The Regression Fallacy: Or Elephants on Parade

Roger Koenker

Economics 536 UIUC

September 28, 2016

Regression to Mediocrity

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

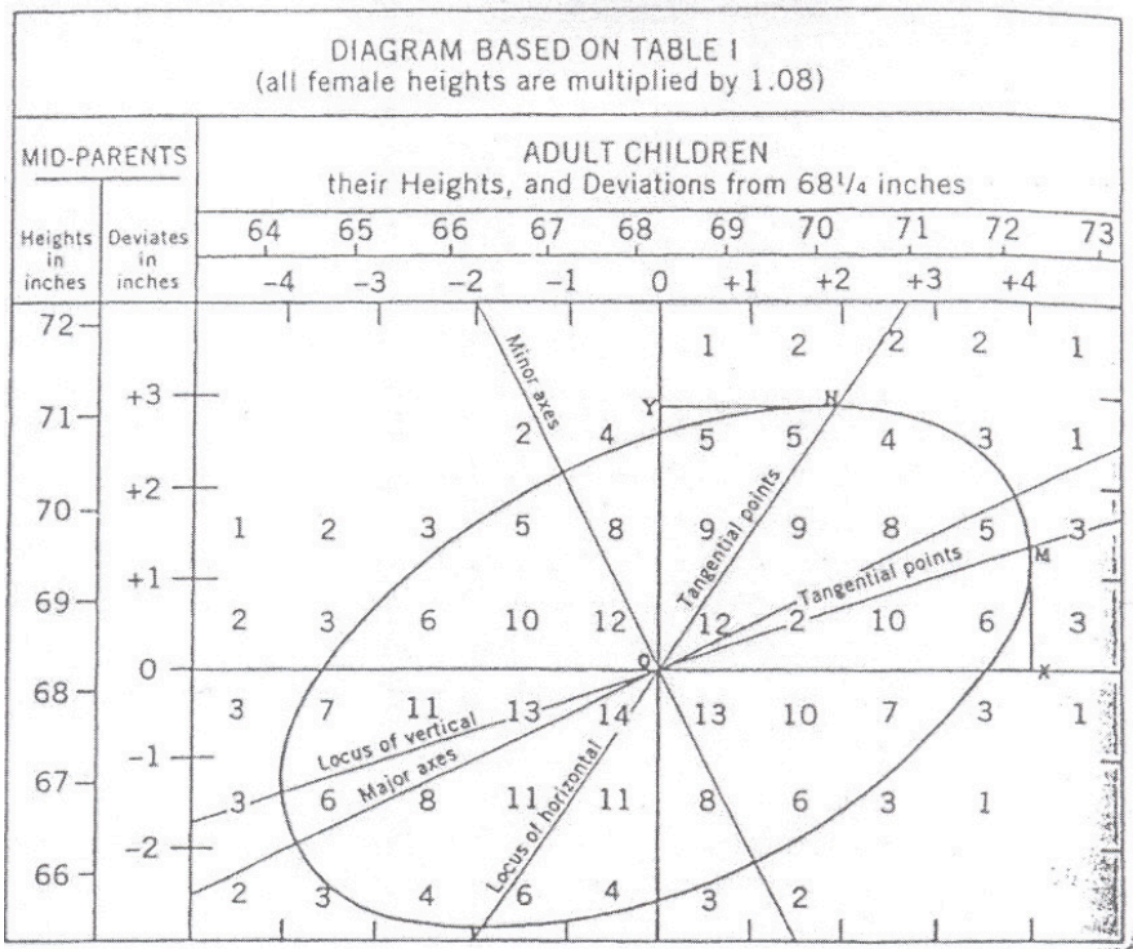
Height of the mid-parent in inches	Height of the adult child													Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7			
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1886a).

