NBER WORKING PAPER SERIES

ESTIMATING THE COVARIATES OF HISTORICAL HEIGHTS

James Trussell

Kenneth Wachter

Working Paper No. 1455

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 1984

The research reported here is part of the NBER's research program in Development of the American Economy. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Working Paper #1455
September 1984

Estimating the Covariates of Historical Heights

ABSTRACT

Data on human height can provide an index that may measure more accurately changes in the standard of living than the more conventional real wage index. Height data, like those on real wages, are relatively abundant and extend back to the seventeenth century. In a previous paper, we developed and tested procedures for estimating the mean and standard deviation of the distribution of human height when the sample is distorted to an unknown extent by missing observations at lower heights. The purpose of this analysis is to extend our techniques so that the covariates of height can be estimated. Such an extension is necessary when trying to draw inferences about the causes of shifts over time in the height distribution so that changes in sample composition can be controlled.

James Trussell
Office of Population Research
Princeton University
21 Prospect Avenue
Princeton, NJ 08544

(609) 452-4946

Kenneth Wachter
Graduate Group in Demography
2234 Piedmont Avenue
Berkeley, CA 94720

(415) 642-9800

In a previous paper, we developed and tested procedures for estimating the mean and standard deviation of the distribution of human height when the sample is distorted to an unknown extent by missing observations at lower heights [Wachter and Trussell, 1982]. The purpose of this analysis is to extend our techniques so that the covariates of height can be estimated. Such an extension is important because the parent project, of which our own research is a small part, is aimed at using data on changes in height over time to infer changes in standard of living [Fogel, et al., 1982; Fogel, Engerman and Trussell, 1982]. In order for such inferences to be valid, we must ensure that observed changes are not caused simply by shifts in sample composition. Other things being equal, for example, we would expect that the average height of Americans would fall when large numbers of eastern Europeans immigrated to the United States after the turn of the century. Before we can make inferences about changes in standard of living, we must control for the shifts in composition. An obvious way to do so is to examine what height trend there would have been if the sample composition had remained fixed.

The Problem

The available data are drawn primarily from military recruitment records, though one sample from a charity called the Marine Society consists of London children [Floud and Wachter, 1982] and another (the only one containing women) is composed of slaves [Trussell and Steckel, 1978; Margo and Steckel, 1982]. A common feature of most of the samples is an underrepresentation of persons at the lower heights. This feature no doubt reflects minimum height standards for entry into the military or charity, but comparison of observed histograms with published minimum height standards reveals that the standards were flexibly rather than rigidly enforced. Not all persons shorter than the standard are missing, and the observed distributions are obviously deficient at heights above the minimum, some nearly as large as the mean.

Our original goal was to make inferences from a sample, which suffers from the obvious shortfall in observations discussed above as well as from other distortions such as heaping on preferred digits, about the true, unknown, underlying distribution. Without further structure, this problem would be insolvable. What makes it tractable at all is the well-documented fact that adult heights are normally distributed. With the knowledge that the deficient sample was drawn from an underlying normal distribution, we were able to devise two techniques for estimating its unknown mean and variance. One of these techniques is based on Quantile-Quantile Plots and will not be discussed further here. The other technique is based on fairly standard results for truncated normal distributions.

The RSMLE

Suppose that h_i is the height of individual i, that all observations below height a and none above it are missing. Then the likelihood function for this truncated normal distribution is

$$L = \prod_i \frac{\phi\left(\frac{h_i - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{a - \mu}{\sigma}\right)} \qquad (1)$$

where μ and σ are the unknown mean and standard deviation of the full (not truncated) underlying normal distribution and ϕ and Φ are the density and distribution functions, respectively, of the standard normal.

