University of Illinois Department of Economics
Fall 1999 Econ 472 Roger Koenker

Revised Appendix A
Concentration of Productivity in Phuzics Scholarship

This appendix presents a brief analysis of the "Parzen effect" in phuzics research. It will be recalled that Parzen (1985) hypothesized that one could discern the degree to which a discipline was "scientific" by studying the tail behavior of the distribution of annual productivity of the research community. Fields with an extremely long right tail, i.e., fields in which a large fraction of the published research is contributed by a small fraction of researchers, are scientific. Non-scientific fields exhibit a more uniform distribution of productivity.

Without judging the usefulness of this theory as a measure of scientific merit, we undertake a brief investigation of whether there is a trend in the tail exponent of phuzics research over the period 1970-1990. As an estimate of the tail exponent we employ Hill's (1975) estimator which is readily shown to possess optimality properties for distributions with a Pareto-Zipf right tail, i.e., for df's of the form

$$F(x) = 1 - (\beta/x)^{\alpha} \qquad x \geq \beta$$

where it provides an estimate of the tail exponent α.

We have a limited sample of only 123 researchers output over the period in question, so we use only the 10 most upper order statistics for each year. Thus,

$$\hat{\alpha}_t = (m^{-1} \sum_{i=1}^{m} \log(Y_t^{(i)}/Y_t^{(m+1)}))^{-1}$$

where $m = 10$, and t ranges from 1970 to 1990. These estimates are plotted in Figure 1, and appear to indicate a strongly positive trend which we estimate there by superimposing a log linear trend model. An increasing trend in α_t would suggest phuzics was becoming less scientific over the period of investigation, a claim supported by some critics, but one which we find difficult to accept on face value.

A serious difficulty with the analysis of the estimates plotted in Figure 1 is that because the field is growing the estimates are based on radically different sample sizes. Thus, for example, in the early 70's we have only about 30 total observations while at the end of the period we have over 100. To adjust for this effect, we began by running a small Monte Carlo experiment in which we generated data from the standard $\mu = 0, \sigma = 1$ lognormal distribution and then applied Hill's estimate for a variety of sample sizes. The results of this experiment are reported in Figure 2.

It is clear that the tail exponent is biased downward substantially for lower sample sizes. To correct for this bias we employed a bootstrap bias correction of the following form: under the null hypothesis of no trend in productivity we constructed a sample of annual productivity over the entire period 1970-1990. From this sample we drew random samples (with replacement) for each of the sample sizes corresponding to the observed annual samples. The annual estimates were then bias-corrected as

$$\tilde{\alpha}_t = \hat{\alpha}_t \cdot \frac{\hat{\alpha}_N}{\bar{\alpha}_{n_t}}$$

where $\hat{\alpha}_N$ denotes Hill's estimator based on the pooled population and $\bar{\alpha}_{n_t}$ denotes the mean of the bootstrapped estimates based on a subsample of n_t observations.

The bias corrected observations appear as circles in, Figure 3, and it is clear from the figure that the trend suggested by the original observations was essentially attributable to the sample size effect.

References

Parzen, I. (1985) Concentration of Scientific Research, *Annals of the Academy of Science of Monte Carlo*, 18, 463-468.

Hill, B. (1975) A Simple General Approach to Inference About the Tail of a Distribution, *Annals of Statistics*, 3, 1163-1174.

