

Staff Papers Series

STAFF PAPER P79-32

SEPTEMBER 1979

User's Guide to SEGREG*

Thomas F. Stinson
Kweiwu Fang
Andrea Lubov



Department of Agricultural and Applied Economics

University of Minnesota
Institute of Agriculture, Forestry and Home Economics
St. Paul, Minnesota 55108

User's Guide to SEGREG*

by

Thomas F. Stinson
Kweiwu Fang
Andrea Lubov

*Research on this project was supported in part by the Economics, Statistics, and Cooperatives Service, U.S. Department of Agriculture; Farmers Home Administration, U.S. Department of Agriculture; and the U.S. Environmental Protection Agency. The authors gratefully acknowledge the assistance of Prof. R. Dennis Cook, Department of Applied Statistics, University of Minnesota.

Staff Papers are published without formal review within the Department of Agricultural and Applied Economics.

Introduction

SEGREG is a computer program which enables users to perform segmented regressions on ordered data sets, given that the boundaries of each segment, or subset of the data, are unknown. The program searches the data for the switching point that defines the two best subsets and for the two switching points that define the three best subsets.

Results obtained from SEGREG are similar to those obtained using spline functions. There is, however, an important difference in the two techniques. With spline functions the user must specify the boundaries of each ordered subset of observations. SEGREG, on the other hand, automatically searches the data to define those ordered subsets. A search for a switching point, or points, is appropriate when the researcher does not know at what point(s) the relationship changes, or if it changes at all.

Segmented regression may be applied to answer a variety of questions. Applications include identifying city populations at which expenditure functions undergo major shifts, determining when the relationship between the amount an animal is fed and its weight gain changes, and finding the length of time a patient's behavior is affected after a drug dosage is administered.

The SEGREG Model

SEGREG fits all possible ordered pairs and ordered triplets of regression equations to a set of data. The set of data must contain less than 300 observations and it must be ordered by increasing size of X_1 , the independent variable.

The ordered pairs of regressions will be of the form

$$(1) \quad Y_i = \hat{a}_1 + \hat{b}_1 X_i + e_i \quad \begin{matrix} n_j \\ i=1 \end{matrix} \quad \begin{matrix} n_j \\ i=1 \end{matrix} \quad \begin{matrix} n_j \\ i=1 \end{matrix}$$

$$(2) \quad Y_i = \hat{a}_2 + \hat{b}_2 X_i + e_i \quad \begin{matrix} n \\ i=n_j+1 \end{matrix} \quad \begin{matrix} n \\ i=n_j+1 \end{matrix} \quad \begin{matrix} n \\ i=n_j+1 \end{matrix}$$

The ordered triplets of regression equations will be of the form

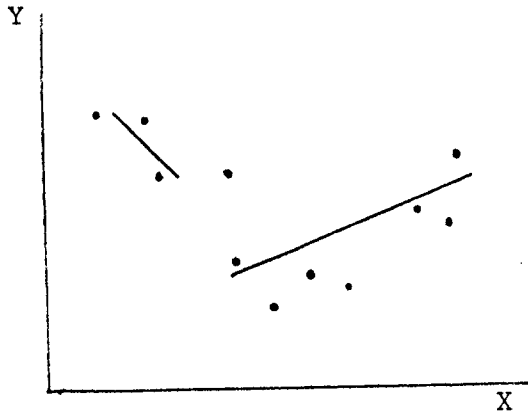
$$(3) \quad Y_i = \hat{a}_3 + \hat{b}_3 X_i + e_i \quad \begin{matrix} n_k \\ i=1 \end{matrix} \quad \begin{matrix} n_k \\ i=1 \end{matrix} \quad \begin{matrix} n_k \\ i=1 \end{matrix}$$

$$(4) \quad Y_i = \hat{a}_4 + \hat{b}_4 X_i + e_i \quad \begin{matrix} n_m \\ i=n_k+1 \end{matrix} \quad \begin{matrix} n_m \\ i=n_k+1 \end{matrix} \quad \begin{matrix} n_m \\ i=n_k+1 \end{matrix}$$

