MAC IIci-Word 5--education example

	(1) SAT	(2) gradrate	(3) teachsalary	(4) pupil.teach.ratio	(5) expendpupil
AL	970	64.4	17948	20.3	2177
AK	914	77.8	34510	13.2	7325
AZ	978	68.4	21119	19.5	2524
AR	1003	76.2	15310	18.2	1971
CA	897	75.1	23614	23.3	2733
CO	979	79.2	23276	18.6	3171
CT	904	77.9	21036	14.8	3636
DE	902	88.9	20625	17.5	3456
DC	823	58.4	25610	17.2	4260
FL	890	65.5	18275	17.8	2680
GA	822	65.9	13040	18.6	2169
HI	869	82.2	24319	22.9	3239
ID	992	77.9	17605	20.7	2052
IL	981	77.1	22972	18.0	3100
IN	864	78.3	20347	19.8	2414
IA	1089	88.0	19402	15.7	3095
KS	1051	82.5	18313	15.6	3058
KY	997	68.4	18384	20.2	2100
LA	980	57.2	18416	18.4	2739
ME	892	76.7	16248	19.5	2458
MD	897	81.4	22800	18.3	3445
MA	896	77.5	21841	16.1	3378
MI	976	73.4	25712	21.9	3307
MN	1020	90.7	22876	18.0	3085
MS	992	63.7	14320	18.6	1849
MO	981	76.2	17521	17.4	2468
MT	1034	83.1	19702	16.0	3289
NE	1041	84.1	17399	15.5	2984
NV	931	74.6	22067	20.9	2613
NH	931	76.5	16549	16.4	2750
NJ	876	82.7	21536	15.8	4007
NM	1014	71.4	20187	18.8	2901
NY	894	66.7	25000	18.8	4686
NC	827	69.3	17585	19.8	2162
ND	1054	94.8	18774	16.6	2853
OH	968	82.2	20004	19.8	2676
OK	1009	79.6	18270	17.0	2805
OR	907	73.0	21746	18.6	3504
PA	887	79.7	21178	17.2	3329
RI	885	75.2	23175	15.7	3570
SC	803	66.2	16523	18.9	2017
SD	1086	85.0	15592	15.5	2486
TN	1009	65.1	17698	20.9	2027
TX	886	69.4	19550	17.9	2731
UT	1045	84.5	19859	24.3	2013
VT	907	85.0	16299	13.9	3051

VA	894	75.7	18535	17.4	2620
WA	968	75.5	23485	21.7	3211
WV	976	77.4	17322	16.9	2764
WI	1007	84.0	21496	17.4	3237
WY	1034	81.7	23822	15.2	4045

Correlation values among the five variables:
1: SAT average, 2:high school graduation rate, 3: average teacher salary
4: pupil teacher ratio in classroom, 5: average expenditures per pupil
table entry the value of the CC, in (---) the estimate of the population correlation, r for the NPCC if the data were normally distributed. Conversion formula is below the table.

CORRELATIONS IN THE EDUCATION DATA

variable	CC	2	3	4	5
	Pearson	.4104**	-.1465	-.0878	-.1577
1	Spearman	.3980**	-.1711	-.1672	-.1354
		(.4137)	(-.1789)	(-.1748)	(-.1416)
	Kendall	.2792**	-.1122	-.1208	-.0917
		(.4246)	(-.1753)	(-.1886)	(-.1435)
	GD	.1800	-.2400*	-.0400	-.0800
		(.2789)	(-.3681)	(-.0628)	(-.1253)
	Pearson		.0970	-.2804*	.1614
2	Spearman		.0824	-.4356**	.3056*
			(.0863)	(-.4522)	(.3186)
	Kendall		.0486	-.2752**	.2102*
			(.0763)	(-.4189)	(.3242)
	GD		.1400	-.3400**	.3200**
			(.2181)	(-.5090)	(.4817)
	Pearson			-.0126	.8273**
3	Spearman			.0891	.7336**
				(.0932)	(.7494)
	Kendall			.0627	.5372**
				(.0983)	(.7472)
	GD			.0400	.5600**
				(.0627)	(.7705)
	Pearson				-.4778**
4	Spearman				-.4850**
					(-.5025)
	Kendall				-.3388**
					(-.5074)
	GD				-.3600**
					(-.5358)