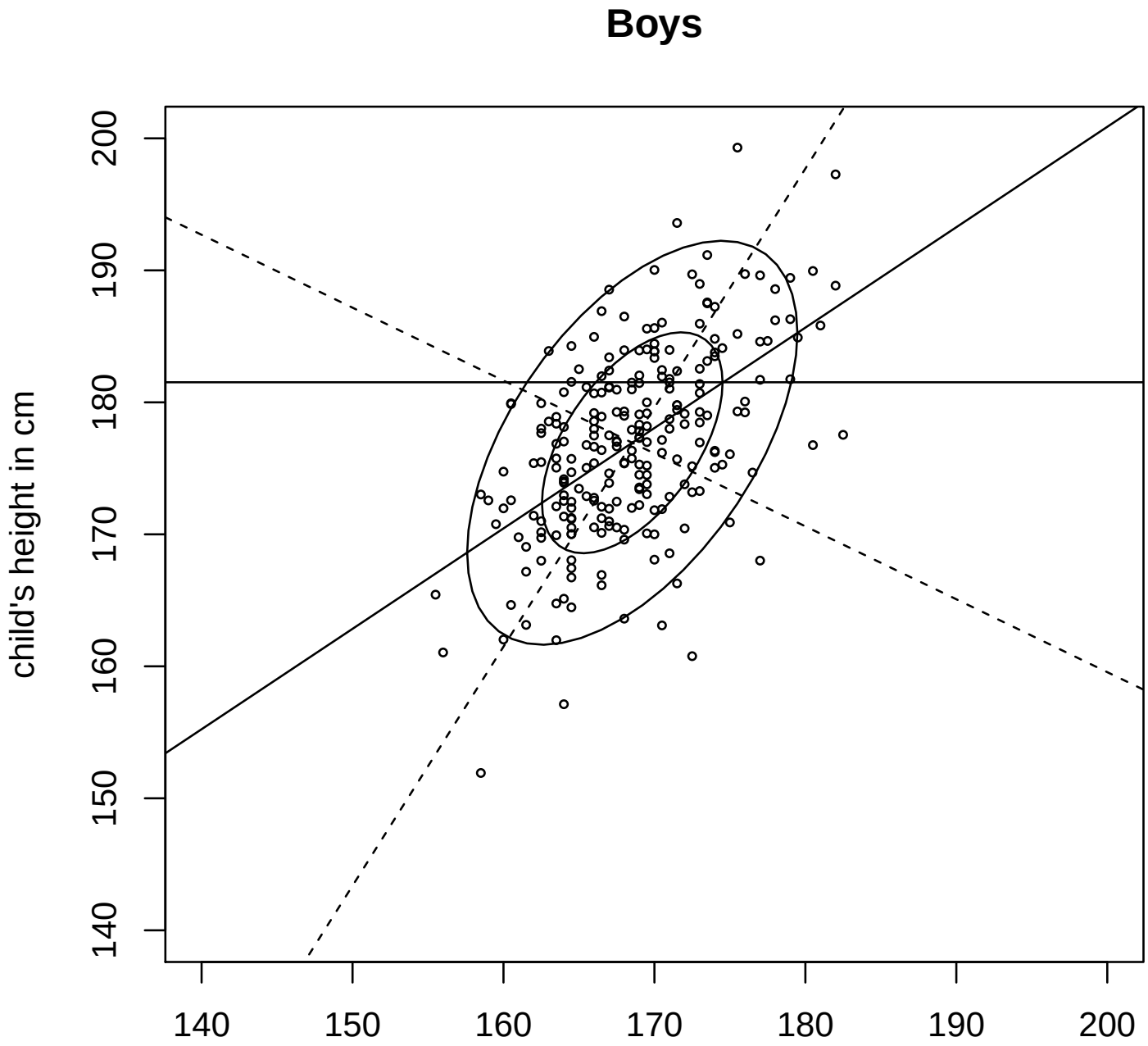
Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties); he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

Galton's (1889) Regression to the Mean Data

Regression to Mediocrity



Galton's (1889) Regression to the Mean Plot


Boys
child's height in cm
140
150
160
170
180
190
200
140
150
160
170
180
190
200