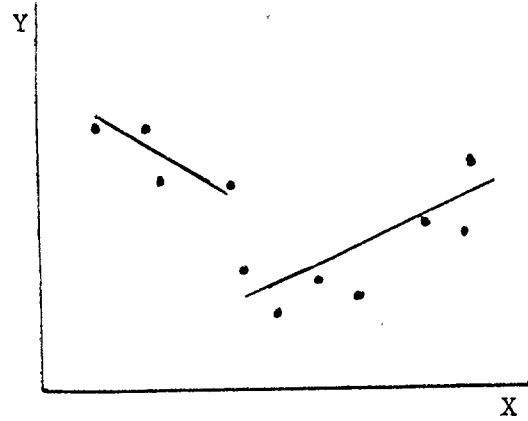
$$(5) \quad Y_i = \hat{a}_5 + \hat{b}_5 X_i + e_i \quad \begin{matrix} n \\ i=n_m+1 \end{matrix} \quad \begin{matrix} n \\ i=n_m+1 \end{matrix} \quad \begin{matrix} n \\ i=n_m+1 \end{matrix}$$

Figure 1 illustrates the calculation of all possible ordered pairs of regression equations when $n = 11$. The first pair of equations is obtained by regressing X on Y for the observations with the 3 smallest values of the X variable and for the observations with the largest 8 values of the X variable; the second pair is the observations with the 4 smallest and 7 largest X values; and so on.

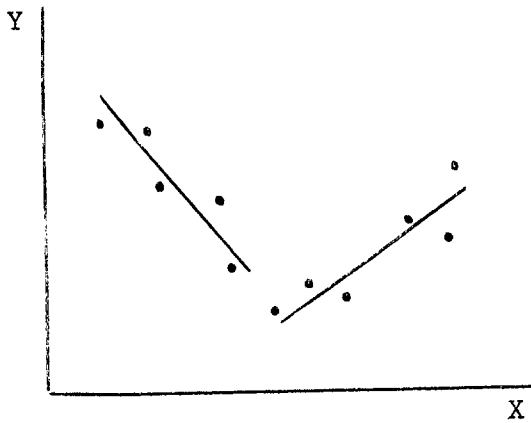
The best switching points are defined by the ordered pair and the ordered triplet with the smallest sum of the residual sum of squares. In all, $(n-2m-1)$ ordered pairs and $(n-2m-1)(n-2m-2)$ ordered triplets of regressions are calculated where m is the user-defined minimum number



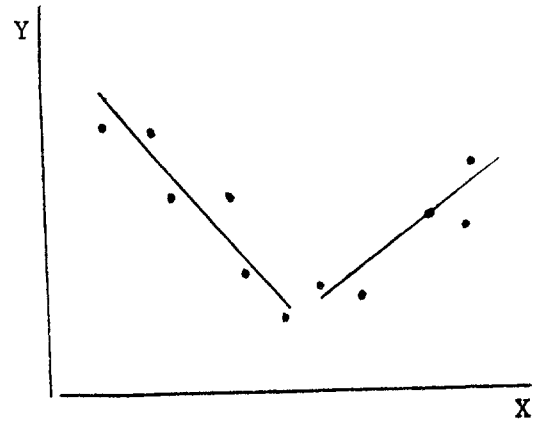
FIRST REGRESSION:
SMALLEST 3, LARGEST 8



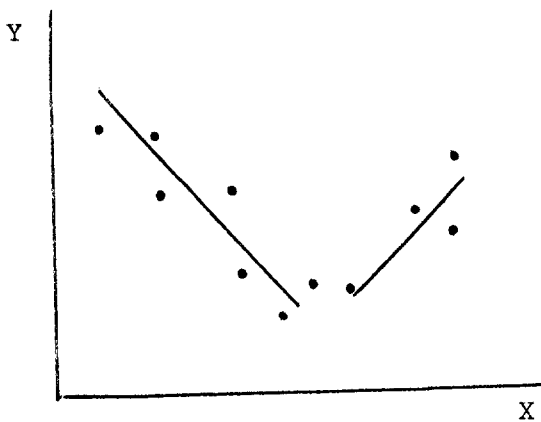
SECOND REGRESSION:
SMALLEST 4, LARGEST 7



