to its unstandardized counterpart if standard deviations for the original X and Y are available. The transformation is derived from (9):

$$B_{yx} = B_{yx} \left(\frac{S_y}{S_x} \right) \quad (10)$$

20.1.1.4 Standard Error of Estimate and Prediction Accuracy

If the researcher wishes to evaluate the accuracy of the prediction equation or, equivalently, to determine the amount of prediction error associated with the predictions, it will be necessary to examine one or more of the statistics that reflect the average size of residuals. The r^2 statistic or its complement $(1 - r^2)$ indicate *proportions* of variation explained and unexplained, respectively. For some purposes, the researcher may prefer to base an assessment of prediction accuracy upon the *absolute* amount of explained or unexplained variation.

A widely used statistic of this sort is the *standard error of estimate* (SEE), which is simply the standard deviation of actual Y values from the predicted Y' values. If the standard error of estimate were to be computed by hand, the following formula would be used:

$$\begin{aligned} \text{SEE} &= \sqrt{\frac{\sum (Y - Y')^2}{N - 2}} \\ &= \sqrt{\frac{SS_{\text{res}}}{N - 2}} \end{aligned} \quad (11)$$

The formula directs one to first divide the residual sum of squares SS_{res} by sample size $N - 2$ to obtain the *average of squared residuals*. The square root of this latter quantity is the standard error of estimate, which may be interpreted as a sort of "average residual" or "average error in predicting Y from the regression equation." The standard error of estimate is normally obtained as part of the computer output in a regression analysis.¹ If it is assumed that actual Y values are

normally distributed about the regression line, the researcher will be able to estimate the proportion of cases that will fall between ± 1 standard error of estimate units from the predicted values, ± 2 standard error of estimate units from the predicted values, and so forth.

20.1.1.5 Standard Error of B

If B is estimated from a sample, the values of B will vary from sample to sample. We know, however, from statistical theory that in the long run the mean of B 's will coincide with the population value β , and we can estimate the standard deviation of the sampling variability of B if certain assumptions are met. The estimate of the *standard error* of B is given by

$$\begin{aligned}\sqrt{\text{Var}(B)} &= \sqrt{\frac{\Sigma(Y - Y')^2/(N-2)}{\Sigma(X - \bar{X})^2}} \\ &= \sqrt{\frac{SS_{\text{res}}/(N-2)}{SS_x}}\end{aligned}\quad (12)$$

If the sample size is large, i.e., greater than 200, the estimates of B from repeated sampling will approximate a normal distribution. Therefore, the researcher can establish the confidence interval for the estimated B . For example, if the estimated B is 2.3 and the standard error of B is .4, the 95 percent confidence interval would be

$$2.3 - 1.96(.4) < \beta < 2.3 + 1.96(.4)$$

If the sample size is relatively small, the B estimates follow the t distribution with $(N - 2)$ degrees of freedom. Therefore, the 95 percent confidence interval for β given the sample size of 62, estimated $B = 2.3$, and the standard error of $B = .4$, is given by

$$2.3 - 2(.4) < \beta < 2.3 + 2(.4)$$

The value (2) is obtained from the table of Student's t distribution with degrees of freedom equal to 60. The standard error of B is routinely provided by the subprogram REGRESSION.

20.1.1.6 Significance Test for B

The significance of B can be tested either by examining the confidence interval or, more conveniently, by evaluating the following F ratio.

$$\begin{aligned}F &= \frac{\Sigma(Y' - \bar{Y})^2/1}{\Sigma(Y - Y')^2/(N-2)} \\ &= \frac{SS_{\text{reg}}}{SS_{\text{res}}/(N-2)}\end{aligned}\quad (13)$$

with degrees of freedom 1 and $(N - 2)$. If the computed F value is larger than the statistical table's critical value for a given level of significance, say .05, the null hypothesis that $\beta = 0$ would be rejected. Otherwise, it would be concluded that the observed B is not significant at the .05 level.

20.1.1.7 Symbol Reference Table

The reference table at the top of page 327 indicates the correspondence between the statistical symbols used in this text and those appearing on the printed output of the SPSS REGRESSION procedure.

20.1.1.8 Illustrative Example of a Bivariate Regression Analysis

To lend substance to the foregoing abstract exposition, it will be helpful to consider a concrete research application for bivariate regression analysis. Suppose a researcher is con-

Symbol used in the text	Symbol used in SPSS output	Meaning
<i>B</i> (<i>italic</i>)	B	Unstandardized regression coefficient
B	BETA	Standardized regression coefficient
β	(not used)	Population parameter of unstandardized regression coefficient

cerned with describing the relationship between Political Tolerance and Education. (Data from a fictitious 100-case file is summarized in Table 20.1. This table provides selected summary statistics that are obtained as part of the output from SPSS subprogram REGRESSION.) If the researcher is merely interested in describing the strength and direction of the relationship, it would only be necessary to examine the correlation coefficient r and the coefficient of determination r^2 . The sign of r would indicate the direction of the relationship, whether positive or negative, while the absolute value of r can be used as an index of the relative strength of the relationship. However, since r^2 indicates the proportion of variation in Political Tolerance explained by Education, it has a clearer interpretation than r as index of the strength of the relationship. The researcher would thus conclude that the relationship between Education and Political Tolerance is positive and that 25 percent of the variation in Political Tolerance is explained by linear regression on the Education variable.

TABLE 20.1 Selected Statistics for Bivariate Regression Generated by Subprogram REGRESSION

Multiple R	.5000	Analysis of variance	DF	SS	F
R^2	.2500	Regression	1	24.75	
Standard error	.8704	Residual	98	74.25	32.6667
Variable	<i>B</i>	B	Standard error <i>B</i>	<i>F</i>	
Education	.1667	.5000	.0292	32.667	
Constant <i>A</i>	3.1667				

The researcher may wish to go beyond the mere description of direction and strength of the Education-Tolerance relationship. In particular, the researcher may wish to determine what Tolerance scores would be predicted for sample respondents with various levels of Education. For this application the A and B statistics will be required. Table 20.1 indicates that $A = 3.1667$ and unstandardized $B = +.1667$. That is, the predicted score on Political Tolerance is 3.1667 when Education = 0, and the predicted score increases by .1667 units on the Political Tolerance scale for each unit (year) increase in Education. To obtain a predicted Political Tolerance score (Y') for any given level of Education (X), the researcher would employ the A and B constants in the linear prediction equation

$$Y' = 3.1667 + .1667X$$

For a person with 12 years of formal education, the predicted Political Tolerance score would be

$$Y' = 3.1667 + .1667(12) = 5.1667$$

By varying the number of years of Education, the researcher could obtain a predicted Political Tolerance score for each level of Education. All predicted scores will, of course, fall directly on the regression line and will not generally be equal to the actual observed Political Tolerance scores.

In some cases, the researcher will prefer to work with data that are standardized before the regression statistics are computed so that both variables are transformed into comparable units. The regression equation would be somewhat simpler since the constant A would be zero. The regression coefficient, in this case B , would indicate the number of standard deviation units change in Political Tolerance that would be predicted when Education changes by one standard deviation unit. Table 20.1 shows $B = +.5$, indicating that the predicted Political Tolerance score increases by .5 standard deviation units for each standard deviation unit increase in Education.

