Social Science History (2020), **44**, 525–554
doi:10.1017/ssh.2020.11

CAMBRIDGE UNIVERSITY PRESS

SPECIAL ISSUE ARTICLE

Recognizing Sample-Selection Bias in Historical Data

Ariell Zimran

Vanderbilt University and the National Bureau of Economic Research
Email: ariell.zimran@vanderbilt.edu

ABSTRACT

Recent research has ignited a debate in social science history over whether and how to draw conclusions for whole populations from sources that describe only select subsets of these populations. The idiosyncratic availability and survival of historical sources create a threat of sample-selection bias—an error that arises when there are systematic differences between the observed sample and the population of interest. This danger is common in studying trends in health as measured by average stature—scholars can often observe these trends only for soldiers and other similar groups; but whether these patterns are representative of those of the broader population is unclear. This article illustrates what simple patterns in a potentially selected sample can be used to recognize the presence of sample-selection bias in a source, and to understand how such bias might affect conclusions drawn from this source. Applying this intuition to the use of military data to describe stature in the antebellum United States, I present several simple empirical exercises based on these patterns. Finally, I use the results of these exercises to describe how sample-selection bias might affect the use of these data in testing for differences in average stature between the Northeast and the Midwest.

Introduction

One of the main challenges facing social science historians is how to use the limited data that historical circumstances caused to be created and that have survived to the present to learn about historical populations. In some cases, censuses and similar records provide a comprehensive view of complete populations—a distinct advantage over many modern data sources (Abramitzky 2015; Collins 2015). But more commonly, important pieces of information were recorded only for specific subsets of the population, and researchers must somehow use these limited records to learn about the broader population. This challenge is present in a broad class of settings, in particular those in which entry to a source depends on an individual's choice (Bodenhorn et al. 2017). The most prominent is the anthropometric history literature, in which population patterns in health are

reconstructed from height data recorded by militaries and prisons, among other sources (Floud et al. 2011; Fogel 1986).[1]

A danger in using such limited data to draw conclusions about historical populations is that conclusions may be incorrect if the available sources are systematically different from the population of interest. For example, in military height data, differences in observed height between two regions might reflect true differences, but might also reflect different labor market conditions in each region that cause the short to be more likely to join the military in one region than another. The error that results from drawing conclusions for a population from an unrepresentative sample is called *sample-selection bias*. The potential of sample-selection bias to generate spurious conclusions has long been recognized in the analysis of historical data, especially in anthropometric history (e.g., Fogel 1986; Fogel et al. 1983; Gallman 1996; Mokyr and Ó Gráda 1996).[2] But a recent debate in anthropometric history (ignited by Bodenhorn et al. 2017) has brought renewed attention to the question of how scholars can determine whether their conclusions are likely to be affected by sample-selection bias and how to work with sources that are suspected of having such a bias.[3]

In this article I use a simple theoretical example from the anthropometric history literature to identify patterns in a data set that are generated by—and thus give evidence of the presence of—sample-selection bias. Specifically, I focus on the use of military data to characterize population average stature and to determine the difference in average stature between the Northeast and the Midwest in the antebellum United States. A height advantage for the Midwest is an important result in American economic history (Komlos 2012) that is surprising in light of greater income in the Northeast than the Midwest (Easterlin 1960; McKeown 1976). But this difference might in part be the product of sample-selection bias that differed between regions (Bodenhorn et al. 2017; Mokyr and Ó Gráda 1996; Zimran 2019).

I then use the intuition coming from these patterns to conduct several exploratory exercises to determine whether and how sample-selection bias is likely to affect analysis in a sample of historical heights and specifically attempts to determine the Northeast–Midwest height difference from this sample. I find that there is suggestive evidence of negative selection into the sample and that the resulting bias may have caused the stature difference between regions to be overstated in the data (c.f., Zimran 2019). The patterns and exercises that I develop apply to

[1]Other examples include the use of data on sold or hospitalized slaves to study marriage behavior among all slaves (Fogel and Engerman 1974; Logan and Pritchett 2018), the use of marriage records or military data to study population literacy (Mitch 1993), and the use of naturalization records to study the consequences of immigrants' name changes (Biavaschi et al. 2017).

[2]It is also common to conduct exercises to determine whether samples are likely to be unrepresentative, or at least to acknowledge the potential danger of such bias (e.g., Logan and Pritchett 2018; Mokyr and Ó Gráda 1996; Steckel and Ziebarth 2016).