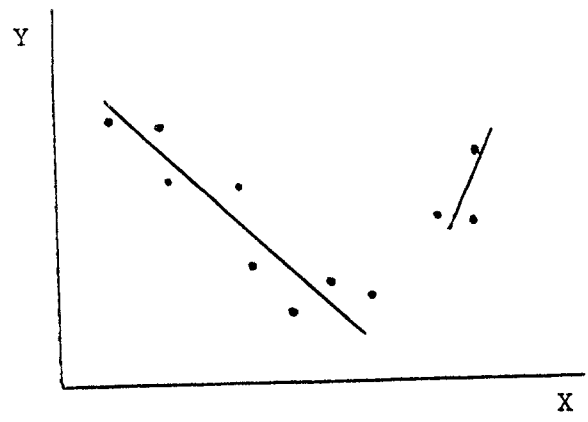
THIRD REGRESSION:
SMALLEST 5, LARGEST 6



FOURTH REGRESSION:
SMALLEST 6, LARGEST 5



FIFTH REGRESSION:
SMALLEST 7, LARGEST 4



SIXTH REGRESSION:
SMALLEST 8, LARGEST 3

Figure 1: All possible ordered pairs of regression equations when $n = 11$.

of observations in a subset. The number of pairs and triplets of regression equations may be quite large, but only the results of the 20 best pairs and the 20 best triplets of equations are reported. The second through the twentieth models are reported to provide information about the stability of the results. If the results are not stable, the statistically-defined switching point may have no meaning.

A continuity correction is available as an option. When that option is exercised the second regression equation in the two- and three-equation models is forced through the coordinates of the fitted value for the largest X in the first subset, and the third equation of the three-equation model is forced through the coordinates of the fitted value for the largest X in the second subset.

Testing the Significance of the Results

An F-test is appropriate to test the null hypotheses that the best two-equation model does not represent any improvement over the standard single-equation model and the best three-equation model represents no improvement over the two-equation model. Computation of the F-statistics is illustrated in table 1. The critical F values that appear in a standard F table, however, are inappropriate comparisons for such tests. Instead, the significance level of the standard F values must be adjusted for the number of regression equations that are computed. The printout shows the F-statistics for tests of both null hypotheses. The computed F-statistics and their significance levels, adjusted for the number of pairs and triplets of equations computed, are also printed out. For 30 observations and a minimum of 3 observations to a subset, the computed F-statistic must be at least 7.85 for the two-equation model (24 regressions) to be statistically

Table 1: Computing F values to test the significance of the two- and three-equation models

Test 1: the one-equation model vs. the two-equation model

$$F = \frac{(\text{SSR}_{\text{one}} - \text{SSR}_{\text{two}})/2}{\text{SSR}_{\text{two}} / n-4}$$

Test 2: the two-equation model vs. the three-equation model

$$F = \frac{(\text{SSR}_{\text{two}} - \text{SSR}_{\text{three}})/2}{\text{SSR}_{\text{three}} / n-6}$$

n = total number of observations

SSR_{one} = sum of squared residuals from the one-equation model

SSR_{two} = sum of the sum of squared residuals from the best two-equation model

SSR_{three} = sum of the sum of squared residuals from the best three-equation model

significant, and it must be at least 12.43 for the three-equation model (253 regressions) to be statistically significant at the .05 level.

Necessary Computer Control Cards

The SEGREG computer program has been written for the CDC Cyber 74 at the University of Minnesota. Before the program can be executed it is necessary to have a data set stored in a sep rate file. Each line of the data set must contain only one XY pair of data with the X variable preceding the Y variable. The format of the data is specified by the user. All cards except the data cards must begin in column 1, and the punctuation and spacing on the cards must be correct. Capital letters will be used to indicate what must be punched as specified; lower case indicates user-supplied arguments.

CARD 1: your name,T5. (T5 is sufficient for $n \leq 50$)

CARD 2: ACCOUNT,xxxxxxxx,password.

CARD 3: RFL,70000.

CARD 4: GET,SEGREG,TAPE1=your file.

CARD 5: MNF(Z,E=4,R=0,L=0,I=SEGREG)

CARD 6: REWIND,TAPE2.

CARD 7: COPYSBF,TAPE2.

CARD 8: 7-8-9 (all punched in column 1).

CARD 9: title of this run (leave blank if no title will be used).

CARD 10: cols 1-10: continuity correction. Punch CONTINUITY in this field if a continuity correction is desired.

cols 11-15: minimum number of observations in a subset. This number must be right justified in the field and may not contain a decimal point. The default value of 3 is used if this field is left blank.

CARD 11: (Data format: value Fortran format items accepted by SEGREG are: F, E, X, T, and /. The format must begin with a "(" and end with a ")" and must be contained on a single card. The X variable must be entered before the Y variable).