If the researcher wishes to evaluate the accuracy of the prediction by examining the

amount of absolute errors in the prediction, the standard error of estimate $(Y - Y')^2/(N - 2)$ may be used. This value is normally obtained as part of the computer output in a regression analysis. For the illustrative data in Table 20.1, the standard error of estimate is given as .8704. This means that the "average" error in guessing Political Tolerance scores from Education is .8704. If the assumption can be made that Political Tolerance scores are normally distributed about the regression line, the researcher will be able to say that the actual Tolerance Score of approximately 68 percent of the individuals will fall within the range $Y' \pm .8704$. For example, for the individuals with 12 years of education, the researcher will be able to say that approximately 68 percent of them will have tolerance scores falling in the interval of $(5.1667 - .8704) < Y < (5.1667 + .8704)$.¹

If the researcher is working with sample data, inferential statistics may be applied to test whether the observed linear association is statistically significant. The F ratios employed in such a test are routinely provided by the REGRESSION subprogram. The two F ratios shown in Table 20.1, one for the overall regression equation and the other for the regression coefficient, are the same in a bivariate regression. For the Political Tolerance example, $F = 32.667$ with 1 and 98 degrees of freedom. The F table indicates that this value is significant at the .001 level.

Finally, one may establish a confidence interval for B . If sample size is relatively large (200 or over), one can assume that the standard error of B (.0292 is our example) is equivalent to a standard deviation of a normal distribution. If sample size is small, Student's t distribution would be used. In our example, the 95 percent confidence interval is

$$.1667 - (1.99) \times .0292 \leq \beta \leq .1667 + (1.99) \times .0292$$

$$.1086 \leq \beta \leq .2248$$

That is, the probability is 95 percent that the true population regression coefficient β is between .1086 and .2248.

20.1.2 EXTENSION TO MULTIPLE REGRESSION

20.1.2.1 Basic Ideas

The basic principles of regression analysis used in the bivariate case may be extended to situations involving two or more independent variables. The general form of the (unstandardized) regression is

$$Y' = A + B_1X_1 + B_2X_2 + \cdots + B_kX_k \quad (14)$$

where Y' represents the estimated value for Y , A is the Y intercept, and the B_i are regression coefficients. The A and B_i coefficients are selected in such a way that the sum of squared residuals $\Sigma(Y - Y')^2$ is again minimized. This least-squares criterion implies that any other values for A and B_i would yield a larger $\Sigma(Y - Y')^2$. Selection of the optimum A and B_i coefficients using the least-squares criterion also implies that the correlation between the actual Y values and the Y' estimated values is maximized, while the correlation between the independent variables and the residual values $(Y - Y')$ is reduced to zero.²

The actual calculation of A and B_i requires a set of simultaneous equations derived by differentiating $\Sigma(Y - Y')^2$ and equating the partial derivatives to zero. A standard form of such equations for two predictor variables is

$$\begin{aligned} A + B_1\bar{X}_1 + B_2\bar{X}_2 &= \bar{Y} \\ B_1(SS_1) + B_2(SP_{12}) &= SP_{Y1} \\ B_1(SP_{12}) + B_2(SS_2) &= SP_{Y2} \end{aligned} \quad (15)$$

¹If sample size is relatively small, the confidence interval for predicted values becomes less reliable the more extreme the X value is. For more on estimation of confidence intervals, see Theil (1971).

²Note that to be precise, B_1 should be expressed as $B_{Y1.23\dots k}$ since it is a partial regression coefficient, that is, it expresses the effect of X_1 on Y when X_2, \dots, X_k are held constant. However, for brevity, we will not use the exact notation unless there is danger of confusing a partial regression coefficient with a simple regression coefficient.

where SS and SP stand for sum of squares and sum of products, or variation and covariation, respectively. For example, $SS_1 = \sum (X_{1i} - \bar{X}_1)^2$ and $SP_{12} = \sum (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$.

The solution of the last two equations in (15) gives

$$\begin{aligned} B_1 &= \frac{SP_{y1}(SS_2) - SP_{y2}(SP_{12})}{SS_1(SS_2) - SP_{12}^2} \\ B_2 &= \frac{SP_{y2}(SS_1) - SP_{y1}(SP_{12})}{SS_1(SS_2) - SP_{12}^2} \end{aligned} \quad (16)$$

and substituting these values into the first equation gives

$$A = \bar{Y} - B_1\bar{X}_1 - B_2\bar{X}_2 \quad (17)$$

It is sometimes more convenient to work with standardized variables and to calculate the unstandardized coefficients indirectly. When standardized variables are used the last two equations in (15), which are called the *normal equations*, become

$$\begin{aligned} B_1 + B_2r_{12} &= r_{y1} \\ B_1r_{12} + B_2 &= r_{y2} \end{aligned} \quad (18)$$

where r_{12} is the Pearson correlation between X_1 and X_2 , and B_1 is the standardized regression coefficient of the independent variable X_1 , etc.

The standardized partial regression coefficients can be expressed as

$$\begin{aligned} B_1 &= \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \\ B_2 &= \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2} \end{aligned} \quad (19)$$

The unstandardized coefficients are simply

$$\begin{aligned} B_1 &= B_1 \left(\frac{S_y}{S_1} \right) \\ B_2 &= B_2 \left(\frac{S_y}{S_2} \right) \end{aligned} \quad (20)$$

where S_i are standard deviations of the sample.

Since computation is performed by machine, there is no particular need to dwell on actual calculations any further. However, there are a few points concerning the normal equations that are worth noting. First, the derivation of the normal equations involving any number of independent variables is simple if the symmetry in (18) is observed. For example, the case of three independent variables would take the form

$$\begin{aligned} B_1 + B_2r_{12} + B_3r_{13} &= r_{y1} \\ B_1r_{12} + B_2 + B_3r_{23} &= r_{y2} \\ B_1r_{13} + B_2r_{23} + B_3 &= r_{y3} \end{aligned} \quad (21)$$

Second, the least-squares solution requires only a set of bivariate correlation coefficients for the solution of standardized regression coefficients, and sums of squares and cross products for the solution of unstandardized regression coefficients. Third, the normal equations lack a unique solution if the sample size is equal to or smaller than the number of variables involved, or if at least one of the independent variables is a perfect linear function of one or more others. Note, for example, that B_1 and B_2 in (19) become undefined when $r_{12} = 1$.

PARTIAL CORRELATION: SUBPROGRAM PARTIAL CORR

Subprogram PARTIAL CORR provides the user with the capability of computing large numbers of partial-correlation coefficients of any order or combination of orders. The subprogram has been designed so that the user may conveniently define multiple levels of control variables and multiple lists of independent and dependent variables on a single PARTIAL CORR procedure card. Up to 25 distinct sets of partials may be specified and each set may itself specify a large number of coefficients.

Input to the program may be either raw data (from a raw-input-data file or an SPSS system file) or one or more matrices of simple correlations. These correlation matrices may be generated by the user's own programs or by SPSS subprograms.

Output from subprogram PARTIAL CORR consists of the desired partial-correlation coefficients, the degrees of freedom, and a one- or two-tailed test of statistical significance. All simple correlations (zero-order partials) used in computing the partial may be printed if the user desires. The means and standard deviations of all variables entered onto the PARTIAL CORR procedure card may also be requested by means of the STATISTICS card. Punched matrices of simple correlation coefficients may also be output for future access on a medium of the user's choice.