Secrist's Convergence to Mediocrity I

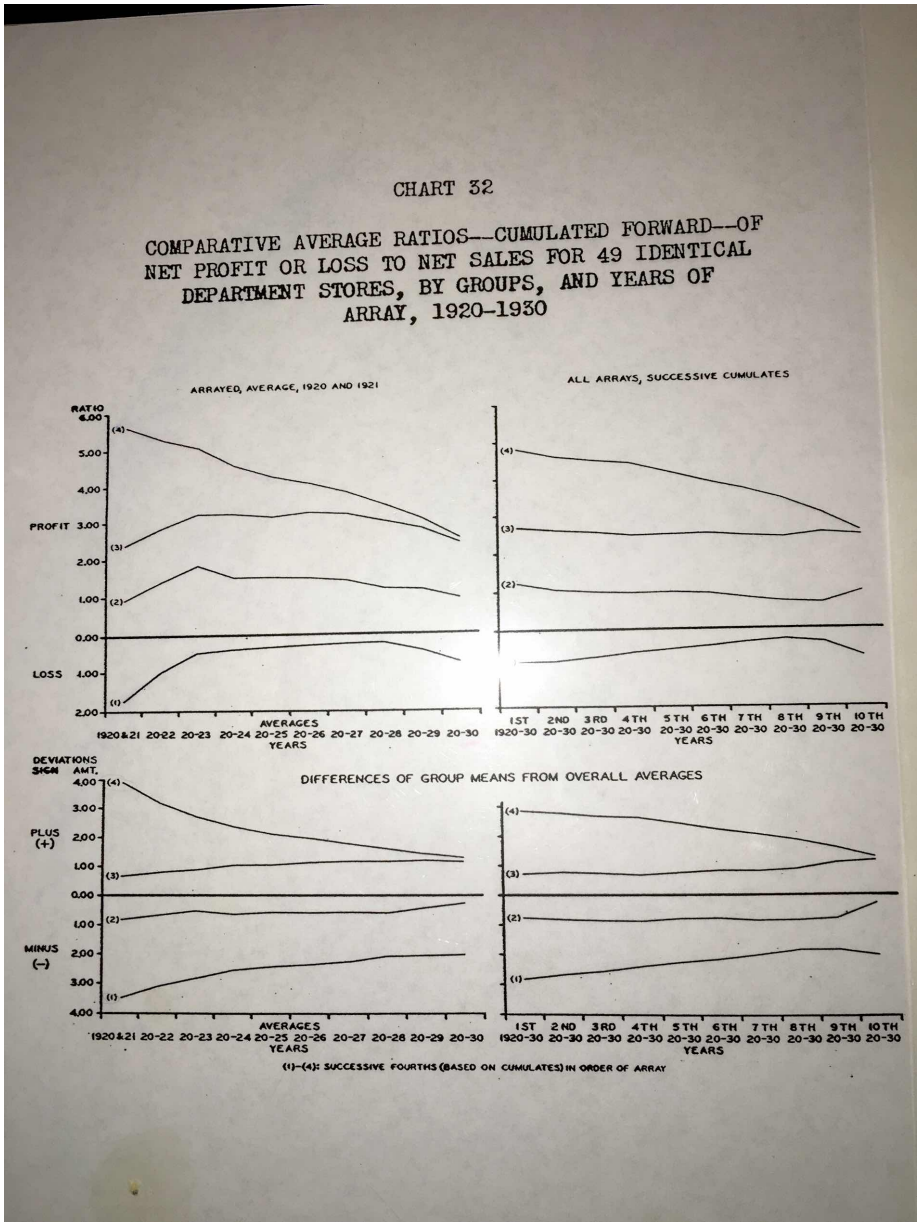