CARDS 12 and 13: 6-7-8-9 (all punched in column 1.) Prepunched orange cards are available at all UCC installations.

Instructions for preparing the first two of these cards may be obtained from the University Computing Center.

Cautions, Special Notes, etc.

The procedure is extremely sensitive to outliers. If an outlier is dominating the results, typically it will be in a subset of the data containing only the minimum number of observations, and the switching points for the next best 5 or 6 models will be approximately the same as those for the best model.

The program will not perform any transformations of the data. If any transformations of the data are desired, they must be made before the program is used.

All 80 columns of the title card are available. When doing several runs using the same data set it is a good idea to make the identification as clear as possible.

If a data set is too small for 20 regressions to be computed $\{(n-2m-1) < 20\}$ the results of all equations will be printed out in the order of their computation. Results will not be sorted unless $(n-2m-1) \geq 20$. A data set may not contain more than 300 observations.

An Example

This example using data with known properties is included to give the user a chance to test the program on her/his own installation, and to provide a guide to analyzing the output provided by SEGREG. The X values in the test data are the numbers 1 to 30. The first 20 Y values were generated by $Y_i = 25 + e_i$ and the last 10 by $Y_i = 64 - 2X_i + e_i$ where e_i is a randomly distributed random variable with a mean of zero and a standard deviation of 1. All of the data are printed in table 2. The computer printout for the single-equation model is in table 3. It is statistically significant.

The computer printout for the two-equation model is printed in table 4. The null hypothesis that the best two-equation model provides no more information than the single-equation model can be rejected at the 2.63×10^{-15} level of significance. The results of the two best models are virtually identical, indicating the stability of this switching point. A further indication of stability is that the 18 alternative switching points identified are tightly grouped around the best switching point.

The printout for the three-equation model is in table 5. The F-statistic to test the null hypothesis that the three-equation model provides no more information than the two-equation model is 4.71, a value which is not statistically significant when adjusted for 253 regressions. Fifteen of the 20 best models have 10 or 11 observations in the third subset. Frequently the switching point defining the first or the third subset will be identical to the switching point defining the second subset in the two-equation model, particularly when the two-equation model is statistically significant. In such cases it may be useful to eliminate the observations

Table 2: Data for regression

<u>Values of X_i</u>	<u>Values of Y_i</u>
1	23.671529988214
2	25.322249489482
3	24.659072392603
4	23.841611899718
5	23.249586046996
6	26.257482774987
7	24.987434469435
8	25.471760428436
9	25.673199955196
10	25.533911568276
11	25.996724843737
12	24.028007968068
13	24.797920759614
14	23.294586285301
15	24.292665615533
16	24.896533394863
17	23.622789259551
18	26.840264205438
19	24.887054300500
20	26.747499146303
21	23.086481721008
22	20.868022183915
23	19.768933387495
24	17.050362727659
25	14.259426066515
26	13.763907284128
27	11.403447321715
28	8.702358284691
29	9.332948471675
30	4.920778658854

Table 3: One phase regression on 30 observations

Total SSR	:Observations:	R Square	:	Beta	:	SD(Beta)	:	Alpha	:	SD(Alpha)	:	SSR(I)
.45398E+03	30.	.57977		-.52790		.08494		29.55681		1.50786		453.98219