Missing data may be excluded from the computation of the coefficients in either a pairwise or listwise fashion. As usual, pairwise deletion causes a case to be excluded from the computation of a given simple correlation (all partials are based on simple correlations) if either of the two variables involved in the computation of that coefficient has a value defined as missing. Listwise deletion causes a case to be deleted if any variable in the entire partial list has a value tagged as missing. Alternatively, missing data may, of course, be included in the computation of the partials, or the user may estimate missing data by recoding missing values to means, medians, etc.

As usual, all data-modification and data-selection procedures of SPSS may be employed while using subprogram PARTIAL CORR. This is not true, however, when the user is inputting

matrices rather than raw data. Section 19.3 describes all the special conventions for matrix input.

Before proceeding to the detailed description of this program and the control cards required to use it, a brief introduction to partial correlation will be presented for users wishing to review this topic. Other users may wish to proceed directly to Sec. 19.2.

19.1 INTRODUCTION TO PARTIAL-CORRELATION ANALYSIS

Partial correlation provides the researcher with a single measure of association describing the relationship between two variables while adjusting for the effects of one or more additional variables. Conceptually then, at least, partial correlation is analogous to crosstabulation with control variables. In crosstabulation the control is accomplished by examining the joint frequency distribution of two variables among two or more categories of one or more control variables, e.g., education's relationship to income, controlling for the effects of age. With crosstabulation the control is literal, i.e., one simultaneously locates each observation according to the values it takes on three or more variables. This is indeed one of the major problems with crosstabulation analysis, for each additional category of each variable in the relationship exerts a tremendous drain on the average cell frequencies. It takes a very large sample to execute even relatively simple controls.

In partial correlation, on the other hand, the control is statistical rather than literal and is based on the simplifying assumptions of linear relationships among the variables. In essence, partial correlation enables the researcher to remove the effect of the control variable from the relationship between the independent and dependent variables without physically manipulating the raw data. In partial correlation the effect of the control variable(s) is assumed to be linear throughout its range, and it is this linear assumption that makes partial correlation possible.

Once one knows the linear relationship among the independent, dependent, and control variables, the partial-correlation coefficient can be calculated by constructing (statistically, that is) new independent and dependent variables with the effect of the control variable(s) removed. This is done by making a prediction (based on the simple correlation coefficients) of both the independent and dependent variables from the knowledge of the effect that the control variable has on them. The new or adjusted independent variable is constructed by taking the difference between the actual value of the original independent variable (for each observation) and its value as predicted by the control variable. This new variable is, by definition, uncorrelated with each and/or all control variables which have been entered. The same procedure is then repeated for the dependent variable.

The linear effect of the control variable(s) has now been removed from both the independent and dependent variables, and the simple correlation between these adjusted variables is the partial correlation. However, since correlation coefficients are a complete description of the bivariate linear relationships among all the variables involved, this procedure can be statistically achieved from the correlation matrix alone, without reference to the individual observations. Therefore, when one computes the partial-correlation coefficient from the correlation matrix, the result is the same as if one had calculated the residuals for each observation [based on the effects of the control variables(s)] and had then computed a new simple correlation between the two sets of residuals. That is what we mean by adjusting the value based on the prediction from the simple correlation.

The basic formula for the computation of partial-correlation coefficients is

$$r_{ij.k} = \frac{r_{ij} - (r_{ik})(r_{jk})}{\sqrt{1 - r_{ik}^2} \sqrt{1 - r_{jk}^2}}$$

where k is the control variable, and i and j are the independent and dependent variables (the order is immaterial, since the correlation of i on j is the same as that of j on i). The extension of this formula to more than one control variable (that is, $n + 1$) is made by replacing the simple

correlation coefficients (or zero-order partials) on the right side of the equation with the n th-order partial coefficients. In this way the preceding formula can be used to recursively define and compute each higher-order partial from the previous one. It can be shown mathematically that the order in which one adds control variables has no effect on the ultimate partial. This is a result of the fact that the preceding formula is simply a computational shortcut of the residual-prediction procedure where the order in which the control variables are entered is clearly immaterial.

Partial correlation can be used in a wide variety of ways to aid the researcher in understanding and clarifying relationships between three or more variables. When properly employed, partial correlation becomes an excellent technique for uncovering spurious relationships, locating intervening variables, and can even be used to help the researcher make certain types of casual inferences.¹ In this brief introduction to partial correlation, we will attempt only to illustrate a few of the many types of conceptual problems for which partial-correlation analysis can be used. We will not, however, attempt to go beyond the simple statistical discussion presented above, and we strongly urge the user to consult one of the many available detailed statistical discussions of partial correlation.²

Partial correlation can be a very helpful tool for enabling the researcher to locate spurious relationships. A *spurious correlation* is defined in a relationship between two variables, A and B for example, in which A 's correlation with B is solely the result of the fact that A varies along with some other variable, C for example, which is indeed the true predictor of B . In this case, when the effects of C are controlled, held constant, etc., B no longer varies with A . As an illustration, let us take a hypothetical study of the determinants of crime rates in a sample of American communities. Let us further assume that the initial investigation has revealed a moderately strong positive correlation between the racial makeup of communities (measured as the proportion of nonwhites living there) and a composite crime-rate index. The researcher suspects, however, that the relationship is spurious and due solely to the fact that two other variables, (1) poverty (measured as the proportion of families with incomes less than \$3,000) and (2) size of community, covary strongly with both racial makeup and crime rates, and therefore the relationship between racial composition and crime rates is purely a function of the former's relationship to both poverty levels and community size. The question is then, does racial composition have any effect on crime rates when the effects of poverty and community size are removed? Let us examine some hypothetical data in order to indicate how partial correlation can address itself to this type of problem. Assume the following correlations existed between the four variables:

	Percent nonwhite	Percent below \$3,000	City size	Crime index
Percent nonwhite	1.00	.51	.41	.36
Percent below \$3,000		1.00	.29	.60
City size			1.00	.49
Crime index				1.00

First, it is clear from the simple correlations that the relationships between poverty and crime rate, and between city size and crime rate, are even stronger than that between racial composition and crime rate. Second, the correlations between racial makeup and the other two independent variables are quite strong. These are the researcher's first indications that the relationship between racial composition and crime rate may be a spurious one. The computation of three partial-correlation coefficients (two first-order partials and one second-order partial) will produce some relatively precise answers to these questions. If the correlation between racial composition and crime rate disappears (i.e., becomes zero) when we control for the effects of

¹Hubert M. Blalock, *Casual Inference in Non-experimental Research*, University of North Carolina Press, Chapel Hill, 1964, and Herbert A. Simon, *Models of Man*, Wiley, New York, 1957.

²M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 2, chap. 27, Griffin, London, 1961.

poverty and city size, we will have considerable evidence that the relationship is indeed a spurious one.

To begin, we will compute the first-order partial between racial composition and crime rate, controlling for the effects of poverty. This partial is .08, indicating that the initial correlation of .36 has been drastically reduced by simply controlling for the effects of poverty. Next we compute the second first-order partial, controlling for the effects of city size. This partial is .20. While the reduction is not as dramatic, it is still quite substantial. Finally, we compute the second-order partial indicating the relationship between racial composition and crime rate while simultaneously controlling for the effects of poverty and city size; this partial is $-.06$, or essentially zero. These relationships have now been clarified considerably: the relationship between racial composition and crime rate is spurious; the effects of both poverty and city size are acting to create the spurious relationship; but poverty is the variable having the greatest contaminating effect and is the major cause of the spurious relationship. Restated, these hypothetical findings suggest that when one controls for the effects of city size and particularly for levels of poverty, crime rates are similar irrespective of the racial composition of the city.