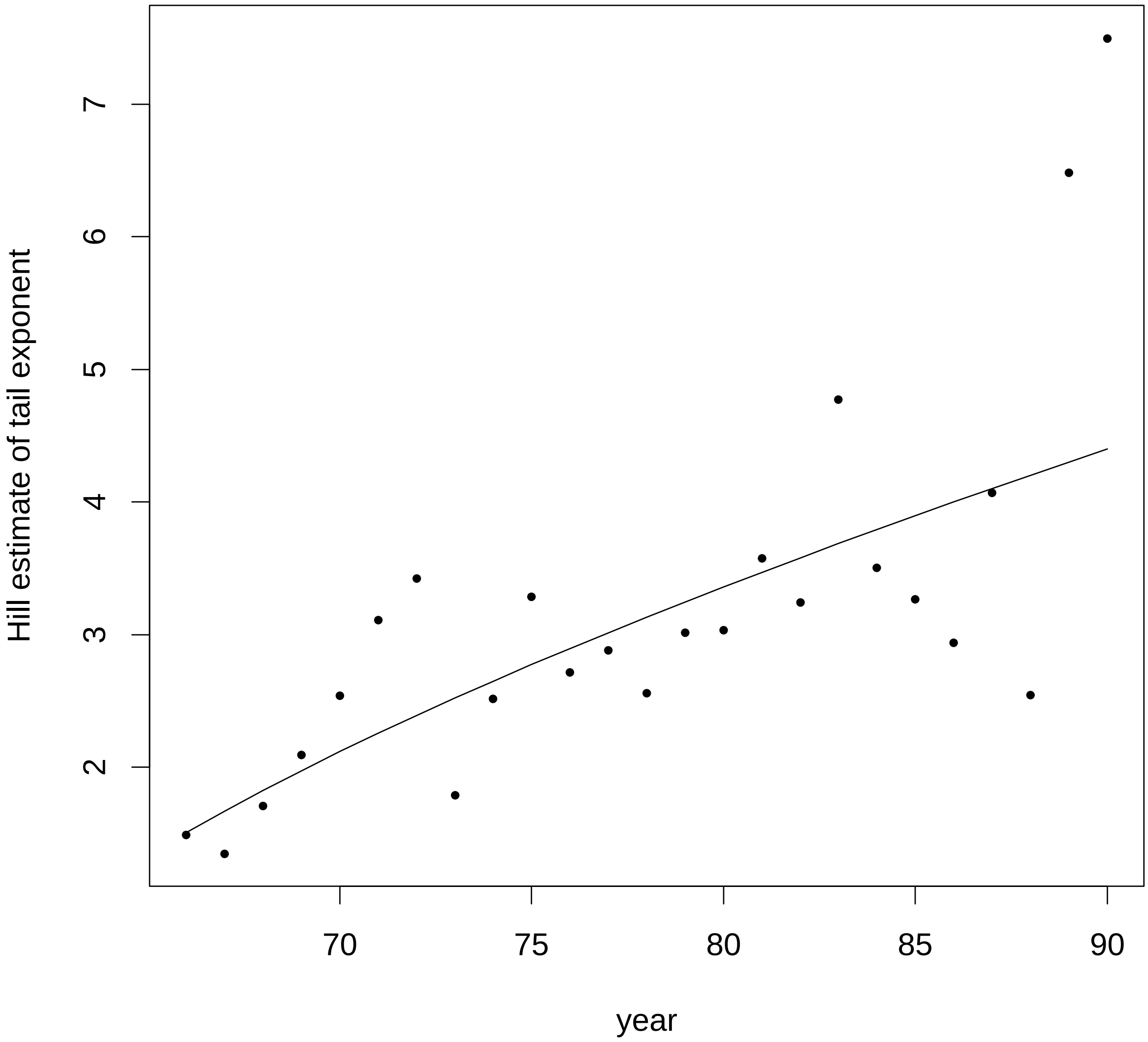

Figure 1: Tail Exponent of Phuzics Productivity Distribution 1970-1990: The plotted points indicate annual estimates of tail exponent of the productivity of phuzicists from 1970 to 1990. The tail exponent is estimated in each year using Hill's estimator based on the 10 largest order statistics of observed productivity. Note that since these years have rather different total sample sizes there may be some bias in this procedure. The curve superimposed on the data is a least squares fit of a log-linear trend.