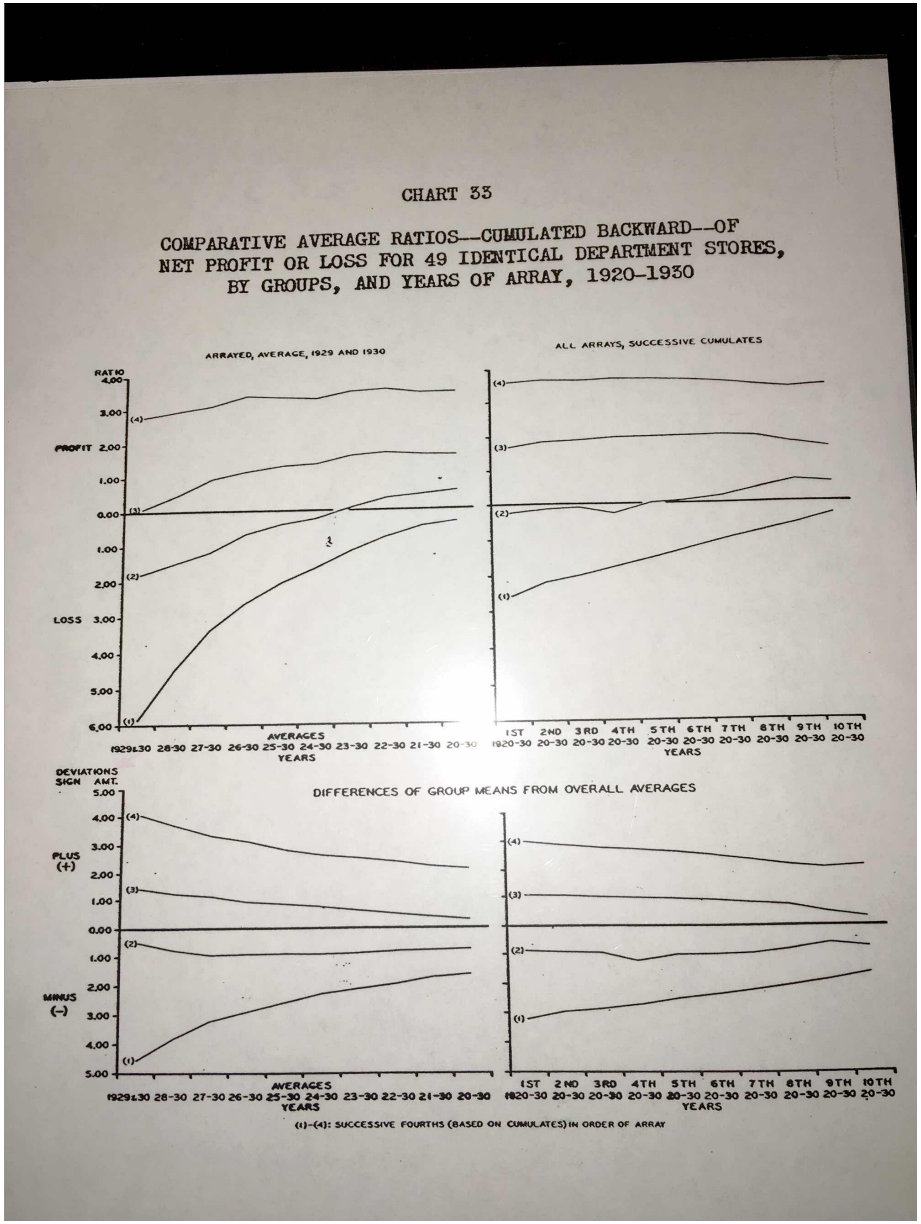