With a relatively small sample of cities, this type of multivariate analysis would have been extremely difficult, if possible at all, with crosstabulation. Partial correlation, on the other hand, provides a relatively easy and quite precise technique for this type of problem.

Another important feature of partial correlation lies in its ability to aid the researcher in a search for intervening linking variables. While there is no statistical difference between the computation of partials employed to locate spurious relationships and those used to determine intervening variables, the conceptual issues are different enough to merit separate treatment. The search for intervening variables is highly related to the issue of causality insofar as the researcher wishes to make statements of the sort: *A* leads to *B* which in turn leads to *C*. While partial correlation can be of great assistance in such problems, the researcher's theory (i.e., the ability to place a time-series ordering to the variables) becomes much more important in these types of situations.

Take, for example, a hypothetical study concerned with the transfer of wealth from parent to offspring. Given a strong correlation between parental wealth and that of their grown children, as well as high correlations between parental wealth and child's educational attainment and between offspring's educational attainment and their own wealth, an important issue might be the determination of the mechanisms which link parental wealth to that of their children. Given a matrix of correlations such as the following one might hypothesize (1) that there is a direct transfer of wealth from parents to offspring and that while the correlation of parents' wealth and child's education is supportive of this relationship, it is not critical; or (2) that the major proportion of the correlation between the wealth of parents and offspring is due to the impact that parents' wealth has on the educational attainment of their children, which is in turn the major predictor of the wealth of offspring.

	Parental wealth	Offspring's education	Offspring's wealth
Parental wealth	1.00	.53	.45
Offspring's education		1.00	.69
Offspring's wealth			1.00

The simple correlations indicate that approximately 20 percent of the variance in the wealth of all offspring sampled was determined by the wealth of their parents.¹ The degree to which this represents direct transfers of wealth, as opposed to the indirect effects via parental wealth's impact on the educational attainment of their offspring, can be determined by computing a partial correlation between parental wealth and that of their offspring. The size of that partial will indicate the proportion of the initial relationship which is due to direct transfers as opposed to educational attainment. This partial is .14, indicating that less than 2 percent out of the

¹The proportion of variance explained is equal to the square of the correlation coefficient.

original 20 percent of the explained variance between wealth of parents and their children is due to direct transfers of wealth, while the remaining 18 percent seems to be the result of the impact that parental wealth has on educational attainment.

The last example usage of partial correlation deals again with a slightly different problem: locating relationships where none appear to exist. Here too the statistical method is identical, but the conceptual issues are a bit different. One sometimes encounters situations where theory or intuitive judgment leads one to believe that there should be a relationship between two variables, but the data simply do not indicate any relationship. When this is the case, there is the possibility that some other variable or variables are acting to hide or suppress the relationship. These suppressor relationships often take the form of "A shows no relationship to B because A is negatively related to C which in turn is positively related to B." Hence A is positively related to B when one controls for the effects of C.

Take the following hypothetical example of a marketing study attempting to determine what types of families purchase second automobiles. The initial investigation of the data surprisingly found that there was almost no correlation ($r = .081$) between a measure of family need for a second car and whether or not a family owned a second automobile. However upon closer scrutiny, the researchers became suspicious of the possibility of a confounding or masking variable. They noticed that family income was strongly related to the purchase of a second car ($r = .55$) and that family income was, on the other hand, somewhat negatively related to need for a second automobile ($r = -.32$). A partial coefficient was then computed between need and purchase, removing the effects of family income from both of these variables in order to determine if income had acted to hide a potentially important relationship.

When this partial was computed, it became clear that family income was indeed masking a rather strong relationship ($r_{12.3} = .32$) between need and a purchase. From this partial the researchers were able to state that at any given level of family income, need for a second automobile explained about 10 percent of the variance in the purchases.

This introduction to partial correlation hopefully indicates how versatile and useful a research tool partial correlation can be. In the first instance it served to help locate a spurious correlation, in the second it enabled us to determine the importance of a particular intervening variable, and in the third its ability to help uncover a relationship where none appeared to exist was demonstrated. The types of analyses which can be accomplished with partial correlation are numerous, and this very brief introduction is not meant to be in any way a substitute for the excellent literature which exists on the subject.

19.2 PARTIAL CORR PROCEDURE CARD

Subprogram PARTIAL CORR is called and activated by a procedure card with the control words PARTIAL CORR followed by a specification field (beginning in or after column 16) containing three types of information which must be entered in order to specify the desired partial correlations. First, one or more pairs of independent and dependent variables for which one or more partials are desired must be entered (these *do not* include the control variables), and they are referred to as the *correlation list*. Second, one or more control variables which are to be used as controls for the variables in the correlation list must be entered, and this portion of the specification field is referred to as the *control list*. Third and finally, the user must enter the *order values* indicating the order of partials desired from the correlation and control lists. The general format of the PARTIAL CORR card is then as follows:

1	16
PARTIAL CORR	correlation list BY control list (order values) / correlation list BY control list (order values) / ...

Because of the complexity of this card, we will break its explanation down into the individual portions of the card.

APPENDIX B

POLITICAL AND DEMOGRAPHIC
CHARACTERISTICS - CITY SEATS

pro- fics	Matr- culants	Pre- School- ers	School- children	Tert- iary Stud- ents	Mobil- ity
6.2	15	7	16	18	44
4.2	12	7	21	25	25
5.2	14	6	24	21	29
1.4	22	14	25	6	61
5.9	32	6	19	18	46
7.3	19	9	28	15	34
4.8	22	12	26	18	44
5	30	7	25	16	41
0	22	10	31	9	44
1.9	28	12	24	17	47
5.5	13	11	24	13	45
7.4	16	9	28	17	34
7.2	24	6	24	16	40
5	21	8	23	16	39
3	13	9	25	19	32
4.2	15	10	24	17	38
9.3	18	16	22	9	57
5.7	28	7	22	19	39
6.5	14	7	25	22	30
5	19	6	21	15	40
0.8	24	13	33	8	56
9.1	26	15	27	5	64
6.2	21	8	21	22	46
3.3	12	8	20	21	35
7.1	24	13	30	6	51
5.7	10	8	20	18	39
5.8	10	7	23	24	28
8.2	15	14	29	12	46
8.3	10	8	22	19	36
3.5	9	7	20	23	31
9.5	19	6	25	9	50
8.9	32	7	16	29	49
7.4	21	8	17	21	47
6.8	19.1	9.1	23.6	16.5	42