Since in our samples heights are ordinarily grouped into one-inch intervals, the likelihood function must be modified slightly. Let n_j be the number of individuals with height j inches in the sample. Then the likelihood function becomes

$$L = \prod_j \left[\frac{\Phi(\frac{j+1-\mu}{\sigma}) - \Phi(\frac{j-\mu}{\sigma})}{1 - \Phi(\frac{a-\mu}{\sigma})} \right]^{n_j} \qquad (2)$$

It is straightforward to maximize the log-likelihood using numerical techniques. We ourselves have employed the algorithm DFP [Powell, 1971] in the numerical optimization package OPT written by Goldfeld and Quandt.[1]

Of course, as described above, our samples are not sharply truncated; hence the likelihood function (2) would appear to be incorrect. Our technique, called the reduced sample maximum likelihood estimator, is based on the following reasoning. We do not know, by observation, how much of the lower tail of the height distribution is defective. But we can let the data tell us. We simply progressively chop off the lower tail, an inch at a time, until the estimates of the underlying mean and standard deviation cease to change.

In practice, of course, this criterion is too rigid. Unless the uncontaminated part of the height distribution conforms precisely to a normal, the sequence of estimated parameters will never converge as successive inches are lopped off. Moreover, because our observed data are only samples (and usually small samples of about 500), the estimated standard errors of the parameters will rise as more observations are discarded. Hence we are faced with compromising between two evils: accepting more of the lower tail increases the risk that estimates will be biased (the mean upward and the standard deviation downward) because there will be relatively too few short observations, while discarding more of the lower tail will decrease the precision of the estimates. Since differences in one inch in height over time or between populations are quite large, it is reasonable to demand that an estimated change this big be significant at the five percent level with samples of 300 to 400 heights. The rule that we developed for determining the discard point satisfies this goal. The discard point is successively raised by one inch until the estimated proportion of observations below it surpasses 28 percent. Discard points on either side are also reasonable candidates. For the purpose of stability, we choose the discard point associated with the median of the three estimated means.

Examples

Examples of the RSMLE are displayed in Table 1. Two data sets are employed. The first, kindly supplied by Georgia Villaflor, is a sample of the US Regular Army in 1850. The second, compiled by Lars Sandberg and Richard Steckel [1980] and generously made available for our use, consists of Swedish conscripts.

Examination of the results shown in Table 1 reveals several important lessons. First, the estimated means from the entire set of observations, shown in the last column for each sample, are considerably higher (by at least one inch) than the preferred RSMLE estimates, which are indicated by a "+" at the head of the column. Inferences based on the sample mean can therefore be very misleading. Second, although the threshold for the discard rule is 28 percent, the estimated proportion of the full distribution below the actual discard point chosen is generally higher and sometimes much higher. In the first two samples it exceeds 50 percent. Third, the estimated proportion of the sample missing entirely is often large, reaching 46 percent and 44 percent for the Swedish cohorts of 1800-09 and 1880-89, respectively.

Estimating Covariates

Estimation of the covariates of height is a natural extension of the RSMLE methodology. Once the discard point has been selected by the rule outlined above, one can simply let the parameters μ and σ of the normal depend on covariates:

$$\mu_i = X_i'b \qquad (3)$$
$$\sigma_i = Z_i'c$$

In equations (3), the mean and standard deviation for an individual i depend on his characteristics. We allow for the possibility that the vector of characteristics Xi determining the mean may be different from the vector Z_i determining the standard deviation. The likelihood function becomes

$$L = \prod_i \left[\frac{\Phi\left(\frac{h_i+1-X_i'b}{Z_i'c}\right) - \Phi\left(\frac{h_i-X_i'b}{Z_i'c}\right)}{1 - \Phi\left(\frac{a-X_i'b}{Z_i'c}\right)} \right] \qquad (4)$$

where h_i is the height of the ith individual and a is the discard point chosen by the RSMLE rule. Equation (4) can be modified to let the discard point a be different for different individuals, so long as it is predetermined.[2]

The implication of the above model is that the height distribution of a population is a mixture of normals. If $\sigma_i = \sigma$ is constant, then we have the truncated, grouped equivalent of regression analysis. In the examples to follow we assume that σ_i is constant, so that the connection with regression can more easily be made.

