University of Illinois
Spring 2006

Department of Economics
Roger Koenker

Economics 471

Lecture 1

Regression to Mediocrity: Galton's Study of the Inheritance of Height

Arguably, the most important statistical graphic ever produced is Francis Galton's (1885) figure illustrating "regression to the mean", reproduced badly below as Figure 1. In it Galton plots childrens' height versus parents' height for a sample of 928 children. He begins by dividing the plane into one inch squares and entering the frequency counts for each square. The resulting "histogram" appeared too rough so he smoothed the plot by averaging the counts within each group of four adjacent squares and plotting the averaged count at the intersection of the cell boundaries. Not content to invent "regression" in one plot, he managed to invent bivariate kernel density estimation, too! After smoothing, the counts appeared more regular and he enlisted the help of the Cambridge mathematician, J.H. Dickson, to draw elliptical contours corresponding to level curves of the underlying population density.

Now suppose we wished to predict children's height based on parental height, say the average height of the parents which we will call, following Galton, the height of the midparent. what would we do? One approach, given the graphical apparatus at hand would be to find the "most likely" value of the child's height given the parents' height, that is, for any given value of the mid-parent height we could ask, what value of the child's height puts us on the highest possible contour of the joint density. This obviously yields a locus of tangencies of the ellipses with horizontal lines in the figure. These conditional modes, given the joint normality implicit in the elliptical contours, are also the conditional medians and means. The slope of the line describing this locus of tangencies is roughly 2/3 so a child with midparent 3 inches taller than average can expected to be (will most probably be) only 2 inches taller than average.

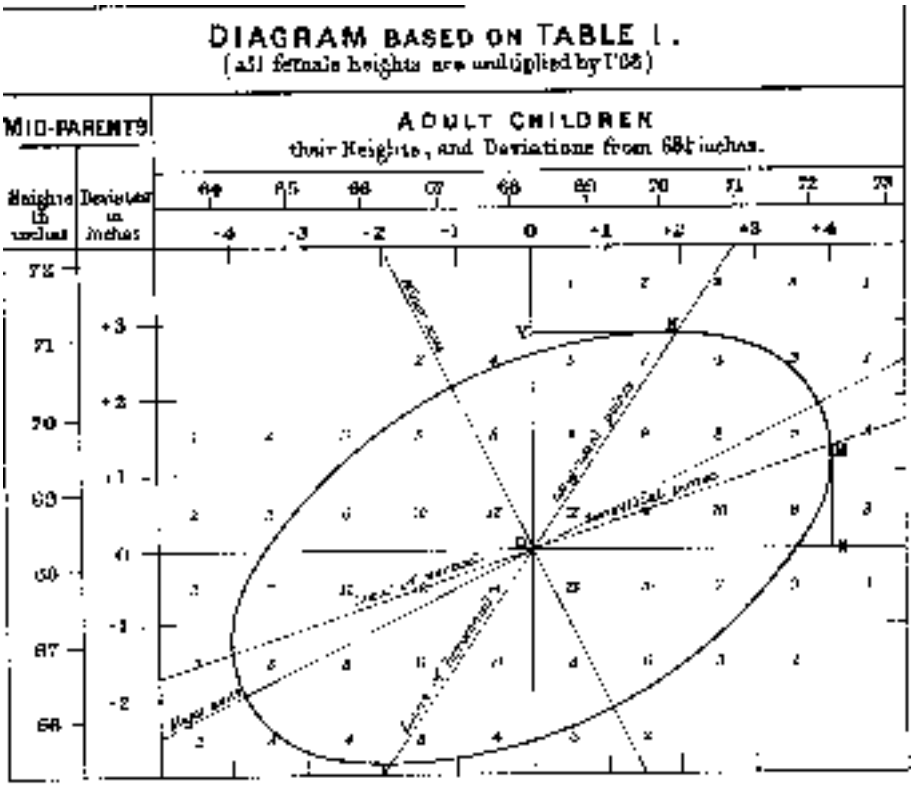

FIGURE 1. Galton's (1889) Regression to the Mean Plot

Galton termed this regression towards mediocrity, and paraphrasing Abraham Lincoln we might strengthen this to regression of the mediocre, to the mediocre, and for the mediocre. Children are more mediocre than their parents: they are tend to be "on average" closer to the mean height, the mean weight, the mean intelligence of the population than were their parents. In the case of height we have seen that children of parents who are one inch taller than the general population tend on average to be only about two-thirds of an inch taller than the population. But before you despair, I should hasten to point out, as Galton did, that parents are also more mediocre than their children – if we run the usual conventions of temporal causality backward and ask: how do the heights of parents compare to the heights of their children we find (looking at the figure) that children who are unusually tall have parents that are closer than they to the mean height of the population, and children who are unusually short also have parents

closer than they are to the mean. Stigler (1997) provides a fascinating guide to Galton's own thinking about this idea, and to the illusive nature of its reception in subsequent statistical research.