Table 4: Two phase regression on 30 observations

TOTAL SSR	OBSERVATIONS	R SQUARE	BETA	SD(BETA)	ALPHA	SD(ALPHA)	SSR(I)
.25961E+02	19. 11.	.01424 .98156	.02174 -2.00602	.04387 .09164	24.58917 65.59632	2.38805 7.62842	18.64712 8.31431
.27009E+02	20. 10.	.06354 .98046	.04636 -1.92125	.04195 .09489	24.41681 63.30750	2.45357 7.90480	21.06612 5.94288
.29768E+02	17. 13.	.00752 .97556	-.01647 -1.84744	.04851 .08761	24.83037 61.38724	2.25684 7.18968	14.40043 15.36752
.31351E+02	21. 9.	.00812 .97359	.01644 -1.91082	.04169 .11871	24.63623 63.02233	2.56392 9.00229	25.43196 5.91893
.31693E+02	18. 12.	.01475 .97548	.02400 -1.90622	.04902 .09557	24.57412 62.93489	2.39666 7.64902	18.63066 13.06228
.45486E+02	22. 8.	.02312 .96313	-.03253 -1.92328	.04728 .15298	25.01167 63.36728	2.85584 10.40337	39.58904 5.89717
.56699E+02	16. 14.	.00115 .93271	.00673 -1.66579	.05287 .12895	24.69122 56.66423	2.22327 8.56206	13.30320 45.39572
.61697E+02	23. 7.	.10600 .94777	-.08135 -1.83220	.05156 .19275	25.40250 60.81860	3.11339 11.88627	56.49616 5.20119
.70527E+02	15. 15.	.06053 .91053	.00394 -1.53791	.06043 .13387	24.70704 53.38207	2.23500 8.56260	13.29084 65.23659
.96149E+02	24. 6.	.21073 .91659	-.14537 -1.79100	.05998 .27050	25.93574 59.65137	3.49949 14.33025	91.02728 5.12199
.10622E+03	14. 10.	.00654 .87806	.01954 -1.40510	.06899 .13995	24.62380 50.01756	2.23653 8.59192	12.99532 93.28279
.14176E+03	13. 17.	.11173 .83270	.08119 -1.26684	.06902 .14699	24.31557 46.56094	2.08525 8.64112	9.53711 132.22191
.15028E+03	25. 5.	.30049 .90021	-.21961 -1.97568	.06987 .37977	26.57910 64.94361	3.92943 17.27712	145.94985 4.32675
.16360E+03	12. 18.	.15328 .81268	.10724 -1.17390	.07970 .14118	24.19401 44.26992	2.07780 8.31029	9.08427 154.51283
.19110E+03	11. 19.	.38577 .78128	.18646 -1.07679	.07844 .13818	23.85061 41.90534	1.89950 8.06734	6.09083 185.01339
.19334E+03	26. 4.	.37803 .80684	-.28034 -1.88174	.07340 .65104	27.12567 62.21952	4.18399 23.01347	189.10091 4.23851
.20454E+03	10. 20.	.30535 .77666	.18016 -1.01873	.09596 .12876	23.87589 40.51198	1.92211 7.64160	6.07765 198.46203
.22339E+03	9. 21.	.26811 .76423	.19213 -.95651	.11998 .12188	23.83201 39.03949	1.94921 7.29540	6.04614 217.33992
.24215E+03	8. 22.	.18443 .75237	.18013 -.90010	.15464 .11547	23.87202 37.72313	1.98576 6.96812	6.02598 236.12836
.25049E+03	27. 3.	.44194 .62705	-.34504 -1.89079	.07755 1.45572	27.72960 62.48493	4.46136 35.00329	246.25388 4.23824