[3]The debate centers on a phenomenon known as the *Antebellum Puzzle*—a pattern of declining health in the antebellum United States in the presence of rising real income and wages (Floud et al. 2011). This pattern is commonly interpreted as indicating that early modern economic growth caused health to decline. Bodenhorn et al. (2017), however, have argued that because the data on stature that are used to establish this pattern come mostly from potentially selected military enlistment records, the pattern may be spurious, driven by changing selection into military service rather than by a true population height decline.

sample-selection bias generated either by *selection on observables*—a difference between the sample and the population on the basis of characteristics observable by the researcher—or by *selection on unobservables*—a difference between the sample and the population on the basis of characteristics unobservable to the researcher.[4]

The intuition and the exercises that I develop can be used to inform analysis in other cases in which there is concern that conclusions might be affected by sample-selection bias. They are thus complementary to Zimran's (2019) formal test and correction for sample-selection bias in historical heights, on which they are based.[5] This method is in turn an elaboration on a well-known procedure introduced by Heckman (1979) and discussed further by Vella (1998), which is based on the principle that studying the process by which individuals came to enter the sample, through the comparison of the sample to the complete population of interest on the basis of observable characteristics, can uncover sample-selection bias from both selection on observables and selection on unobservables. Although this method is well known, the intuition of the test and correction that it provides is often not well understood.[6] Moreover, its implementation is potentially costly in terms of data requirements and estimation.

There is thus a need for empirical approaches that test for sample-selection bias that are less data intensive and more intuitively straightforward. Bodenhorn et al. (2017) propose such a test that can be implemented using only the potentially selected sample, based on the logic that if the composition of military enlisters in a particular birth cohort responded to changes in the state of the economy over time, then long-run improvements in living standards must have affected the composition of height data over birth cohorts, and thus inferences from such data. This test has invited some criticism from contributors to the anthropometric history literature (e.g., Komlos 2019, 2020; Komlos and A'Hearn 2019; c.f., Bodenhorn et al. 2019). It is also limited in that it cannot provide definitive evidence of selection or information on its likely direction, and can test for only selection driven by one particular force.[7] Nonetheless, it is valuable in that it provides a simple test that can be applied using only the potentially selected sample.

The patterns and exercises presented in this article, in addition to providing clearer intuition for how sample-selection bias can be detected, are able to go further than the test of Bodenhorn et al. (2017) toward diagnosing and understanding the likely direction of sample-selection bias. They are able to do this through the addition to the selected sample of two additional pieces of information. The first and most important is a variable affecting selection into the sample, but that has no

[4]For instance, selection on observables might arise if urbanites were both shorter and more likely to enlist than ruralists and sector is observed. Selection on unobservables might arise if childhood health affected an individual's labor market outcomes (and thus the attractiveness of military enlistment to him) and his terminal height. Bodenhorn et al. (2017) and Zimran (2019) discuss these concerns in detail.

[5]Zimran (2019) is able to determine the role of sample-selection bias in military data in creating the Antebellum Puzzle. He finds that the data used to establish this result did in fact suffer from sample-selection bias. But he also finds that the magnitude of the bias was not sufficient to be solely responsible for the puzzling patterns discovered in this context.

[6]Bushway et al. (2007) discuss a number of problems that arise in the application of these methods in criminology, but, like most economics treatments of the subject, provide no intuition as to what the method does.

[7]That is, it is able to test only for selection arising from different lifetime labor market opportunities experienced by successive birth cohorts.

effect on the outcome. This variable enables the researcher to determine whether the likelihood of entering the sample is associated with the outcome—a hallmark of sample-selection bias generated by selection on unobservables. The second is a sample describing the population of interest, which enables the researcher to describe the determinants of entry into the potentially selected sample by comparing the sample to the population of interest. This piece of information is not crucial if the researcher is willing to make certain assumptions regarding the role of observable characteristics in determining entry into the sample.[8]

Despite the benefits that these patterns and exercises provide, it must be kept in mind that they are not formal tests or corrections for sample-selection bias. Only implementing the procedure proposed by Zimran (2019) or other variants of Heckman's (1979) procedure can provide such a test and correction. But they can be used by researchers to better understand in a transparent way whether sample-selection bias is likely to affect conclusions drawn from a suspect data source and, if so, how. Informed by the results of these exercises, researchers can decide whether and how to qualify their conclusions or even to implement a formal correction.

Theory

Selection on Observables

Consider the example of trying to determine the average height of the (northern) US population and the unconditional difference in average stature between Midwesterners and Northeasterners from a sample of military data.[9] Without information on how the sample was formed, it is impossible to determine whether the average stature of the military reflects that of the population, or whether any observed difference in average stature between regions reflects a true difference in the heights of the populations of each region, sample-selection bias induced by differences in selection into the sample across regions, or some combination of these two forces. That is, in the absence of information about how individuals came to enter the military, it is impossible to draw conclusions regarding population average stature from the average stature of the sample. In some countries' data, this challenge is overcome by conscription: if everyone (or a randomly selected group) were required to serve in the military, then observed heights could be taken as representative of those of the population.

However, if military enlistment was the product of individual choice, as in Britain and the United States in the nineteenth century, then the translation of the average stature observed in the sample to the average stature of the population, and thus the determination of the Northeast–Midwest difference in average stature, is less straightforward. To see this, consider the following example: (1) each region is divided into an urban and a rural sector; (2) ruralists are, on average, taller than urbanites; (3) there is

[8]Clearly the data on the population of interest need not include information on the outcome of interest (e.g., height). If they did, there would be no sample-selection problem. This source must simply describe some observable characteristics of the population that can be compared to those of the potentially selected sample to identify the determinants of entering the sample.

