sociological science

Supplement to:

Maralani, Vida, and Douglas McKee. 2017. "Obesity Is in the Eye of the Beholder: BMI and Socioeconomic Outcomes across Cohorts." Sociological Science 4: 288-317.

Supplemental Material on Data and Methods

Appendix Table A1 shows the full distribution of BMI across the two samples. At ages 19 to 24, most respondents in the 1979 cohort had BMIs concentrated in the normal range. Because the 1979 cohort has very few respondents with BMI values above 35 and statistical methods can be sensitive to extreme outliers, we restrict our analyses for this cohort to respondents with BMIs between 17 and 35. This omits 0.8% of the 1979 sample at the low end of the distribution (n=45) and 1% of the sample at the top end of the BMI distribution (n=60). The 1997 cohort had higher BMIs in early adulthood, reflecting the increase in BMI in the national population. At ages 19 to 23, the 1997 cohort has BMIs concentrated in the normal and overweight range. For this cohort, we censor BMI at 17 and 40 (rather than 35) in order to keep the censoring at the upper part of the distribution similar to the previous cohort. This omits 0.4% of the 1997 sample at the low end of the distribution (n=23) and 2% of the sample at the top end of the BMI distribution (n=129). Our substantive results do not change, however, if we instead censor the 1997 cohort at a BMI of 35 or even 45. Similarly, our substantive results do not change if we extend the 1979 cohort to a BMI of 40, although this introduces substantial uncertainty in the results for the top of the BMI distribution because there are very few respondents in this range in the 1979 cohort.

Methods

The partial linear regression model is specified as:

(1) $Y = f(BMI) + X\beta + \varepsilon,$

where f(BMI) is assumed to be a smooth nonparametric function with bounded first derivatives, $X\beta$ is a vector of independent variables that enter linearly, and ε is an independent and identically distributed mean-zero error term (Yatchew 1998). This model has two parts: a part that represents the nonparametric association between the outcome and BMI (f(BMI)) and a part that is a parametric function of the other covariates ($X\beta$).

The parameters of this model are estimated by first ordering the observations from the smallest to largest value of BMI. All the variables of the first two observations are then differenced and this difference becomes the first observation of a "new" dataset. Next, observations 2 and 3 are differenced, and then 3 and 4 are differenced, and so on, and each time the difference is kept as an observation in the new dataset. The idea is to compare cases that are very close in BMI values but may differ on the other covariates. Using these first-differenced

observations, the outcome is regressed on all covariates except BMI. This method gives unbiased estimates of the coefficients for the control variables, β in equation (1) (Yatchew 1998).

These coefficients are then used to net out the association between the control variables and the original dependent variable:

(2) $Y - XB \approx f(BMI) + \varepsilon$

The difference on the left is the original outcome residualized with respect to the control variables (but not BMI). This residual is then used in a standard bivariate locally weighted regression model with the original BMI variable. When the predicted value of f(BMI) is added back to the mean value of the original dependent variable, this traces out a semiparametric curve of the relationship between each socioeconomic outcome and BMI, net of set of covariates, in the original metric of each outcome. We do not show coefficients from these regressions because our interest is the relationship between each outcome and f(BMI), net of a set of basic controls, rather than the association between the controls and the outcomes (β).

The patterns traced out by any locally weighted regression depend on the bandwidth used for setting the local window in which the relationships are described. Because we use these semiparametric curves as a way of visualizing how the patterns across the entire distribution of BMI map to those between the standard thresholds (18.5, 25, 30), we choose narrow bandwidths for displaying our semiparametric graphs. We use a bandwidth of 0.2 for whites and 0.3 for blacks, which given the differences in sample sizes by race, produce similarly detailed curves for the groups. This level of detail ensures that the relationships around specific thresholds are measured close to that actual location, rather than being averaged with BMIs that are further away.

Appendix Table A1. Full Distribution of BMI in the 1979 and 1997 NLSY Cohorts

	NLSY-1979				NLSY-1997			
BMI	White Men	Black Men	White Women	Black Women	White Men	Black Men	White Women	Black Women
≤14	0	0	1	0	1	0	0	1
15	0	1	6	0	0	0	2	2
16	2	2	26	7	4	1	11	1
17	13	3	89	23	16	3	32	14
18	42	18	246	52	35	13	89	30
19	102	42	343	93	95	29	163	47
20	177	89	354	132	148	64	236	65
21	284	135	326	128	184	83	230	79
22	289	149	208	103	213	114	212	86
23	271	127	162	89	238	114	184	91
24	176	115	94	79	201	105	127	79
25	168	61	78	49	209	89	118	76
26	113	32	61	37	159	68	79	40
27	76	28	39	28	114	65	68	41
28	55	20	27	20	91	60	62	46
29	44	10	22	28	77	44	56	45
30	25	6	20	12	66	30	35	43
31	21	8	18	14	58	26	34	31
32	13	1	9	11	49	29	37	34
33	10	5	11	11	37	11	19	23
34	9	2	11	6	25	16	22	20
35	2	3	4	4	21	10	18	17
36	2	1	5	6	15	10	13	9
37	3	1	2	2	10	6	13	8
38	1	0	4	5	8	9	17	12
39	1	1	2	3	9	4	8	18
40	0	0	1	1	4	5	7	8
41	0	1	0	5	3	2	6	16
42	0	0	2	0	4	3	4	7
43	1	1	1	1	2	1	3	10
44	0	0	0	0	1	2	3	11
45	0	1	0	0	2	1	1	8
≥46	0	0	4	2	5	7	12	15
N	1900	863	2176	951	2104	1024	1921	1033
BMI distribution in standard categories (proportion shown)								
<18.5	0.02	0.01	0.10	0.05	0.02	0.01	0.04	0.03
18.5-24.9	0.70	0.78	0.75	0.69	0.52	0.50	0.63	0.45
25-29.9	0.24	0.18	0.10	0.17	0.31	0.32	0.20	0.24
≥30	0.05	0.04	0.04	0.09	0.15	0.17	0.13	0.28