Estimates of the covariates of height for the US Army in 1850 and selected cohorts of Swedish conscripts are shown in Tables 2 and 3, respectively. These are illustrative calculations in which the mean depends on place of birth for the US sample and place of residence (and whether the conscript was urban born) for the Swedish samples. Note that all effects estimates are relative to the omitted category, Great Britain in the US sample and East and rural-born in the Swedish samples. Results are shown not only for the discard point selected by the RSMLE rule (as found in

Table 1) but also for the untruncated sample and for the other two discard points that enter into the RSMLE rule.

Several observations emerge from scrutiny of these results. First,and not surprisingly, as the discard point is raised, the standard errors of the estimated parameters rise. Second, the results of naive regression (including a grouping correction)[3] shown in the last column for each sample are quite different from the results obtained when the RSMLE discard rule is employed. By comparison, estimates from the three truncated samples implied by the three discard points that enter the RSMLE rule are generally closer to one another. Third, as likelihood ratio tests reveal, covariates add significant explanatory power.

Substantively, being born in the South significantly and substantially increases height among US soldiers in 1850, while being born in Germany significantly reduces it. Among Swedish conscripts, the estimated parameters change markedly over time. Being urban born is significant only in the 1800-09 cohort, where it has a strong negative effect on height. Coming from the West has a significant effect only in the 1880-89 cohort; residence in Stockholm and the North have significant effects only in the middle (1850-59) cohort.

Summary

In this paper we develop a method for estimating the covariates of height from samples that may be relatively deficient in observations at lower heights. Estimates derived from our technique are shown to differ sharply from those that would be obtained by naive regression analysis.

The difference between the two sets of results is of course data dependent. Regression analysis is appropriate only when observations at lower heights have not been selectively removed from the available sample. But this condition will in general never be met in samples of historical heights, for those persons with characteristics resulting in short stature will be precisely those who will be underrepresented.

FOOTNOTES

[1] Available from the Econometric Research Program, Department of Economics, Princeton University, Princeton, NJ 08544.

[2] For example, to achieve a large enough sample size one may wish to combine observations for several years, whose discard points are known to differ.

[3] Actually, even in the last column there is some truncation, since the discard point in a regression would be $-\infty$, but the lowest discard point in the table is the lowest observed height. Setting the bottom cutoff even lower changes the estimates only trivially, however, in these examples, so that the last column is indeed the grouped equivalent of regression.

REFERENCES

Floud, Roderick and Kenneth Wachter [1982]. Poverty and phsyical stature: evidence on the standard of living of London boys 1770-1870. Social Science History 6(4):422-452.

Fogel, Robert, et al., [1982]. Changes in American and British stature since the mid-eighteenth century: a preliminary report on the usefulness of data on height for the analysis of secular trends in nutrition, labor production, and labor welfare. Working Paper No. 890. National Bureau of Economic Research: Cambridge, Mass.

Fogel, Robert, Stanley Engerman and James Trussell [1982]. Exploring the uses of data on height: the analysis of long-term trends in nutrition, labor welfare and labor productivity. Social Science History 6(4):401-421.

Margo, Robert and Richard Steckel [1982]. The heights of American slaves: new evidence on slave nutrition and health. Social Science History 6(4):516-538.

Powell, M.J.D. [1971]. Recent advances in unconstrained optimization. Mathematical Programming 1:26-57.

Sandberg, Lars and Richard Steckel [1980]. Soldier, soldier, what made you grow so tall? Economy and History 23:91-105.

Trussell, James and Richard Steckel [1978]. The age of slaves at menarche and their first birth. Journal of Interdisciplinary History VIII (3):472-505.

Wachter, Kenneth and James Trussell [1982]. Estimating historical heights. JASA 77 (378):279-293.

Table 1. Reduced sample maximum likelihood estimates (RSMLE) for Swedish and U.S. data.