[9]The unconditional difference in average stature is the difference between the average height of all Northeasterners and the average height of all Midwesterners, not taking into account differences in urbanization, occupation, or any other characteristics. This is distinct from the coefficient on a regional indicator in a regression, which would capture the conditional difference.

Table 1. Example height distributions

	Average Heights		*Fractions*		*Average Height*
	(1)	(2)	(3)	(4)	(5)
Region	Urban	Rural	Urban	Rural	All
Panel A: Population (Actual)					
Northwest	67.00	69.00	0.75	0.25	67.50
Midwest	67.00	69.00	0.25	0.75	68.50
Panel B: Military (Observed)					
Northeast	67.00	69.00	0.90	0.10	67.20
Midwest	67.00	69.00	0.50	0.50	68.00

Notes: Panel A describes the population of interest. Columns 1 and 2 describe the average heights of each region-sector, and columns 3 and 4 describe the distribution of each region's population across these sectors (so that each row sums to one). Column 5 of panel A shows the true average height of each region, and thus the true difference in average heights between regions. Columns 3 and 4 of panel A are observed, but the other columns are not. Panel B describes the observed population—the military enlisters. The contents of all five columns are observed, but because the greater tendency of urbanites to enlist causes columns 3 and 4 to differ from panel A, the observed average height of each region and the difference between them does not match the true difference in panel A.

no difference in the average heights of individuals of the same sector across regions; (4) the only other determinant of height is genetic variation that is the same in each region-sector and averages away in random samples; and (5) the fraction of the population that is rural is greater in the Midwest than in the Northeast, implying greater average stature in the Midwest than in the Northeast. Panel A of table 1 presents an example of average heights satisfying these conditions. These are the true average heights—what the researcher wishes to learn but does not observe.

Consider first the extreme case in which only urbanites enlist in the military. Both regions' average heights would thus be understated, leading the researcher to underestimate the average stature of the population. A less extreme case allows both urbanites and ruralists to enlist, but retains the greater tendency for urbanites to enlist relative to ruralists. Such an example is illustrated in panel B of table 1. As the observed data would overrepresent urbanites relative to the population, the average stature of each region as observed in the enlistments would again understate the true stature, leading to an underestimate of the average stature of the population. If the urban status of enlisters is observed, then this is an example of selection on observables because the variable driving the nonrepresentativeness of the data (urban or rural status) by impacting both height and the probability of enlistment is observed.

The first pattern that sample-selection bias creates in a data source is evident from this example.

Pattern 1. *Selection on observables occurs whenever an observable characteristic that affects the outcome of interest is over- or underrepresented in the sample relative to the population of interest—that is, whenever an observable characteristic that affects the outcome also affects entrance into the sample.*

Such selection on observables would also affect the estimated Northeast–Midwest height difference. In the extreme example of enlistment only by urbanites, the observed heights of Northeasterners and Midwesterners would be the same despite the true Midwestern height advantage. In the less extreme example

of panel B of table 1, the regional differences in the sample would also not reflect regional differences in the population. In this example, selection on observables causes the observed difference in the heights of Midwesterners and Northeasterners (0.80 inches in panel B) to differ from the actual difference (1.00 inch in panel A).[10]

The researcher must make two determinations to ascertain whether selection on observables is likely present. The first is whether any given characteristic affects entry into the sample. Comparing the potentially selected sample to the random sample of the population (one of the two additional pieces of information discussed in the preceding text) enables the researcher to determine the factors affecting entry into the sample. If such a population sample is not available, it is possible to compare sample fractions to population fractions,[11] or to use theoretical or other knowledge of the environment in question if no other data are available.

The second is whether a given factor affects the outcome of interest. This is often known on theoretical grounds. It can also be determined from the observed data. If there is no selection on unobservables, then, as in the example, the sample is random conditional on the observables, and a simple regression analysis can reveal the relationship between observables and the outcome.[12] The lack of selection on unobservables in this example implies that the sample within each region-sector is random and observed heights represent actual heights in that region-sector. The only problem is that the fractions of each sector in the sample differ from those in the population. If the population fractions are known (as in this example), then it is possible to compute true average stature by combining observed stature for each region-sector with its population fraction. That is, in table 1, the researcher can compute population average heights using panel B's height data in columns 1 and 2, and panel A's fractions of the population in columns 3 and 4.[13]

[10]The contents of table 1 can be generated by the model

$$h_i = 67.00 + 2.00R_i + 0.00N_i + \varepsilon_i$$

$$P(y_i = 1) = 0.75 + 0.00N_i - 0.50R_i$$

where h_i denotes the height of individual i; R_i is an indicator equal to one if the individual lives in the rural sector; N_i is an indicator equal to one if the individual lives in the Northeast; y_i is an indicator equal to one if an individual enters the military (and his height is observed); and ε_i is a mean-zero stochastic error term that is uncorrelated with R_i and N_i (and thus is unrelated to military enlistment). Height is observed only if $y_i = 1$. The N_i is included in the model despite its zero coefficients in both equations to emphasize that the researcher is trying to learn the difference between the average height of each region, and so must allow for region-specific differences in height and enlistment probability (i.e., must include a Northeast indicator in regressions).