APPENDIX 8

POLITICAL AND DEMOGRAPHIC
CHARACTERISTICS - COUNTRY SEATS

	1975 ALP 2PP Vote	1973 ALP 2PP Vote	1973- 75 Swing	Blue- Coll Urban	Agri- cult ural Work Force	Blue- Coll Rural	Mid- White Coll Urban	Pers- onal Vote	Matri- culants	Pre- School ers	School Children	Tert- iary Stud- ents	SAHT Tenants	18-24 Year Olds	25-34 Year Olds	35-44 Year Olds	45-54 Year Olds
Alexandra	22.3	30	7.7	31	46.1	13.6	5.3	3.3	8.7	8.7	22.3	0.88	0.47	14.4	16.1	16.2	19
Chaffey	31.1	37.5	6.4	38	39.1	14.3	6.1	5.1	6.5	10.8	23.9	0.22	5.47	17.6	19.9	19.3	18.7
Eyre	38.3	45.1	6.8	51.7	20.9	15.9	4.9	5.3	12.5	12.7	21.1	0.16	6.4	18.8	26.3	21	16
Flinders	29	32.7	3.7	36.1	56.2	9.9	5.6	3.2	5.7	12.5	24.6	0.14	5.8	18.1	22.8	19.1	16.9
Goyder	23.1	28.9	5.8	26.7	51.8	12.3	4.1	2.9	5.1	10.7	23.3	0.16	0.61	15.1	18.6	18	18.1
Kavel	23.2	29.6	6.4	43.2	29.9	9.9	6	5.8	7.6	9.3	21.8	0.6	2.64	15.3	17.5	16.8	18.6
Light	32	34.4	2.4	39.2	31.9	9.8	6.9	2.3	7.1	8.8	21.7	1.27	4.84	15.3	15.8	16.3	18.4
Mallee	21.4	35.7	14.3	29.1	55.2	15.6	4.1	3.3	5.9	11.7	24.7	0.11	1.75	17.6	21.5	19.4	17.4
Mt. Gambier	42.9	59	16.1	46.6	20.8	7.4	7.6	1	7.4	11.1	24	0.17	11.29	19.5	20.8	18.6	18
Murray	35.8	37.7	1.9	47.7	24.7	6.5	6.7	2.2	6.6	10.5	22.8	0.31	6.76	17.6	19.5	17	17.8
Rocky River	31.5	37.8	6.3	39.2	33.8	7.6	5.5	4	4.8	10.5	22.9	0.15	5.98	14.4	18	17.8	17.1
Stuart	66.9	75.2	8.3	63.6	4	1.7	8.1	0	6	10.4	23.5	0.22	16.62	18.9	19.8	18.9	17.7
Victoria	28.3	44.3	16	38.5	37.8	12	5.5	5.8	8.6	11.8	25.2	0.19	12.17	18.7	22.3	18.9	17.8
Whyalla	69.7	78	8.3	66.1	0.5	1.4	7.6	1	12.7	13.5	26.9	0.33	56.54	21.5	28.7	22.1	14.5
MEAN	35.4	43.3	7.9	42.6	32.3	9.4	6	3.2	7.5	10.9	23.5	0.35	9.81	17.3	20.5	18.5	17.6
SEAT	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31	V32	V33	
	55-64 Year Olds	65+ Year Olds	O'seas Born	UK- Born	Greek- Born	Italian Born	German Born	Yugo- slav Born	"Other Europe" Born	Catho- lics	Church of England	Luth- eran	Metho- dist	Pres- byter- ian	Agnos- tic	Mobil- ity	
Alexandra	17	17.2	9.1	5.7	0.04	0.7	0.6	0.06	0.2	9	24.6	4.1	33.3	1.6	12.5	36.6	
Chaffey	13.3	11.2	12.7	3.3	3.6	1.3	0.9	1	0.4	13	19.2	16.9	18	1.6	7.7	39.2	
Eyre	10.2	7.7	20.3	9.9	1.1	0.7	1.2	1.5	1.2	19.7	21.3	6.2	18.5	4.2	19.2	55.2	
Flinders	12.8	10.3	7.4	3.8	0.3	0.3	0.6	0.6	0.3	15	21.9	7.9	32	1.3	13.3	38.9	
Goyder	15.2	15.1	17.7	2.7	1.3	1.5	0.4	0.3	0.6	13.6	16.5	7.7	41.9	1.1	8.8	28.7	
Kavel	15.1	16.6	9.6	2.2	0.07	0.9	1.2	0.3	0.5	7.5	15.3	35.8	16	2.6	10	31.2	
Light	16.5	17.7	8.2	4.6	0.09	0.7	0.8	0.2	0.3	16.7	16.5	21.5	25.7	2.1	9.3	32.4	
Mallee	13.6	10.4	5.3	2.8	0.1	0.3	0.5	0.1	0.2	11.1	20.3	11.7	29.7	3.7	11.3	34.3	
Mt. Gambier	4.7	10.5	12.5	5.1	0.4	1.5	0.9	0.7	0.6	23.6	18.8	3.4	13.6	19.8	12	39.9	
Murray	14.3	13.8	9.4	4	0.1	1.4	0.9	0.1	0.3	12.1	17	17.4	37.3	3.8	12.4	40.2	
Rocky River	15.1	17.6	5.3	2.8	0.6	0.2	0.4	0.1	0.1	15.8	14.4	4.7	41.2	1.8	9.8	34.4	
Stuart	13.7	11	12.3	5.5	0.8	2.2	1.1	0.3	0.3	23	22.5	2.9	22.9	3.5	13.4	38.5	
Victoria	12.7	9.6	10.9	28.7	1.2	1.2	0.6	0.5	0.5	17.3	21.9	5.2	21.1	12.8	11.2	39.8	
Whyalla	8.8	4.3	41.5		0.9	1.1	2	2.4	2.1	22	25.2	3.2	14.8	6	16	60.2	
MEAN	13.1	12.4	12.3	6.2	0.8	1	0.9	0.6	0.5	15.7	19.7	10.6	26.1	4.7	11.9	39.3	

APPENDIX C

METHOD:

For each State seat, the Two Party Preferred (2PP) vote received by the Labor group in the Legislative Council has been subtracted from the 2PP vote received by the Assembly candidate.

ASSUMPTION:

- | | | | |
|----|--|---------------|--|
| 1. | (ALP Legislative Council vote by Seat) | <u>equals</u> | ALP Party support in that seat. |
| 2. | ALP Assembly vote by seat | <u>equals</u> | ALP Party support plus personal vote of the candidate. |

CONCLUSION:

- | | | | | | |
|----|-------------------|--------------|------------------------------|---------------|---|
| 1. | ALP Assembly vote | <u>minus</u> | ALP Legislative Council vote | <u>equals</u> | ALP Assembly Candidate's personal vote. |
|----|-------------------|--------------|------------------------------|---------------|---|

In 1975 however, the Labor Team in the Legislative Council received the donkey vote. Therefore, whenever the Labor Assembly candidate did not receive the donkey vote, this had to be estimated and added to his original vote. Thus, the donkey vote appeared in both sides of the equation and cancelled itself out. The donkey vote was calculated as follows :-

$$\text{DONKEY VOTE} = 0.5\% + (\text{ALP 1975 Leg. Council 2PP vote} - 45) \times 0.1$$

(Where the 1975 Legislative Council 2PP vote fell below 45% the donkey vote was assumed to equal a constant, 0.5%)

Thus, a 40% and 45% Labor vote indicated a 0.5% donkey vote, while a 55% Labor seat was assumed to have a donkey vote of :-

$$0.5\% + (55 - 45) \times 0.1 = 1.5\%.$$

1. It should be noted that any persons who cast a personal vote for a Labor Assembly candidate and who then (for reasons of convenience) also cast a vote for the Labor Council team are not included in that Labor Assembly candidate's personal vote score. This factor (personal vote 'leakage') has disturbing implications for the Labor vote in the new marginal seats of Newland, Mawson and Coles.

A good deal of time was spent trying to find a more rigorous method of calculating the donkey vote. In particular, I used census data by electorates to measure education, unemployment, number of persons who had never attended school - even the number of persons who had inadequately completed census forms - to no avail.