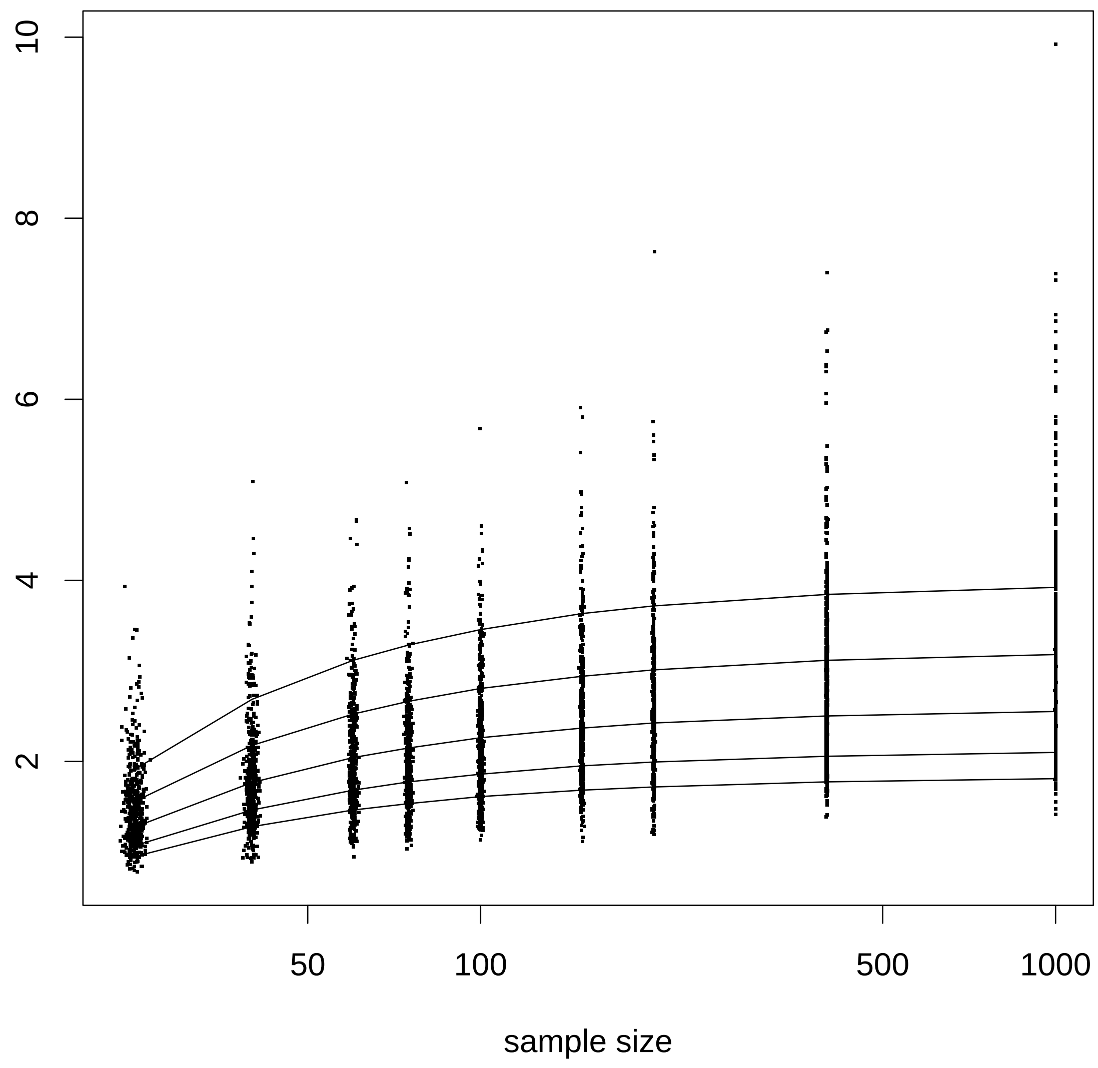

Figure 2: Monte Carlo Evaluation of Hill's estimator: This figure illustrates the result of a small Monte Carlo experiment designed to evaluate the performance of Hill's estimator of the tail exponent. The plotted points indicate 500 evaluations of Hill's estimator based on the 10 largest order statistics (m=10) for 9 different sample sizes: 25,40,60,75,100,150,200,400,1000 from the standard log-normal distribution. Clearly there is considerable bias and heteroscedasticity over this range of sample sizes. To provide an adjustment for these effects five quantile regression estimates have been superimposed on the plotted data corresponding to the quantiles .1,.25,.5,.75,.9 . The quantile regression model is linear in the reciprocal of the sample size.