[11]For instance, the researcher might compare the fraction of individuals in the sample who are from urban areas to the fraction of individuals in the population who are from urban areas. A difference would suggest a role for sector in determining entry into the sample. In the absence of a sample of the population at risk for entry into the sample, these fractions might be available from census publications or other similar sources.

[12]The assumption of no selection on unobservables is the typical (and often implicit) assumption made in anthropometric history (Bodenhorn et al. 2017).

[13]The same logic applies when more than one variable affects selection or if the variable or variables affecting selection are continuous. If all variables affecting both height and entrance into the sample are observed, then selection conditional on these variables is random and the average stature of individuals with any given set of observables is known. Again, information on the distribution of these observables in the population enables the calculation of true average heights by providing the correct weights.

As a result of the relatively small data requirements to do so, sample-selection bias induced by selection on observables is relatively simple to recognize and address. Indeed, this is commonly done in anthropometric history (e.g., Fogel 1986; Fogel et al. 1983).

Selection on Unobservables

If enlisters' sector were not observed by the researcher, then the example in table 1 would be a case of selection on unobservables because an unobserved factor (in this case, sector) affects both height and entrance into the sample. The researcher would observe only the height difference in column 5 of panel B of table 1, and would not know how much of this difference is a true difference and how much is the product of selection on unobservables. More fundamentally, the researcher would have no information on whether the average height of the sample reflects that of the population. Such selection can arise even if there is no selection on observables, or even if a sample overrepresents portions of the population that the researcher is interested in studying.[14]

This bias can be better illustrated with another example. For this example, remove the urban–rural distinction so that all individuals are in the same sector and the distribution of heights is the same in each region. Instead, assume the following: (1) individuals differ in their wages, and only those with wages below a particular threshold enlist in the military; (2) lower wages imply lower stature; (3) average wages are higher in the Northeast; and (4) the relationship between height and wages is the same in each region once accounting for regional differences in wages. Figure 1 illustrates this example.[15] Higher wages in the Northeast are evident from the rightward shift of its wage–height relationship relative to that of the Midwest. The same distribution of heights in each region is illustrated by the same range of each line (and an implicit assumption of a uniform distribution along the line).

The most important assumption made in this example is that only individuals below a certain wage threshold are observed.[16] The intuition would be analogous

[14]For instance, even if the researcher is interested only in studying the working classes, a sample to which only the working classes were selected might also overrepresent the poorer members of the working class relative to the better-off ones (Bodenhorn et al. 2017). Moreover, even if a data set is identical to the population of interest on observables, this does not imply that it would be identical on unobservables (Kosack and Ward 2014).

[15]This example can be generated by the model

$$h_i = \beta_0 + \beta_1 N_i + \varepsilon_i$$

$$w_i = \alpha_0 + \alpha_1 N_i + u_i$$

$$y_i = \mathbf{1}\{\mathrm{w_i} < \overline{\mathrm{w}}\}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function; h_i is the height of individual i; N_i is an indicator equal to one for Northeasterners; w_i is the wage of individual i; y_i is an indicator equal to one for military enlisters; and ε_i and u_i have means of zero, are uncorrelated with N_i and $\mathrm{corr}(\varepsilon_i, u_i) = 1$. Height is observed only if $y_i = 1$. The perfect correlation between the errors is helpful for illustration, but is not necessary. If there is a nonzero positive correlation between them, the lines in figure 1 can be taken to represent regression lines, and the intuition is the same. The equivalence of average heights in the two regions implies that $\beta_1 = 0$, but β_1 is included to emphasize that the researcher is looking to learn this difference and cannot do so if region is excluded from the height equation.

[16]All the other assumptions can be reversed with the same basic result as long as the correlation between wages and stature is nonzero.

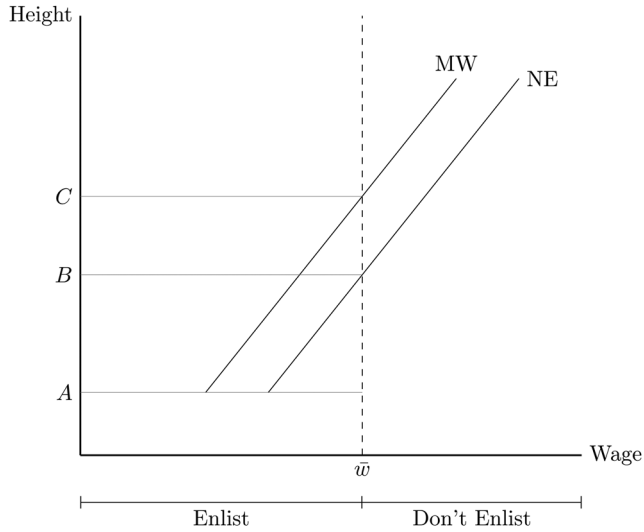

Figure 1. Hypothetical relationship of wages, heights, and military enlistment.

if it were instead assumed that enlisters were all above a particular threshold (this would generate positive selection rather than negative selection as in this example). However, it is crucial to assume that enlistment comes from one extreme or the other of the wage distribution. All the analysis in the following text (as well as the method of Heckman 1979) would fail if enlistment, for instance, came only from both extremes of the wage distribution (but not its center), or excluded its extremes.