For country seats, I felt the donkey vote may have been related to sparsity of population, but census data, and figures for the average number of voters per polling booth produced inconsistent results. I finally settled on the same formula for all city and country seats after calculating that country members who received the donkey vote received an average personal vote 0.52% higher than their less-fortunate colleagues. This 0.52% figure for the country seats (excluding Mt. Gambier, Pt. Pirie, Whyalla and Stuart), deviated from the formula based result by only 0.02%.

Note: The personal vote scores which follow represent the personal vote of the Labor Candidate, relative to his/her non-Labor opponents.

RESULTS: TABLE 1

CANDIDATE'S PERSONAL VOTE - STATE.

Rank	Seat	Personal Vote	Rank	Seat	Personal Vote
1	NORWOOD	5.0	27	LIGHT	-1.6
2	ELIZABETH	4.4	28	HANSON	-1.6
3	TEA TREE GULLY	3.4	29	GLENELG	-2.1
4	MAWSON	2.7	30	BRAGG	-2.2
5	BRIGHTON	2.4	31	MURRAY	-2.2
6	ASCOT PARK	2.4	32	HEYSEN	-2.3
7	UNLEY	2.0	33	PRICE	-2.5
8	MILLICENT)	2.0	34	MITCHAM	-2.7
9	COLES)	1.6	35	GOYDER	-2.8
10	SEMAPHORE	1.4	36	GOUGER	-3.0
11	SALISBURY	1.3	37	FLINDERS	-3.2
12	WHYALLA	1.0	38	ALEXANDRA	-3.3
13	MITCHELL	0.9	39	FISHER	-3.5
14	ADELAIDE	0.9	40	FROME	-3.6
15	MT. GAMBIER	0.8	41	ROCKY RIVER	-4.0
16	HENLEY BEACH	0.4	42	CHAFFEY	-5.1
17	GILLES	0.3	43	EYRE	-5.3
18	ALBERT PARK	0.1	44	MALLEE	-5.6
19	TORRENS	0.0	45	VICTORIA	-5.8
20	SPENCE	-0.1	46	KAVEL	-5.8
21	FLOREY	-0.2	47	PIRIE	-23.2
22	PLAYFORD	-0.7			
23	STUART	-0.9			
24	PEAKE	-1.0			
25	DAVENPORT	-1.6			
26	ROSS SMITH	-1.6			

INTRODUCTION

Two questions were posed with regard to the newly-created seat of Rocky River. They were :

- 1) Is there any chance that the Labor Candidate can win the seat at the next State election?
- 2) Has there been any basic changes to the structure of the class-vote in the Moonta-Kadina-Wallaroo areas during the past decade?

METHOD

Statistical summaries for the Moonta, Kadina and Wallaroo Urban areas were obtained from the 1966 Census and the 1971 Census results. A statistical summary for the new Rocky River electorate was also obtained from the 1971 results.

Regression formulae for the 1973 ALP vote (see "A Model of South Australian Political Behaviour") were applied to the Rocky River statistical summary to obtain a predicted vote for the new seat, based on 1973 demographic alignments (i.e. "on 1973 figures"). The same formulae (for the 1973 ALP vote) were applied to the 1966 and 1971 Census Urban summaries for Moonta, Kadina and Wallaroo.

RESULTS

	1966 ALP vote on 1973 voting figures (%)	1971 ALP vote on 1973 voting figures (%)
MOONTA	38.9	23.2
KADINA	45.1	43.1
WALLAROO	60.0	59.1
AVERAGE	48.0	41.8
ROCKY RIVER	not available	38.0

DISCUSSION

Question 1.

The Labor candidate has no chance of winning Rocky River in 1977/78, even if the ALP regains its 1973 levels of support in the country. The Rocky River 1973 vote of 38% is based on a Regression Equation which explains 100% of the variance and has a plus or minus error of only 0.18%.

In other words, there is only a 5% chance that the ALP candidate will obtain a vote as high as 38.18%. The chances of an "error" in the predication as high as 12% (the error needed for a 50% ALP vote) are impossible to predict using conventional statistical tables. However, I would guess the odds would be several hundred million to one against.

The only remaining chance for the ALP candidate would be to win the seat with a 12%-plus personal vote against the "sitting" Liberal Member. Again, the odds against this are prohibitively-high. There is simply no scope for a personal vote of this magnitude in Rocky River.

Question 2.

The class-vote in the Moonta-Kadina-Wallaroo area declined by an average of 6.2% between the 1966 Census and the 1971 Census. If this rate of decline continued at the same rate between 1971 and 1976, the class-vote in Moonta-Kadina-Wallaroo would have dropped by well over 10% during the past decade.

CONCLUSION

The new seat of Rocky River is a hopeless electoral proposition for the Labor Party at the next State election.

JUNE 14, 1977

APPENDIX E.A.N.O.P. SURVEY - POLITICAL INTERPRETATION

Swinging voters, like any other political or demographic group, are distributed unevenly across electorates. In South Australia, swinging voters tend to live in our marginal electorates. Therefore, it is natural to expect a sample of marginal State seats to contain a bias towards political volatility.

This is certainly the case with the A.N.O.P. 1976 sample, as can be seen in the table below. The table lists the seat/subdivision sampled, the estimated A.N.O.P. Two-Party-Preferred Swing, and the Volatility Index. The Volatility Index measures the concentration of swinging voters - and subsequently the potential for swing - in each State electorate. City and country seat indices are measured in relation to city and country mean swings respectively.

SEAT/SUB-DIVISION	1976 ANOP % 2PP SWING	SWING RANK	VOLATILITY INDEX	VOLATILITY RANK
Norwood	1.4	1	84	1
Semaphore	5.4	2	89	2
Hanson East	7.1	3	111	3
Mawson	7.2	4	168	4
Modbury North	7.8	5	204	5
CITY SAMPLE	5.8		131	
Mt. Gambier	7.5	1	148	1
COUNTRY SAMPLE	7.5		148	

TABLE 1

As Table 1 shows, the swing back to the Labor Party has varied in direct relation to the Volatility Index. Table 1 also shows that the A.N.O.P. sample is biased towards high levels of volatility by the inclusion of Hanson East, Mount Gambier, Mawson and Modbury North. The estimate of the pro-Labor swing actually indicated by the A.N.O.P. survey therefore must be adjusted for this bias before it can be applied to either the city or country areas.

The raw swing figures given in the A.N.O.P. report - especially the crude first-preference figures which have been distorted by the destruction of the LM - provide an overly-optimistic assessment of the electoral recovery made by the Labor Party since 1975.

A.N.O.P. SURVEY - 2

Once the A.N.O.P. swing figures have been adjusted for the volatility of the sample, the pro-Labor swings in the city and country are shown to be 4.4% and 5.1% respectively. Thus, the level of support for the State Branch of the Labor Party in late 1976 was higher than it was in the elections of 1975 and 1970, but still lower than our 1973 peak. This is shown below in Table 2.

REGION	1970 ALP 2PP %	1973 ALP 2PP %	1975 ALP 2PP %	1976 ALP 2PP % (ANOP)
City	56.8	59.5	54.3	58.7
Country	43.2	43.3	35.4	40.5
State	52.9	54.9	49.8	53.5

TABLE 2

The next step is to determine the impact of these city and country swings on key electorates. This is done below in Table 3, where I have applied the mean swing figures to key electorates via their individual Volatility Indices. This produces a result for sampled seats slightly different to the A.N.O.P. figure, because of the "averaging" effect of my calculations.