Spearman: $\hat{r} = 2\sin(\frac{p r_s}{6})$; Kendall ($r = r_k$) and GD ($r = r_{gd}$): $\hat{r} = \sin(\frac{p r}{2})$

Correlation Coefficient	Two-sided critical values (n= 51) 5% = one *	1% = two **'s
Pearson	.279	.361
Spearman	.277	.364

Kendall	.188	.247
GD	7/25: 6/25 (.58)	9/25: 8/25 (.76)

x = ERA,   y = fraction of games won

	x	y
Boston	3.98	0.549
Detroit	3.72	0.543
Milwaukee	3.45	0.537
Toronto	3.81	0.537
New_York_Y	4.23	0.528
Cleveland	4.16	0.481
Baltimore	4.54	0.335
Oakland	3.43	0.642
Minnesota	3.93	0.562
Kansas_City	3.66	0.522
California	4.32	0.463
Chicago_WS	4.13	0.441
Texas	4.07	0.435
Seattle	4.20	0.422
New_York_M	2.91	0.625
Pittsburgh	3.47	0.531
Montreal	3.10	0.500
Chicago_C	3.88	0.475
St_Louis	3.49	0.469
Philadelphia	4.16	0.404
Los_Angeles	2.96	0.584
Cincinnati	3.35	0.540
San_Diego	3.28	0.516
San_Francisco	3.42	0.512
Houston	3.40	0.506
Atlanta	4.11	0.338

Chapter 4: Examples

This chapter continues the baseball examples by giving the CC estimates of the standard error about the regression line, and an example is added which illustrates why more than the Pearson and Spearman CC's should be run on bivariate data. The data comes from the United States Department of Education and appeared in the International Edition of USA Today on December 20, 1984. The data is educational data from 50 states and Washington D.C. State averages on five variables were recorded related to high school education; (1) SAT scores, (2) Graduation rate, (3) Teacher salary, (4) Pupil-teacher ratio, and (5) Expenditures per pupil. In such data, researchers may be interested in seeing if there are relationships among the states on these variables. This data cannot be considered independent and identically distributed because each state is somewhat unique. Because of this a researcher should want to treat each state with equal importance in searching for relationships. The Greatest Deviation CC does treat data points on an equal basis and it will be seen that GDCC does show some relationships that are missed by Pearson and Spearman and even Kendall. The data and the correlations appears at the end of this chapter and should be read prior the discussion.

The Pearson, Spearman, Kendall, and GD CC's were computed on the ten pairs of data and noted whether or not they were signicant at the 1(**) or 5 (*) % levels. Exact critical values were used for Pearson and GD and asymptotic values for Spearman and Kendall. In order for GD to have an exact 5% critical point, one rejects if GD equals or exceeds 7/25 and randomly rejects 58% of the time if it equals 6/25. Since 58% is greater than 1/2, GD values get a * if 6/25 is obtained. A similar remark goes for GD and its 1% critical value.

There was agreement on six of the relationships; variable pairs (1,4), (1,5), (2,3), (3,4) were nonsignificant whereas (3,5) and (4,5) were significiant at the 1% level. On pair (2,3) all the NPCC's were more significant than Pearson. On pair (2,4) GD was more significant than the other three CC's. However, a remarkable difference occurs for pairs (1,2) and (1,3). On (1,2) GD is not significance whereas the other three CC's are, but for pair (1,3) GD is significiant at the 5% level while none of the others are. Which is right? One can study the bivariate plots and use influence measures to delete data of an unusual nature relative to the rest of the states. If this is done, then the data for GA, MN, and WY are deleted for pair (1,3) and GD becomes more significant and all the other CC's become significant. Thus, three states masked a negative CC except for GD which possibly avoided a Type II error. For pair (1,2) delete DC, IA, MN, and SD and GD becomes even closer to zero and all three of the other CC's become nonsignificant. In this case, three correlations are being made significant by just four areas and only GD gave a result consistent with most of the data and possibly avoided a Type I error. Note that different states were deleted for these two pairs, and hence, it is unclear what conclusion should be draw for all the data with three of the CC's. However, the CC GD made the correlational analysis easy and consistent conclusions could be drawn.

With all the multivariate data being analyzed in complex problems by many of today's researchers, this example makes it clear how a small segment of the data can lead one to dubious conclusions. Since Least Squares estimation techniques are closely related to the Pearsons CC as shown in early chapters for regression, it is clear that without a parallel robust NPCC analysis, many conclusions could be draw which do not represent the majority of the data.