Figure 1 shows that the relationship between wages and stature implies that only the shorter members of each region tend to join the military, leading the researcher to understate the average stature of the population from the observed data. Moreover, figure 2 shows that Midwesterners are more likely to enlist, as evidenced by the greater share of the Midwest's line that is below the cutoff for enlistment $\overline{w}$. Higher wages in the Northeast imply that there is a range of heights such that the wages are low enough to enlist in the Midwest but not in the Northeast (the range B to C). This would make the Midwest appear taller in the enlistments data—Midwesterners would have an observed average height of $\frac{1}{2}(C+A)$ while Northeasterners would have an observed average height of $\frac{1}{2}(B+A)$—even though it was assumed above that the distribution of height is the same in each region. If wages are not observed, then this is a case of selection on unobservables.[17]

[17]If wages were observed in both the population and the sample, then this would be another example of selection on observables. Because there are a range of wages whose heights are not observed (because they do not enlist), it is necessary to exploit the linear structure of the model to learn the true average heights. In a more realistic case in which the lines of figure 1 represent regression lines rather than true data (in the language of note 15, if the correlation of ε_i and u_i is positive but not equal to one), then this is simply a generalization of the example in table 1 with a continuum of values of the observables.

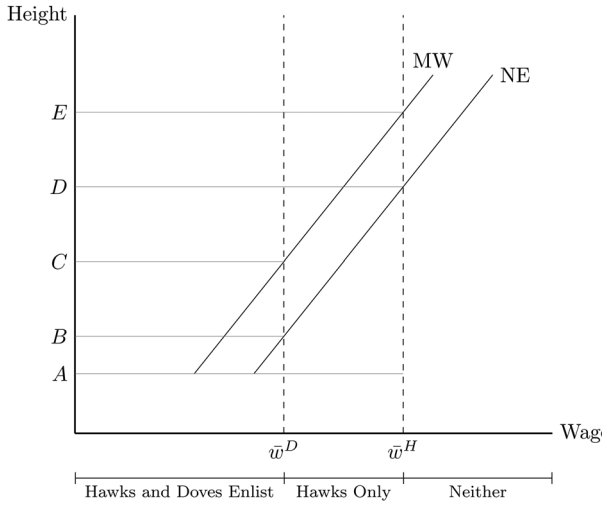

Figure 2. Hypothetical relationship of wages, heights, and military enlistment with hawks and doves.

This example provides the second pattern generated by sample-selection bias in the data.

Pattern 2. *Selection on unobservables that differs across groups creates differences across groups in the probability of entering the sample.*

It is important to note that this pattern is merely suggestive. Different probabilities of entering the sample (i.e., different enlistment probabilities) need not imply selection on unobservables. If wages were unrelated to height, then there would be a different enlistment probability in each region but no selection on unobservables. But under the assumption of selection from only one extreme, there cannot be selection on unobservables without such a difference in the probability of entering the sample.

The value of this pattern is that it is typically easy to check for it in most data sources. It is essentially the same as determining whether selection on observables has occurred, but the observable is the indicator of group. It is not possible, however, to determine what the likely direction of selection bias is, though it may be possible to guess based on outside knowledge of the institutional environment.

But in the absence of additional information, there is ultimately no way for the researcher to know whether the average stature of the sample over or understates that of the population. There is also no way for the researcher to know whether the difference in stature observed between regions in this example is because of different incentives to enlist (the true reason) or because of differences in health between regions, as the literature on historical heights has usually interpreted such results. All that the researcher knows is that Northeasterners in the data are shorter on average than Midwesterners in the data.

A more definitive check for the presence of sample-selection bias, as well as the ability to determine the direction of the bias induced by selection on unobservables, is possible with additional information if the information satisfies certain conditions. Continuing the example depicted in figure 1, make the following additional

assumptions: (1) the population is divided between hawks and doves; (2) hawk–dove status in the population and in the military is observable; (3) the division between hawks and doves is independent of height and wage so that the distribution of heights and wages in each region is the same between hawks and doves; and (4) the threshold wage for hawks' enlistment is higher than that of doves.[18] This is a simplification of the idea that ideology played a role in driving military enlistment in the Civil War (Zimran 2019). The crucial assumption here is that hawks and doves differ only in their likelihood of entering the sample and not in their heights. Hawk–dove status is known as an *excluded variable*.[19] This excluded variable is the essential piece of information that enables the researcher to uncover selection on unobservables.