F VALUE (BEST 2 PHASE VS SINGLE PHASE) = 205.89675 SIGNIFICANCE LEVEL (ADJUSTED FOR NO. OF REGRESSIONS) = .00000

Table 5: Three phase regression on 30 observations

TOTAL SSR	OBSERVATIONS	R SQUARE	BETA	SD(BETA)	ALPHA	SD(ALPHA)	SSR(I)
.19363E+02	11.	.38577	.18648	.07844	23.85061	1.89950	6.09083
	9.	.41235	.29278	.13211	20.13854	5.89085	7.38978
	10.	.98036	-1.92125	.09489	63.30750	7.90480	5.94288
.20724E+02	13.	.11173	.08119	.06902	24.31557	2.08525	9.53711
	7.	.55351	.48183	.19353	16.74908	7.53029	5.24373
	10.	.98086	-1.92125	.09489	63.30750	7.90480	5.94288
.20823E+02	11.	.38577	.18648	.07844	23.85061	1.89950	6.09083
	8.	.22739	.21207	.15958	21.29546	6.25922	6.41769
	11.	.98136	-2.00602	.09164	65.59632	7.62842	8.31431
.22130E+02	12.	.15328	.10724	.07970	24.19401	2.07780	9.08427
	8.	.39615	.33303	.16788	19.42665	6.82548	7.10245
	10.	.98036	-1.92125	.09489	63.30750	7.90480	5.94288
.22202E+02	5.	.19520	-.23245	.27250	24.84617	1.73134	2.22776
	12.	.52635	-.18924	.05676	27.08070	2.86052	4.60666
	13.	.97536	-1.84744	.08761	61.38724	7.18968	15.36732
.22554E+02	5.	.19520	-.23245	.27250	24.84617	1.73134	2.22776
	14.	.09037	-.07265	.06633	25.94955	3.38266	12.01208
	11.	.98136	-2.00602	.09164	65.59632	7.62842	8.31431
.22767E+02	17.	.00702	-.01647	.04851	24.83037	2.25684	14.40043
	3.	.00177	-.04633	1.10091	27.03954	19.95398	2.42399
	10.	.98036	-1.92125	.09489	63.30750	7.90480	5.94288
.22851E+02	13.	.11173	.08119	.06902	24.31557	2.08525	9.53711
	8.	.36934	.40947	.26724	17.88275	8.57536	4.99937
	11.	.98136	-2.00602	.09164	65.59632	7.62842	8.31431
.23472E+02	11.	.38577	.18648	.07844	23.85061	1.89950	6.09083
	6.	.00379	-.02092	.16963	24.45875	6.01327	2.01420
	13.	.97536	-1.84744	.08761	61.38724	7.18968	15.36732
.23742E+02	10.	.00115	.00673	.05287	24.69122	2.22327	13.30320
	4.	.37930	.74203	.67054	11.79570	15.17669	4.49630
	10.	.98036	-1.92125	.09489	63.30750	7.90480	5.94288
.23756E+02	12.	.15328	.10724	.07970	24.19401	2.07780	9.08427
	7.	.20035	.23883	.21309	20.83947	7.44341	6.35725
	11.	.98136	-2.00602	.09164	65.59632	7.62842	8.31431
.23904E+02	10.	.30535	.18010	.09596	23.87589	1.92211	6.07765
	10.	.13341	.14891	.13418	22.63227	5.77447	11.08338
	10.	.98036	-1.92125	.09489	63.30750	7.90480	5.94288
.24025E+02	5.	.19520	-.23245	.27250	24.84617	1.73134	2.22776
	15.	.00471	-.01637	.06600	25.36806	3.51929	15.84423
	10.	.98036	-1.92125	.09489	63.30750	7.90480	5.94288
.24107E+02	14.	.00634	.01954	.06899	24.62380	2.23655	12.99532
	6.	.39730	.44181	.27174	17.48286	9.16585	5.16897
	10.	.98036	-1.92125	.09489	63.30750	7.90480	5.94288
.24300E+02	10.	.30535	.18010	.09596	23.87589	1.92211	6.07765
	9.	.01745	.05410	.15359	23.92717	5.96513	9.90835
	11.	.98136	-2.00602	.09164	65.59632	7.62842	8.31431
.24333E+02	15.	.00033	.00394	.06043	24.70704	2.23500	13.29084
	5.	.32601	.49662	.41226	16.45967	11.59303	5.09887
	10.	.98036	-1.92125	.09489	63.30750	7.90480	5.94288
.24728E+02	19.	.01424	.02174	.04387	24.58917	2.38805	18.64712
	0.	.98131	-2.33273	.15831	72.78467	8.97795	1.75423
	5.	.90031	-1.97560	.37977	64.94361	17.27712	4.32675
.24891E+02	19.	.01424	.02174	.04387	24.58917	2.38805	18.64712
	3.	.98033	-2.93974	.41643	85.30184	13.56184	.34683
	8.	.96343	-1.92328	.15298	63.36728	10.40337	5.89717
.25011E+02	9.	.26811	.19213	.11998	23.83201	1.94921	6.04614
	10.	.00012	-.00393	.12703	24.87609	5.26849	10.65091
	11.	.98136	-2.00602	.09164	65.59632	7.62842	8.31431
.25129E+02	8.	.18443	.18013	.15464	23.87202	1.98576	6.02598
	9.	.47533	-.23759	.09431	27.77047	4.07021	3.75542
	13.	.97536	-1.84744	.08761	61.38724	7.18968	15.36732

F VALUE (BEST 3 PHASE VS BEST 2 PHASE) = 4.70862 SIGNIFICANCE LEVEL (ADJUSTED FOR NO. OF REGRESSIONS) = .99999

known to constitute a subset--in this case the observations containing the largest 10 X values--and rerun SEGREG using the reduced data set. Then the results of the two-equation model can be used to test the hypothesis that there are two additional subsets to the data, given that the existence of one subset is known.