Thus, in our education example only GD pointed to a possible relationship between teacher salary and SAT scores, and it was negative. The GDCC regression gave

$$S\hat{A}T = 1130.84 - 0.008939 * teachersalary$$

whereas Pearson's CC (slope and intercept same as least Squares) gave

$$S\hat{A}T = 1008.27 - 0.002906 * teachersalary.$$

Note that in the 5% significant regression an increase of $1000 in average teacher salary points to a decrease of 8.9 in average SAT score, but that Pearson's regression is nonsignificant and the decrease is only 2.9. The contradiction in that higher salaries lead to lower Sat scores lends itself to interesting speculation, but one thought might be that such data involving state averages cannot be used to draw any meaningful conclusions about high school education.

The baseball examples include average major league team statistics for the 1989 season, x was the team pitching earned run average (ERA) per game, and y was the final fraction of games won (winpct). The data is included in this chapter. By using the CC location and scale estimation techniques of the last chapter on the regression of y on x, an estimate of the residual standard error of the regression is obtained. For GDCC the regression was $\hat{y} = 0.8092 - 0.08353x$. Let the vector of residuals be $res = y - \hat{y}$ and $reso$ the vector of ordered residuals, with q the vector of normal quantiles with ith component $q_i = \Phi^{-1}(\frac{i}{27}), i = 1, 2, \cdots, 26$. The regression of $reso$ on q gives a line and Q-Q plot whose slope is the estimate of the regression standard deviation. The slope is $\hat{s} = 0.0571$. This Q-Q plot also indicates that the residuals show no deviation from normality.

For least square (Pearson CC) on this data, the classical estimate of s is $\hat{s} = 0.0553$ from the residuals on the regression line $\hat{y} = 0.9276 - 0.1145x$. The GD estimate 8.35% fewer wins with an increase of one in the ERA while LS estimates 11.45%. The ERA's in 1989 were between 4.54 (Baltimore) and 2.91 (N.Y. Mets) The residual variation estimate is essentially the same (~0.056).

The second baseball example consisted of estimating the average number of hits to produce a run in the 1992 Atlanta Braves games. The response variable y is runs in a game, and x, the predictor variable, is the number of hits. Two separate regression were run for the 175 Braves games. Braves hits and runs (x, y) and their opponents (x_1, y_1). The data is given in this chapter. The regressions were done in Chapter 2 and now the residual variation can be estimated. The reader should remember that with 175 games and discrete data, there are an extreme number of ties in the data, and so this example illustrates the versatility of the GD NPCC in simple linear regression(or any other NPCC). The GD estimate of s comes from the sope of the regression of the ordered residuals $(y - \hat{y})$ on vector q where $q_i = \Phi^{-1}(\frac{i}{176}), i = 1, 2, \cdots, 175$. A summary of the results

	Braves	Opponents
GD	$\hat{y} = 0 + 0.5000x$	$\hat{y}_1 = -1.6750 + 0.6125x_1$
Pearson, LS	$\hat{y} = -1.6155 + 0.6828x$	$\hat{y}_1 = -1.7881 + 0.6554x_1$
GD, $\hat{s}$	1.655	1.946
Pearson, LS, $\hat{s}$	1.804	1.941

The Pearson regression line is also the least squares (LS) regression line but the LS estimate of s is not the same as would be the Pearson estimate of s as it would come from the slope of the regression line of $(y-\hat{y})$ on q but using the Pearson CC to fit the line.

Apart from the discrete nature of the hit-run data, the Braves Q-Q plot from which 1.655 was obtained with the residuals reveals a lack of fit only at the upper end ($q_i > 1.4$) with the data moving upward from the line in an arc. The Opponents plot shows only two points upper end points not on the GD regression line(Q-Q plot). Note that the data is certainly not normal and so the correct method of analysis to obtain the hits to runs estimate is uncertain. How should one draw conclusions in such a problem? Also note the the GD method yields smaller residual error than does LS for the Braves and is essentially the same for their opponents. In general GD gives a smaller slope as the estimate of runs per hit than does LS. From the x-y plots of the data and their regression lines it appears that this is due to the few extreme games with more than 12 hits per game.