SEAT	1976 Est. ALP 2PP%	VOLATILITY INDEX	MEAN % SWING NEEDED IN CITY OR COUNTRY TO CHANGE HANDS
Todd	64.2	168	8.5
Norwood	60.7	84	12.7
Newland	60.4	204	5.1
Mawson	59.3	185	5.0
Unley	58.0	79	10.1
Hartley	56.9	75	9.2
Mt. Gambier	52.1	195	1.0
Coles	50.7	123	0.6
Morphett	49.4	106	0.6
Hanson	47.5	111	2.3
Glenelg	45.5	85	5.3
Eyre	42.6	92	8.0

TABLE 3

The reader will note that the swing needed for a seat to change hands is not simply a function of the margin by which it was won or lost at the preceding election. One must also take into account the volatility of the electorate. For example, Molly Byrne in 1976 would have polled some 64% in Todd, and Des Corcoran would have obtained about 57% in Hartley. Yet, because Todd is almost three times as volatile as Hartley, a 9% anti-Labor city swing on A.N.O.P.'s 1976 figures would have unseated Molly Byrne and left Des Corcoran as the member for Hartley.

In Table 4, below, I rearranged the data from Table 3, to emphasise vulnerable new seats currently "held" by both parties. Seats not actually listed in Table 4 have been intentionally excluded (e.g. Torrens, Chaffey).

RATIO SEATS LIB : LABOR	LABOR SEATS	PRO-LIB SWING %	LIBERAL SEATS	PRO-LAB SWING %	RATIO SEATS LIB : LABOR
21:26	Norwood	12.7			
22:25	Unley	10.1			
23:24	Hartley	9.2			
*** LIB GOVERNMENT ***			Eyre	8.0	32:15
24:23	Todd	8.5			
25:22	Newland	5.1	Glenelg	5.3	31:16
26:21	Mawson	5.0	Hanson	2.3	30:17
27:20	Mt. Gambier	1.0			
28:19	Coles	0.6	Morphett	0.6	29:18

TABLE 4

Table 4 shows the State Labor Government should "win" 28 new State seats on the 1976 A.N.O.P. results. I should point out that the "results" in Coles and Morphett do not fall within a 95% confidence interval. In other words, both Coles and Morphett were too close to call. Given average luck, we would have won one of the two, probably Coles. A lot will depend here on the competence of our two candidates and the efficiency of their "grassroots" campaigning.

The seat of Hanson is a dubious one for Labor. We need a city pro-Labor swing of 2.3% on top of the swing already recorded by A.N.O.P. to win Hanson. This would mean the Labor Party would have to poll 61% of the metropolitan vote - 1.5% higher than our 1973 vote (see Table 2). The only alternative for the Labor candidate is to completely whittle away Heini Becker's Personal Vote lead of some 1.6%. Becker should also be insulated by a donkey vote of about 0.7%. In short, the odds favour Becker and the Liberal Party.

The evidence indicates that the new seats of Eyre and Glenelg do not warrant serious examination as prospects for Labor.

The new seat of Mount Gambier should have been won by Labor on the 1976 A.N.O.P. results. The new seat is even more volatile than the old seat, so the 2PP pro-Labor A.N.O.P. swing of 7.5% in the old Mount Gambier would be recorded as almost 10% across the new Mount Gambier. Any increase in our A.N.O.P. 1976 country vote would "sew up" Mount Gambier for the Labor Party. It should be remembered that a country pro-Labor swing of only 1% on the 1976 A.N.O.P. figures would reach 2% in Mount Gambier. Unfortunately I cannot be too precise about the outcome in Mount Gambier as A.N.O.P. did not print their questionnaire with the report. The unknown factor is, of course, "did the Mount Gambier sample state its voting intention, bearing in mind the identity and popularity of the local candidates." If they did, we have little to worry about. If they didn't take account of a new Personal Vote of about 2% for Liberal MP Harold Allison, then there is cause for concern, particularly if there is some stagnation in our country vote.

The new seats of Mawson and Newland will be won or lost together. If we wish to regain a working majority in South Australia, both seats have to be won. This we would have done even on 1975 figures. If, however, our vote fell 1% below 1975 levels, we would lose both seats. As I will point out below, the two seats are demographic and political "twins". The same campaign issues should therefore apply equally to both seats.

If we lose the seat of Todd, we will lose Government. With Molly Byrne's personal vote of 3.4%, this is difficult to imagine.

The remaining seats of Hartley, Unley and Norwood are all - by the conventional definitions - "marginal". And yet when one allows for their extreme stability, they are all more accurately classified as "safe". In fact, the Labor vote in all three is probably still undergoing long-term improvement, due to the naturalisation of resident aliens.

A.N.O.P. SURVEY - KEY DEMOGRAPHIC GROUPS

The key metropolitan seats which will decide the fate of the Government - and the size of its majority - are Todd, Newland and Mawson. These three seats are remarkably similar: politically, demographically, even geographically.

This homogeneity ensures that the Party can economise on its investment in electioneering: in terms of content, quantity and style of advertising and the range and cost of issues canvassed. One or two election issues, carefully targetted and attractively presented, will win an equally-large share of votes in all three key seats. The reverse, of course, also applies.

What sort of swinging voters, then, do we have in Newland, Todd and Mawson? The profile below is based on general and seat-specific information contained in the A.N.O.P. 1974 Swinging Voter Survey, my Political and Demographic Analysis of new city seats and my S.A. model of Political Behaviour.

- CLASS: Not significant; very slight bias toward Upper-White Collar and Middle-White Collar Workers.
- HOUSING: Private housing (there is some public rental housing in Todd). Residents are typically in the first few years of home purchase. They will usually have lived in their house for only a relatively short period - about five years.
- FAMILY STATUS: Parents of two pre-school children, or very young school-age children.
- EDUCATION: Not significant for electoral volatility. (Newland voters are better-educated.)
- AGE: 25-34 years (Very important) There are large concentrations of this age group in all three seats.
- RELIGION: Religion is not a significant guide to electoral volatility in the city. (Both Newland and Mawson are strongly Protestant.)
- ETHNICITY: Not important. (Both Newland and Mawson contain

large concentrations of British-born persons.)

By far the most interesting insights into the character of the outer-urban swinging voters are provided by an examination of the interaction between Class, Sex, Age and Workforce Participation

First, let us examine what we already know about each of these variables, in isolation.

Class: Taking Newland as an example, the Class-Vote is 54%. This is a figure based on the percentage of the total workforce engaged in "blue-collar" or "working-Class" jobs. Class in S.A. is an extraordinarily-powerful indicator of future voting intention - almost as strong in fact as previous vote.

Age: Age is another very useful indicator of voting behavior. In the city, the percentage of those 18 years and over in the 25-34 year age group is the strongest single indicator of electoral instability. Older people tend to be politically stable.

Workforce Participation: About 67% of voters in seats such as Newland are in the workforce. This comprises 91% of male voters and 43% of female voters.

Sex: Females are politically more volatile than males. 57% of the 1974 A.N.O.P. swinging voter panel were female.

Now, let us consider how these variables interact in, for example, Newland:

Class x Sex: Women work in occupations typified by political conservatism. By class, working women in Newland vote 33% Labor; working men vote 63% Labor (Mean Labor vote equals 54%). The conservative voting behavior of Australian women therefore can be best understood in light of their occupational class.