CHART 32

COMPARATIVE AVERAGE RATIOS--CUMULATED FORWARD--OF NET PROFIT OR LOSS TO NET SALES FOR 49 IDENTICAL DEPARTMENT STORES, BY GROUPS, AND YEARS OF ARRAY, 1920-1930

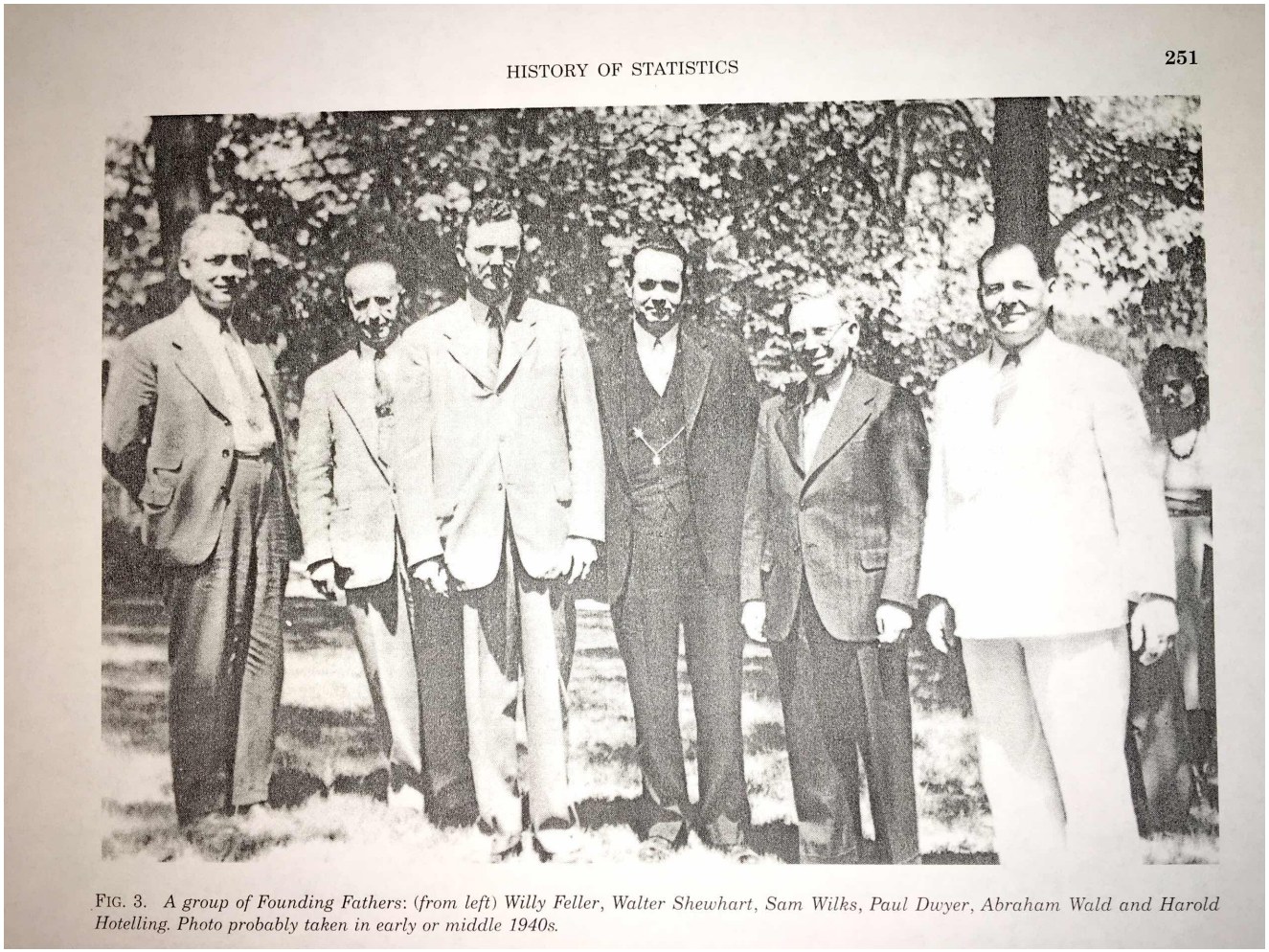
Secrist's Convergence to Mediocrity II



Harold Hotelling

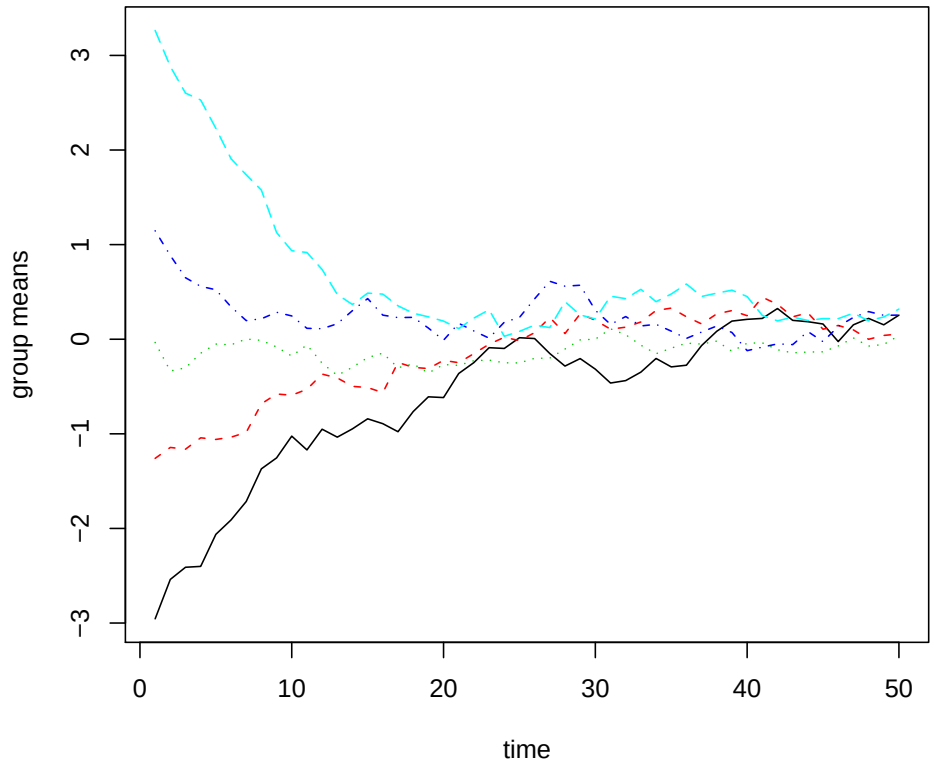
Harold Hotelling was born in Fulda Minnesota in 1895, grew up in Seattle, and was educated at the University of Washington, and Princeton University. He taught at Stanford from 1924-31, Columbia from 1931-46, and U. of North Carolina.