	Bottom Truncation Point			
	U.S. Regular Army, 1850			
	67"+	66"	65"	61"
Mean	66.55	66.34	67.28	67.86
S.D.	2.73	2.80	2.40	2.00
Prop. < bottom	.57	.45	.17	.00
Prop. missing	.30	.35	.13	.00
-ln likelihood	2487.6	3647.7	4522.6	5058.2
# Obs.	1490	2018	2300	2398
	Swedish Conscripts, 1800-1809 Cohort			
	67"+	66"	65"	58"
Mean	65.55	65.38	66.77	67.39
S.D.	2.57	2.62	2.09	1.71
Prop. < bottom	.71	.59	.20	.00
Prop. missing	.46	.50	.17	.00
-ln likelihood	962.9	1608.7	2157.7	2452.3
# Obs.	661	1005	1202	1246
	Swedish Conscripts, 1850-1859 Cohort			
	67"	66"+	65"	58"
Mean	66.89	67.00	67.81	68.15
S.D.	2.52	2.47	2.08	1.80
Prop. < bottom	.52	.34	.09	.00
Prop. missing	.31	.28	.09	.00
-ln likelihood	1379.9	1983.7	2329.7	2434.2
# Obs.	843	1104	1197	1204
	Swedish Conscripts, 1880-1889 Cohort			
	67"	66"+	65"	63"
Mean	65.04	66.39	67.50	68.01
S.D.	3.22	2.81	2.33	1.95
Prop. < bottom	.73	.44	.14	.01
Prop. missing	.60	.37	.13	.01
-ln likelihood	606.8	870.7	1053.0	1126.6
# Obs.	365	479	533	543

+ Preferred by RSMLE rule.

Table 2. Reduced sample maximum likelihood estimates of the covariates of height, U.S. Regular Army, 1850.

	Bottom Truncation Point			
	67"+	66"	65"	61"$
Ireland	-.49*	-.71	-.57	-.26
Northeast	.23*	.30*	-.06*	.14*
Mid-Atlantic	.27*	.65*	-.01*	.09*
South	1.51	1.79	1.31	1.06
Midwest	.58*	.98	.49*	.44
Germany	-.92	-.60*	-.75	-.49
Other	-.84*	-.43*	-.64*	-.43
Great Britain	66.82	66.67	67.60	67.97
S.D.	2.66	2.70	2.33	1.94
-ln likelihood	2472.2	3619.6	4485.7	5015.8

+ Preferred by RSMLE rule.

* Not significant at 10% level, 2-tailed test.

$ Approximates naive regression, with no truncation. See footnote 3.

Table 3. Reduced sample maximum likelihood estimates of the covariates of height, selected cohorts of Swedish conscripts.

Bottom Truncation Point

Birth Cohort, 1800-1809

	67"+	66"	65"	58"$
West	-.40*	-.04*	.20*	.14*
North	.69*	.81	.47	.40
Stockholm	.19*	.16*	-.77	-.31
Urban born	-3.15	-2.85	-1.03	-.45*
East	65.78	65.50	66.83	67.35
S.D.	2.53	2.55	2.05	1.69
-ln likelihood	959.6	1602.4	2142.0	2440.3

Birth Cohort, 1850-1859

	67"	66"+	65"	58"$
West	.62	.32*	.00*	.05*
North	1.01	.81	.58	.36*
Stockholm	-1.54	-2.21	1.64	-1.04
Urban born	1.05*	1.12*	.87	.50*
East	66.70	67.07	67.92	68.24
S.D.	2.47	2.40	2.04	1.78
-ln likelihood	1374.0	1969.5	2311.2	2419.0

Birth Cohort, 1880-1889

	67"	66"+	65"	63"$
West	-.92*	-1.53	-1.09	-.79
North	-1.11*	-.61*	-.60	-.46
Stockholm	-.93*	-1.36*	-1.23	-1.31
Urban born	-.15*	.96*	.65*	.45*
East	65.64	67.18	68.06	68.44
S.D.	3.18	2.72	2.27	1.91
-ln likelihood	605.7	864.9	1044.7	1118.0

+ Preferred by RSMLE rule.
* Not significant at 10% level, 2-tailed test.
$ Approximates naive regression with no truncation. See footnote 3.