Class x Age: Older people are more conservative, partly because of the psychological process often termed "senescence". Other, probably stronger, factors include death and promotion. Deaths due to accidents, terminal illnesses, drug addiction and violence strike down a disproportionate number of Working-class men and women in Adelaide. Middle-class and Upper-class Adelaide citizens lead more healthy and trouble-free lives in prosperous suburbs overflowing with doctors, lawyers, teachers and social workers. Promotion, also provides for a steady increase in the Liberal vote in the older age groups.

Sex x Age x Workforce

Participation: The following table (Table 5) uses 1971 Census data for the new seat of Newland to illustrate the age distribution of males and females in the workforce in a fast-growing, volatile, outer-metropolitan electorate.

<u>Age</u>	<u>Male</u>	<u>Female</u>
15-19	58	61
20-24	95	52
25-29	98	33
30-34	98	41
35-39	98	46
40-44	98	53
45-49	96	53
50-54	97	42
55-59	97	29
60-64	83	12
65+	19	3
TOTAL	90.6	43.1

Table 5 Male and female workers as a percentage of the population (by age groups).

We can clearly see from the Table 5 the impact of the child-rearing cycle on working wives. The percentage of wives in the workforce reaches a low of 33% for the 25-29 year olds and slowly climbs back up to a relatively-high 53% for 40-49 year olds.

To sum up, I present below in Table 6, a simplified picture of a small six-home street in a typical Newland suburb. The street contains six houses, and six married couples.

Four of the husbands (all of whom have jobs) normally vote Labor, two normally vote Liberal. However Husband Number 4 two years ago was promoted to foreman in a small plastics factory. He was previously a fitter and turner and a staunch Labor Voter. Now he is not so committed. He voted Liberal in December 1975. His wife, formerly a nurse, usually votes Labor in State elections and Liberal in Federal Elections. Of the remaining five wives, two are at home, looking after their pre-school-aged children. One of these was formerly a clerk, and a Liberal, then LM voter. The other was a cleaner and usually always votes Labor, except for the 1975 Federal Election. The remaining three wives are all in the workforce and vote - with their husbands - on normal class lines: two Liberal, one Labor.

HOUSE NO.	1	2	3	4	5	6	
H U S - B A N D	EMPLOYED	EMPLOYED	EMPLOYED	EMPLOYED	EMPLOYED	EMPLOYED	JOB STATUS
	Metal Worker	Motor Mechanic	Carpenter	Foreman	Employer	Clerk	Normal Occupation
	Labor	Labor	Laborer	Pro-Labor Swinger	Liberal	Liberal	Normal Vote
W I F E	EMPLOYED	HOUSE-WIFE	HOUSE-WIFE	HOUSE-WIFE	EMPLOYED	EMPLOYED	JOB STATUS
	Cook	Cleaner	Clerk	Typist	Steno.	Typist	Normal Occupation
	Labor	Pro-Labor Swinger	Pro-Liberal Swinger	Pro-Liberal Swinger	Liberal	Liberal	Normal Vote

TABLE 6. An example of possible interaction between the Class-Vote, the Swinging Vote and Workforce Participation (by sex) in six typical Newland households. Abstracted from available census and market research data.

Table 6 provides a fair summary of the presently-available knowledge on the voter variously described as "swinging", "switching", "undecided" or "volatile". In a State where class-voting links are very strongly established, the swinging voter can best be identified as being temporarily without an occupational class (due to child-rearing) or in transition from one class to another (due to promotion).

In addition, he - or rather she - will be young (late twenties) with a very young family. She will live in a new home, in a new, developing suburb. Perhaps most importantly, from a communications point of view, the swinging voter will tend to be found in the home during the day. There, she can probably be best contacted by commercial radio in the morning and commercial television in the afternoon.

(The advertising agency should have some useful information on this last point.)

A.N.O.P. SURVEY - ISSUES

GENERAL ISSUES:

Unemployment and Inflation emerge as the two dominant issues for all voters, Labor, Liberal or uncommitted. Two out of every five swinging voters consider Unemployment to be the State Government's "most serious" or "second most serious" problem. For inflation, one out of every four swinging voters feels the same concern.

When dealing with this problem in speeches, press releases, leaflets etc. you should consider the following :

Four out of ten voters think the State Government is doing a bad job of reducing unemployment. And yet half of these persons believe the State Government should reduce unemployment by implementing measures which are beyond State control - legally or financially. We could solve both these problems by a series of statements designed to convince swinging voters we are generally in favour of their unemployment remedies, but that the Federal Government has to foot the bill. Such a speech could be made for example, after the release of higher unemployment figures, calling on the Federal Government to initiate stimulatory fiscal measures (e.g. tax cuts).

* * *

SEAT-SPECIFIC ISSUES:

In addition to the General Issues outlined above, the Party should stress the following specific issues in the seats of Newland, Mawson, Coles and Todd.

Education

Federal funding cuts and their impact on the construction of new primary schools in fast-growing

areas.

Roads and

Public Transport Improvements planned for roads to fast-growing areas (e.g. North-East Transport Corridor); Fares and general public transport services to outer, developing suburbs.

Housing

The availability and costs of housing for young married couples; interest rates, general building costs. It should be stressed that the State Government has no control over interest rates.

Water

Construction of filtration plants and the attention paid by the State Government - despite the apathy of present Federal Government - to this problem in South Australia.

* * *

For the older seats of Morphett, and Glenelg, Torrens and Hanson, the Party would gain more votes by concentrating on the General Issues of Unemployment and Inflation, at the expense of specific issues. However, some specific issues should have limited impact on these four older "marginal" seats. They are:

Housing

Large numbers of voters in these seats are young married flat-dwellers, saving to obtain a deposit on their first homes. Recent moves by the State Government to make houses cheaper, and more readily available should be stressed here.

Pensions and
Social
Problems

Unfortunately the sort of voters helped by the State Government's moves in this area are much older, and more set in their political ways. From an opportunistic point of view, \$1 spent on a reduction of interest rates is probably worth \$10 spent on pension increases.

Law and Order
and

Permissiveness This appears to be a relatively-strong issue among

older voters. However, any general campaign on this sort of "cardboard issue" may well lose us votes among volatile voters in the key seats. Therefore, any mention of this issue should be contained - if used at all - in our older, more stable electorates (within the Adelaide Inner Fire Ban District).

* * *

In our marginal country seats, the General Issues of Unemployment and Inflation could usefully be stressed. However, Labor campaigns in these seats have traditionally been highly individualistic and entrepreneurial in nature. I feel therefore we could usefully be guided on the specific issues in Mount Gambier and Eyre, largely by reports from our ALP candidates.

This point aside, our Eyre candidates should come out very strongly against the Federal Government's determination to dismantle our shipbuilding industry. Lower tariffs and their implications for general Industrial Development in South Australia could also be mentioned.

It should also be remembered that the country swinging voter appears to be younger, better-educated and more middle-class than the country non-swinging voter. Also, there appears to be some "overlap" between the stereotype of the country Labor voter and the country swinging voter. Labor's country candidates can therefore pitch their specific-issue campaigns simultaneously to both swinging voters, and the party faithful. In so doing, they could stress their concern with traditional Labor issues such as job security, the cost of rental, or private-purchase housing, and the availability of schools.