Founding Fathers



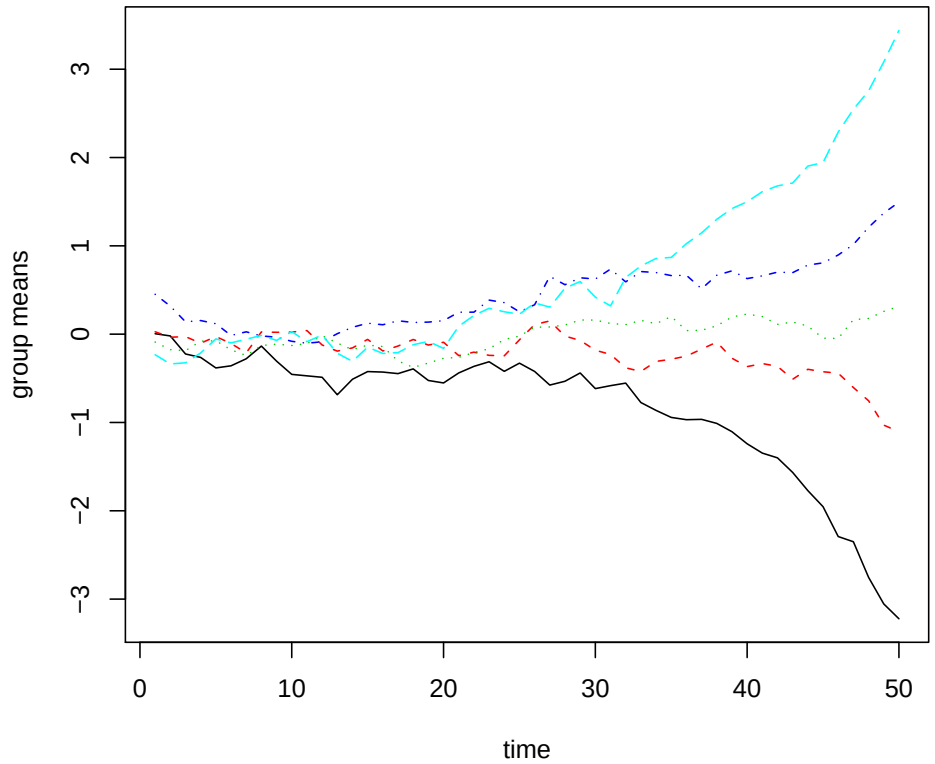
Founding fathers of probability and statistics in post-war U.S. From left to right: William Feller, Walter Shewhart, Samuel Wilks, Paul Dwyer, Abraham Wald and Harold Hotelling.

Electronic Elephants on Parade



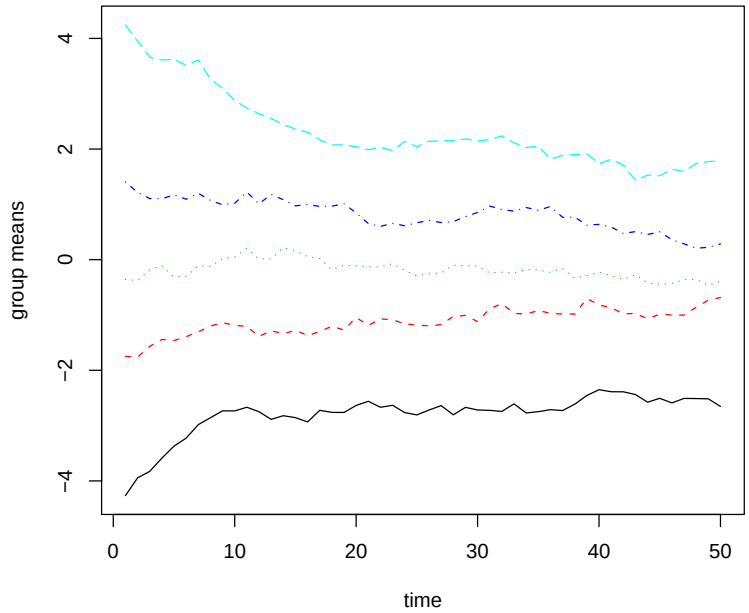
Electronic Elephants on Parade: This figure illustrates a simple AR(1) version of the Hotelling-Secrist regression fallacy. 500 AR(1) series of length $T = 100$ were generated with $\rho = .9$. At time $t = 50$ the series were grouped into quintiles and the group means of these quintiles were plotted following Secrist's approach.