The value of the hawk–dove division in addressing the sample-selection problem stems from the following insight, illustrated in figure 2. While doves enlist only if their wages are below $\overline{w}^D$, hawks with wages in the range $[\overline{w}^D, \overline{w}^H]$ also enlist (as well as hawks with wages below $\overline{w}^D$). This implies that hawks have a higher probability of enlistment. It also implies that observed hawks (in the military) would be taller than observed doves (again, in the military) in each region despite there being no relationship of hawk–dove status with height in the population: hawks in the Northeast who are observed in the military include individuals of heights A to D, while observed doves in that region include only those of heights A to B; similarly, observed hawks in the Midwest include individuals of heights A to E, while observed doves in that region include only those of heights A to C. That is, hawk–dove status has no relationship to height in the population; but because it affects the military enlistment decision, bringing individuals with wages between $\overline{w}^D$ and $\overline{w}^H$ into the sample, the observed average height of hawks is greater than that of doves.

This observed difference in height between hawks and doves is the third pattern created by sample-selection bias.

Pattern 3. *Selection on unobservables causes a variable that affects selection into the sample but is unrelated to outcome of interest in the population to be related to the outcome in the sample.*

Pattern 3 is essentially the logic underlying the "diagnostic test" proposed by Bodenhorn et al. (2017). The implicit assumption made in that case is that, within a birth cohort, year of enlistment should be unrelated to population height, but is

[18]This example can be represented by the model

$$h_i = \beta_0 + \beta_1 N_i + \varepsilon_i$$

$$w_i = \alpha_0 + \alpha_1 N_i + \left(\overline{w}^D - \overline{w}^H\right) H_i + u_i$$

$$y_i = \mathbf{1}\{\mathrm{w_i} < \overline{\mathrm{w}}^{\mathrm{D}}\}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function; h_i is the height of individual i; N_i is an indicator equal to one for Northeasterners; w_i is the wage of individual i; y_i is an indicator equal to one for military enlisters; H_i is an indicator equal to one for hawks; and ε_i and u_i have means of zero, are uncorrelated with N_i and H_i, and $\mathrm{corr}(\varepsilon_i, u_i) = 1$. The perfect correlation between these errors is helpful for illustration, but is not necessary. If there is a nonzero positive correlation between them, the lines in figure 2 can be taken to represent regression lines, and the intuition is the same. The equivalence of average heights in the two regions implies that $\beta_1 = 0$, but β_1 is included to emphasize that the researcher is looking to learn this difference and cannot do so if region is excluded from the height equation.

[19]This nomenclature comes from the fact that it is excluded from the equation determining height.

related to the probability of entering the sample. But pattern 3 is more general than that of Bodenhorn et al. (2017), which would capture selection on unobservables only from the mechanism on which they focus—changing incentives for enlistment over birth cohorts.[20]

As with Bodenhorn et al.'s (2017) test, pattern 3 can be identified in the selected sample alone under the assumption that the excluded variable affects entry into the sample and not the outcome. That is, only one of the two added pieces of information—the excluded variable—is necessary. But data on the population of interest can be used to test the assumption that the excluded variable affects entry into the sample. Specifically, comparison of the excluded variable in the sample and the population can more definitively determine that the variable in question is related to the probability of entering the sample. It is not possible to determine whether this variable has no relationship to the outcome in the population, so it must be assumed. This assumption is important because if the two were related in the population, then the hawks' height premium in the sample could also reflect an actual height premium for hawks in the population as much as a role of ideology in driving enlistment and thus selection on unobservables.

This excluded variable can also reveal the direction of the bias induced by selection on unobservables. As illustrated in figure 2, military enlisters are negatively selected on their unobservables—that is, only the shortest enlist. But this is not directly observed. Nonetheless, the fact that hawks are more likely to enlist than are doves, and that observed hawks are taller than observed doves within each region, indicates that the selection on unobservables uncovered by pattern 3 must be negative. Under the crucial assumption of selection from a single extreme (of the wage distribution), hawks draw in a greater fraction of their respective populations to enlistment. The fact that doing so brings in taller individuals implies that enlistment must be primarily from the bottom of the height distribution. Had hawks instead been observed to be shorter than doves, that would indicate that selection into military service was positive.[21] This is the fourth pattern created by sample-selection bias in the data.

Pattern 4. *If individuals whose value of the excluded variable makes them more likely to enter the sample are observed to be taller in the selected sample, then selection on unobservables is negative, and vice versa.*

This pattern is again something that can be determined from only the selected sample if the effect of the excluded variable on the probability of entering the sample

[20]Bodenhorn et al.'s (2017) test is based on the assumption that year-of-enlistment indicators (or year-of-enlistment indicators interacted with birth cohort indicators) affect (or capture forces that affect) the probability of enlistment but not population stature. This is likely a valid assumption in the case of changing economic conditions over time affecting military enlistment. But in other cases, such as the cross-sectional comparison studied in this article, this approach may not be effective. The more general focus in this article on the excluded variable thus makes the proposed exercises based on pattern 3 (and pattern 4 in the following text) more general.