47 2 2 11 5

	x	y	x1	y1
1	7	2	2	0
2	8	3	5	1
3	8	4	15	11
4	8	5	9	3
5	6	0	6	3
6	12	6	8	2
7	9	4	7	5
8	7	4	6	5
9	4	1	8	3
10	10	3	2	0
11	7	5	12	7
12	8	3	10	7
13	7	2	8	4
14	13	10	13	4
15	5	2	11	4
16	7	4	15	9
17	5	2	7	4
18	5	2	4	0
19	9	3	8	2
20	7	5	2	0
21	3	1	3	0
22	12	8	7	0
23	11	7	12	8
24	9	3	9	0
25	5	0	8	7
26	12	6	5	1
27	9	3	7	4
28	12	3	16	4
29	10	4	8	2
30	9	2	7	1
31	13	11	15	12
32	11	5	12	6
33	9	3	13	8
34	10	4	6	2
35	15	10	21	11
36	10	3	7	4
37	7	4	6	2
38	6	1	11	7
39	6	4	9	5
40	12	5	6	1
41	8	2	11	7
42	10	6	5	3
43	2	1	9	7
44	13	6	14	7
45	7	2	6	1
46	6	1	6	4

94 1 1 5 0

	x	y	x1	y1
48	15	9	7	3
49	10	5	7	1
50	14	6	6	1
51	11	7	9	6
52	10	5	11	3
53	7	1	9	4
54	8	3	7	2
55	15	5	2	1
56	10	9	6	4
57	7	4	8	2
58	5	2	6	3
59	6	2	6	1
60	12	6	10	4
61	10	4	6	2
62	12	4	7	2
63	4	2	5	0
64	12	9	10	8
65	8	4	5	3
66	10	5	9	7
67	6	3	9	2
68	9	2	11	1
69	9	2	6	0
70	17	7	5	0
71	7	5	2	0
72	7	4	9	7
73	6	3	12	12
74	7	5	6	6
75	7	4	8	3
76	6	1	5	2
77	9	3	5	0
78	6	4	7	2
79	6	0	10	8
80	4	1	9	3
81	5	4	8	5
82	6	2	10	1
83	7	2	3	0
84	10	4	4	0
85	9	3	4	1
86	11	7	14	4
87	9	4	8	2
88	8	5	5	0
89	10	3	6	0
90	9	3	6	2
91	13	9	13	7
92	7	2	7	0
93	7	4	4	3

	x	y	x1	y1
95	8	4	9	5
96	8	1	12	5
97	7	5	12	7
98	9	5	9	3
99	4	0	11	5
100	10	3	6	4
101	11	5	9	3
102	6	3	5	0
103	11	8	10	5
104	7	7	11	5
105	9	5	10	1
106	10	5	6	3
107	10	6	6	2
108	18	12	9	2
109	11	10	7	3
110	6	3	15	5
111	8	4	12	8
112	6	4	9	3
113	22	15	6	0
114	9	7	12	5
115	7	2	7	4
116	12	5	5	4
117	9	5	4	1
118	7	4	7	2
119	7	2	8	3
120	7	2	10	5
121	6	3	4	2
122	11	3	13	8
123	8	0	10	6
124	7	4	13	5
125	7	3	7	7
126	13	7	7	6
127	5	2	12	10
128	15	8	9	6
129	12	7	13	5
130	9	4	8	1
131	11	5	9	6
132	7	2	14	11
133	4	1	6	2
134	9	6	7	5
135	6	4	12	3
136	16	7	5	1
137	10	7	13	5
138	11	12	11	7
139	6	3	6	2
140	13	7	4	0
141	11	9	10	3

	x	y	x1	y1
142	10	9	7	2
143	4	2	7	4
144	5	3	10	2
145	6	2	9	3
146	8	3	18	13
147	10	2	8	3
148	15	16	7	1
149	7	4	6	2
150	5	1	7	4
151	13	7	6	0
152	3	0	7	4
153	6	0	7	1
154	9	2	3	1
155	8	2	6	1
156	8	6	8	0
157	5	0	6	1
158	11	6	8	5
159	7	4	4	1
160	10	7	10	2
161	3	1	4	0
162	7	3	8	4
163	8	5	5	1
164	14	13	7	5
165	5	2	8	3
166	11	6	6	4
167	3	1	13	7
168	9	4	13	13
169	7	3	7	2
170	4	3	4	1
171	5	4	9	5
172	9	2	6	3
173	5	1	6	2
174	13	7	6	2
175	8	3	14	4