It is a remarkable feature of the conditional densities of jointly Gaussian random variables that the conditioning induces what we may call a "pure location shift". In Galton's original example the height of the midparent alters only the location of the center of the conditional density of the child's height; dispersion and shape of the conditional density is invariant to the height of the midparent. This is, of course, the essential feature of the classical linear regression model – the entire effect of the covariates on the response is captured by the location shift

$$E(Y|X = x) = x'\beta$$

while the remaining randomness of Y given X may be modeled as an additive error *independent* of X. Just to confirm that this empirical regularity hasn't been repealed over the intervening 100 years I illustrate in Figure 2 a similar plot for a sample of Finnish boys. The results are quite similar to those obtained by Galton.

What does this have to do with econometrics? To answer this question we need to step forward in time about 50 years and consider a book published in 1933 by Horace Secrist. Secrist was a professor of economics at Northwestern University trained at Chicago and an expert in what we would now refer to as Industrial Organization. The book was titled *The Triumph of Mediocrity in Business* and had occupied 10 years of this research. In it Secrist showed in excrutiating detail if you grouped firms into performance categories in some initial year, and then followed them in subsequent years, that the initially most successful tended to do worse over time, while the least successful tended to improve. Secrist's conclusion was that American business was "converging to mediocrity." We will come back to this example later in the course.

References

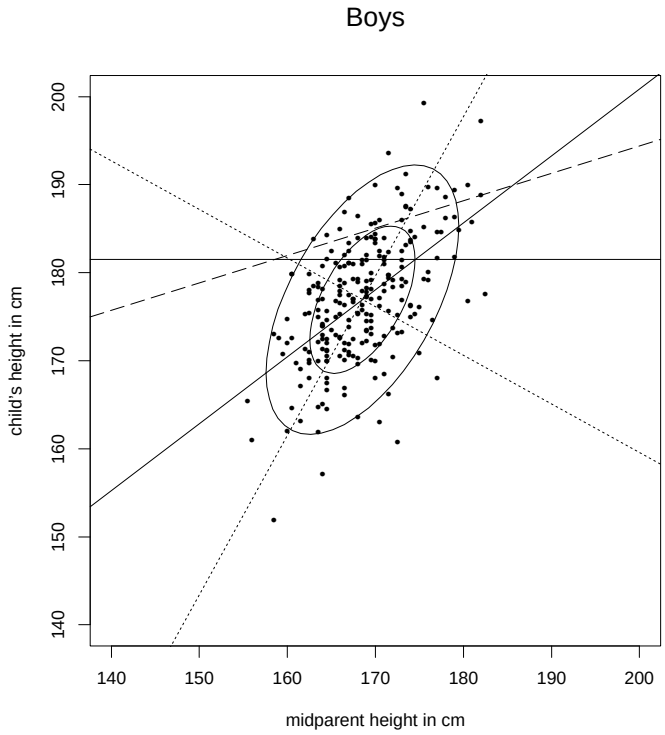

FIGURE 2. A Modern Galton Inheritance-of-Height Plot: The plot depicts the heights of 236 Finnish boys at age 17 versus the mean height of their parents. The two ellipses represent 50 and 90 percent confidence regions estimated for the pairs of points based upon the conventional bivariate Gaussian model for the data. The dotted lines depict the major and minor axes of the ellipses. The solid line represents the least squares fit; that is, it is the line that minimizes the sum of squared vertical distances from the points to the line. The slope of the line is $\beta = .76$, somewhat larger than the slope of $\frac{2}{3}$ obtained by Galton. What is the dashed line?

Stigler, S. (1996) The History of Statistics in 1933, Statistical Science, 11, 244-252.

Friedman, M. (1992) Do Old Fallacies Ever Die? J. of Econonomic Literature, 30, 2129-2132.

Hotelling, H. (1933) Review of Secrist, JASA, 28, 463-4.

Hotelling, H. (1934) Response to Secrist, JASA, 29, 198-9.
Secrist, H. (1933) *The Triumph of Mediocrity*, Northwestern U. Press.