[21]It is important to note that this is driven by selection from the bottom of the wage distribution. In this example with $\text{corr}(\varepsilon_i, u_i) = 1$, this implies observation exclusively of the bottom of the height distribution; but in a more realistic example in which $\text{corr}(\varepsilon_i, u_i) \in (0, 1)$, selection exclusively from the bottom of the wage distribution implies observation primarily but not exclusively of individuals from the bottom of the height distribution.

is known or assumed. But as with pattern 3, data on the population at risk to enter the sample enable the researcher to compare the selected sample to the population and more definitively to determine whether and how the excluded variable affects entrance into the sample.

Finally, suppose that there is a third group of individuals ("zealots") who enlist regardless of their wage, and continue to assume that the membership in the hawks, doves, or zealots is observed and unrelated to height. In this case, it is possible to learn the true heights of each region simply from the zealots. More generally, the bias in the observed height of each region is decreasing as the probability of entering the military increases from doves to hawks to zealots (where there is no selection on unobservables) and a greater fraction of the group is observed. The fifth pattern induced by sample-selection bias is generated by this example.

Pattern 5. *The more predisposed individuals are to be observed on the basis of their observable characteristics, the less is the sample-selection bias induced by selection on unobservables among these individuals. If there are individuals whose observable characteristics so strongly predispose them to enlist that their unobservables are unimportant, then there is no sample-selection bias among these individuals.*

This pattern shows that it is sometimes possible to solve the sample-selection problem by using only a limited portion of the data, though inference from this smaller sample will be less precise due to the smaller sample size. In general, a sample of the population at risk for observation is necessary to determine if any portion of the sample has sufficiently high population of entering the sample to perform such an analysis.

Patterns 2–4 can also shed light on how bias from selection on unobservables affects the Northeast–Midwest height difference in the sample. Pattern 2 showed that selection on unobservables that differed between regions created differences in the probability of entering the sample across regions. But the researcher could not be certain that such differences indicated selection on unobservables because there was no way of knowing whether the lines in figure 1 were upward sloping (leading to negative selection on unobservables) or flat (implying no selection on unobservables)—that is, whether higher wages are associated with greater height. However, with patterns 3 and 4 revealing negative selection on unobservables into observation, the researcher can conclude that the lines are upward sloping—wages and height are positively correlated. The greater probability for Midwesterners to be observed than Northeasterners thus draws in people of higher wage in the Midwest, and implies more negative selection into observation in the Northeast than in the Midwest. The Midwest's height premium is thus overstated. Indeed, despite there being no such premium by assumption, the different patterns of enlistment cause observed Midwesterners to appear taller in the observed data.

Multivariate Settings

In the preceding discussion, I have made the simplifying assumption that there are no observable characteristics that affect both the outcome and the probability of entering the sample. This assumption is helpful in clarifying the intuition and deriving the patterns, but it is unrealistic in practice. Relaxing this assumption requires some clarification of patterns 2, 3, and 4 to fit a multivariate context.

Pattern 2 used a difference between regions in the probability of entering the sample to suggest the presence of selection on unobservables that differs between regions. In a multivariate setting, it is possible for selection on unobservables to differ between regions without a difference in the probability of entering the sample by region. Instead, the difference that would arise would be in the *conditional selection probability*—the probability that an individual is observed given his observable characteristics. Selection on unobservables that differs between regions would result in a different distribution of these probabilities by region, which would generally, but not necessarily, result in a difference in the fraction of each region that is observed in the sample. Thus, in a multivariate setting, pattern 2 should be considered suggestive, and more information can be gleaned from examining the conditional selection probabilities, as will be done in the empirical exercises in the text that follows.

Pattern 3 allows the researcher to detect selection on unobservables by looking for a correlation in the sample between the outcome and the excluded variable. In a multivariate setting, in which variables other than the excluded variable affect entrance into the sample, it is the relationship between this variable and the outcome, conditional on all observables affecting the outcome, that is important.[22] Failure to control for observables might spuriously create a relationship. This is a relationship that can be tested using only the selected sample, though again data on the population can establish the relevance of the excluded variable to entrance into the sample.

Pattern 4 allows the researcher to determine the direction of the sample-selection bias induced by selection on unobservables from the sign of the correlation of the outcome and the excluded variable. With other observables, the relationship, as mentioned in the preceding text, is conditional on other observables.[23]

Data

I use these patterns to develop suggestive evidence regarding the presence and likely direction of sample-selection bias in a sample of US military data from the Union Army. I then explore how this bias, if present, might affect attempts to determine the Northeast–Midwest height difference. This analysis mirrors Zimran's (2019) formal investigation of this difference, though with somewhat different data. Revisiting this question enables me to demonstrate how the theoretical patterns derived in the preceding text can be used in practice.

Sources

The data for this analysis are taken from four main sources. The first is the potentially selected sample including the stature data (the outcome of interest) and

[22]For developing a general sense of whether or not selection on unobservables is present, it is generally sufficient to simply control in a linear sense. In a more formal correction (e.g., Heckman 1979; Zimran 2019) the precise way in which the controlling is performed is important.