Electronic Elephants on Parade II



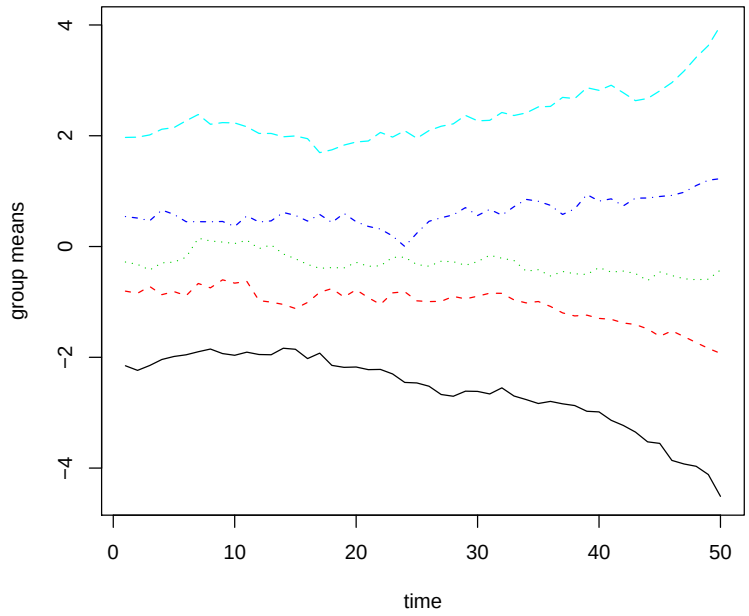
This figure illustrates a simple AR(1) version of the Hotelling-Secrist regression fallacy. 500 AR(1) series of length $T = 100$ were generated with $\rho = .9$. So this time the elephants are walking backwards.

Electronic Elephants on Parade III



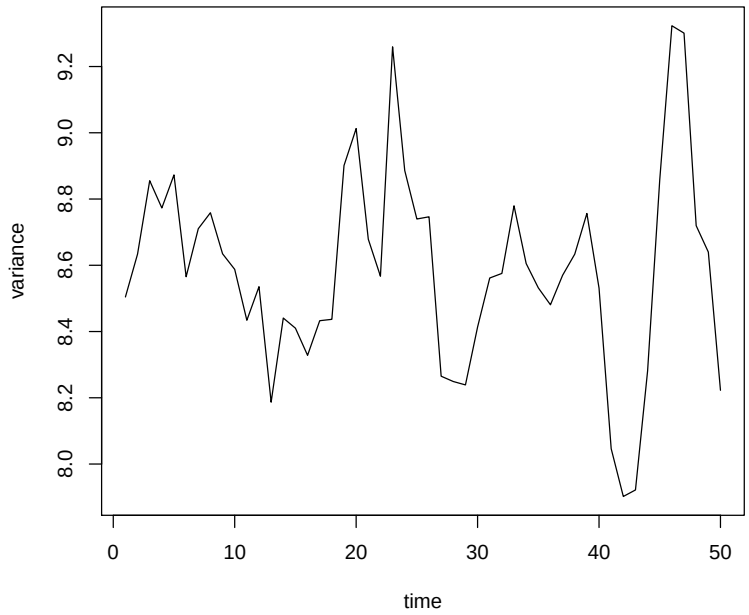
In contrast to the previous example, now each cross-sectional unit has a fixed effect which is drawn from a $\mathcal{N}(0, 2)$ distribution. As in Secrist, the group means still "converge to mediocrity". But note that now the convergence is less pronounced than in the previous case since the groups tend to different group means because of the fixed effects.