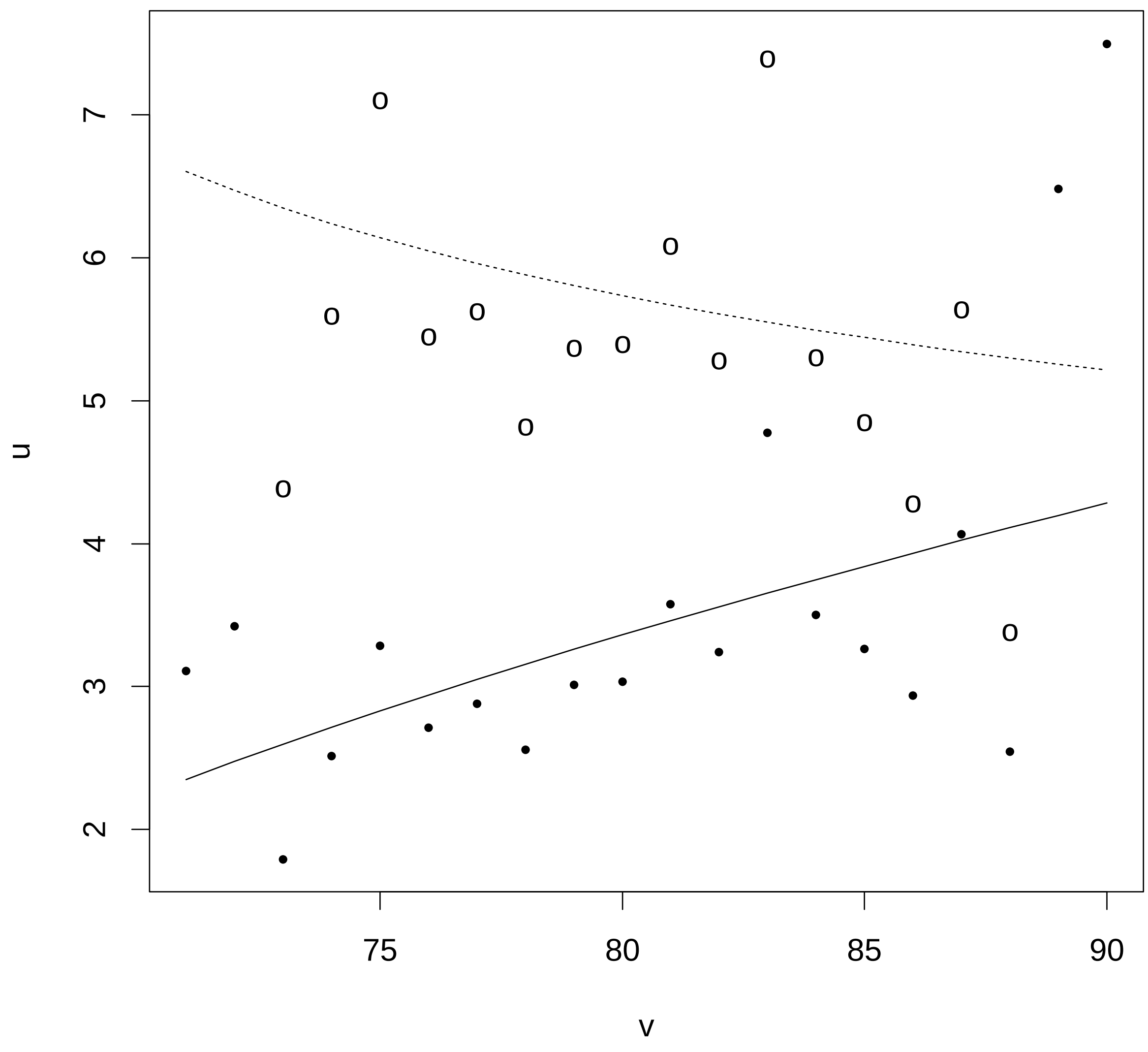

Figure 3: Tail Exponent of Phuzics Productivity Distribution 1970-1990: The plotted solid dots indicate annual estimates of tail exponent of the productivity of phuzicists from 1970 to 1990; the plotted circles denote the same estimates after bootstrap bias adjustment for bias due to sample size discrepancies across years. The tail exponent is estimated in each year using Hill's estimator based on the 10 largest order statistics of observed productivity. The curves superimposed on the data are least squares fits of a log-linear trend to the two samples. In the case the bias corrected observations the (dotted) fit is weighted least squares based on the variance of the bootstrap replications.