Electronic Elephants on Parade IV



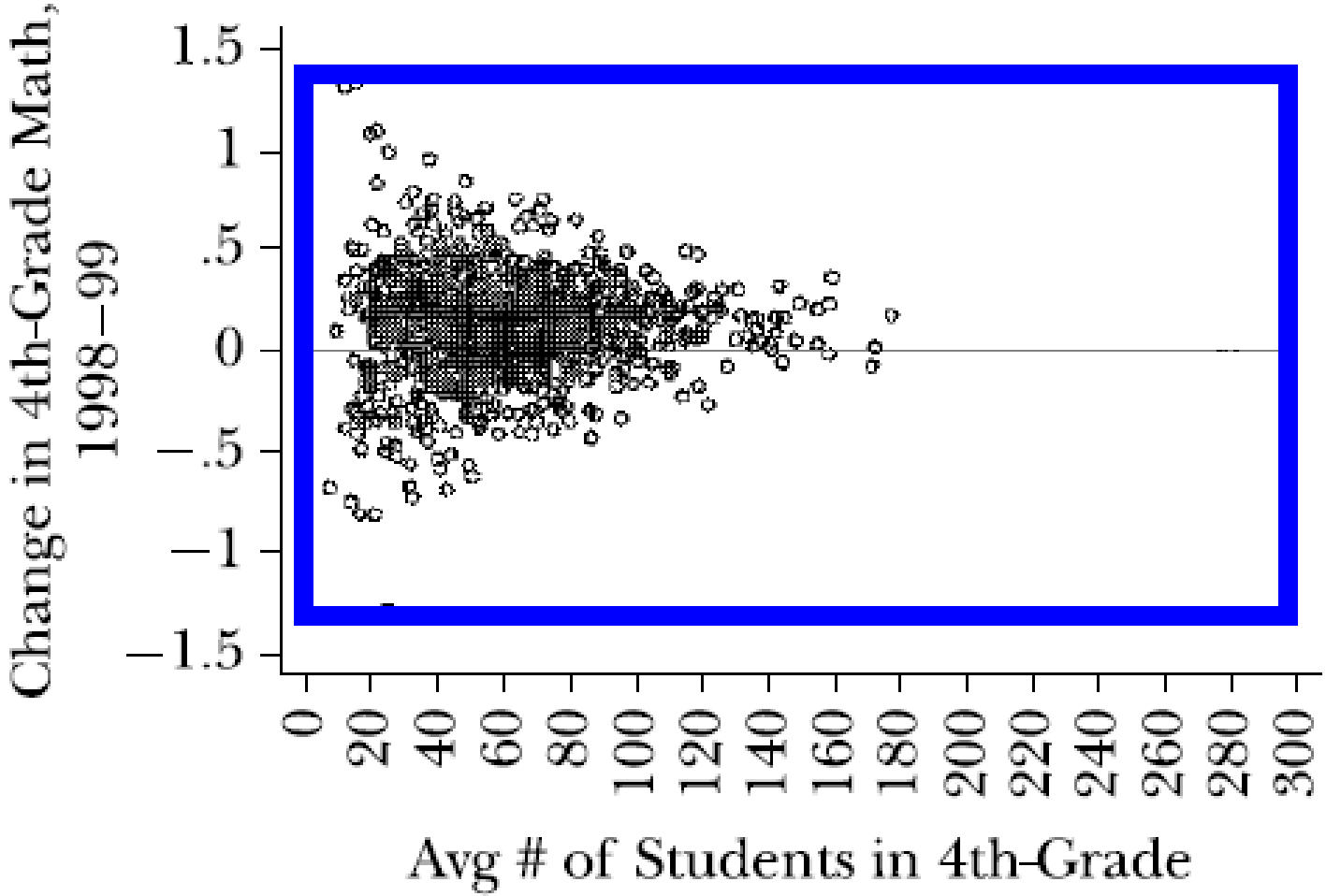
This time, again they are going backwards, there are fixed effects and at time $t = 100$ the series were grouped into quintiles and the group means of these quintiles were plotted following Hotelling's suggestion for evaluating Secrist's approach with department stores' profitability.

Electronic Elephants on Parade V



A better way to investigate whether there is really convergence to mediocrity over time would be to look at the evolution of the variance – is the variability of profitability getting smaller? Here we plot that evolution for the elephant parade and find that it fluctuates around a constant value of about 8.5. So no convergence tendency is revealed.

Trial of the Pyx for School Performance



Mean 4th Grade Math Score Changes by School Size: Source Kane and Staiger (2002)