[23]This is a simplification (and identical in intuition) to the Heckman (1979) approach of controlling for (a function of) the conditional enlistment probability. I will use the conditional enlistment probability rather than the excluded variable in some of the following exercises to more clearly illustrate the effect of sample-selection bias on the estimated Northeast–Midwest height difference.

covariates for military enlisters. It is based on Records of the Adjutant General's Office (1861–65). Data from this source are the products of two collections, each of which provides a random sample of enlisters in the Union Army, including data on stature, age at enlistment, date of enlistment, place of birth, place of enlistment, and occupation at the time of enlistment. The first is Fogel et al.'s (2000) Union Army Project, which provides information on a random sample of 16,285 enlisters. The second is Cuff's (2005) data set, which adds information on an additional 10,304 enlisters from the state of Pennsylvania.[24] The total number of observations is thus 26,589.[25] The oversampling of Pennsylvanians is a form of selection on observables. Because this bias is not that which typically concerns scholars in historical heights (because it is not generated by individuals' choices regarding enlistment), I simply weight all analyses so that the distribution of states of enlistment in the data matches that of the Union Army (Gould 1869).[26] I limit the data to native-born white males in the birth cohorts of 1820 to 1846, who were born and lived in the Northeast and Midwest, and who were at least 18 years old at the time of enlistment. Because the place of enlistment will be treated as the place of residence in the following analysis, I also exclude individuals who enlisted in a state other than the state of their regiment.[27]

The second source provides the description of the observable characteristics of the population at risk for military enlistment but not their height. Specifically, it is the 1 percent sample of the 1860 US Census (Ruggles et al. 2015). When applying the same filtering criteria as applied to the military data, this data set includes 28,205 individuals. It provides information on age, place of residence, and occupation.[28]

The third source is a collection of county-level data from the Census of 1860, provided by Manson et al. (2017), which gives information on county-level agricultural and manufacturing production and capital stocks, wealth, and population density. This information is assigned to individuals in the census sample based on their county of residence and to individuals in the military data based on their county of enlistment.

The fourth and final main source (ICPSR 1999) provides data on the excluded variable—voting patterns in the presidential election of 1860. The main variable of interest in this case is the share of each county's vote cast for Abraham Lincoln, the Republican candidate. These data are assigned to individuals in each sample in the same way as the county data from Manson et al. (2017). As the variable affecting entrance into the sample but assumed to have no direct effect on the outcome, these voting data are crucial to the exercises that follow and to implementing the insights

[24]These sources of height data also underlie Zimran's (2020) analysis of the effects of transportation on height in the antebellum United States.

[25]This is the sample size after all restrictions described in the following text.

[26]Due to small sample sizes, I omit enlisters from Minnesota, Missouri, and Rhode Island.

[27]For example, I assume that anyone enlisting in an Ohio regiment must have lived in Ohio, and omit anyone for whom there is a disagreement.

[28]It also provides information on outcomes such as school attendance and other household characteristics, but these are not used because they are not observed in the military records and comparisons of the military and the population at risk for enlistment on these dimensions is thus possible. Their impact on the military enlistment decision thus cannot be determined. Data linking enlistment records to the census (as in Zimran 2019) would enable this kind of analysis.

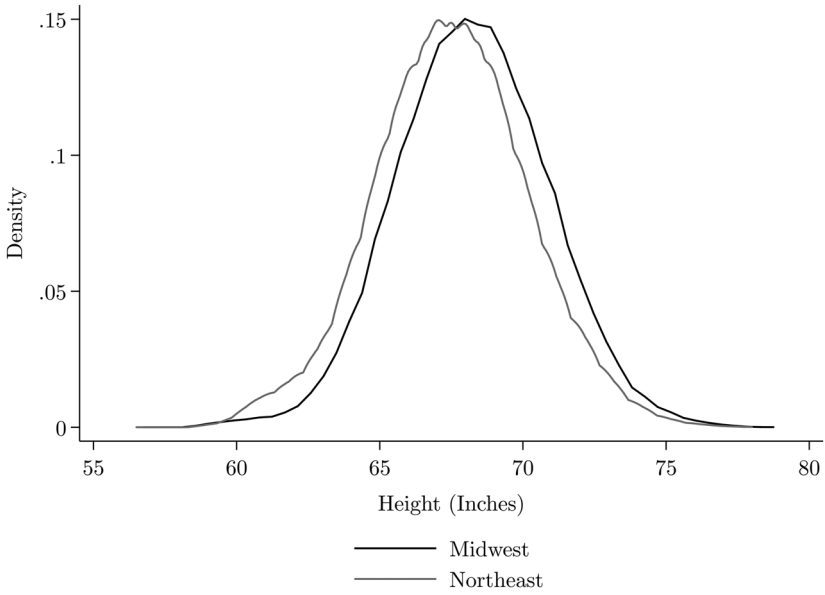

Figure 3. Height distributions by region.
Note: This figure presents the estimated height distributions for each region, combining both sources of military data and weighting observations to match Gould's (1869) distribution of states.

of patterns 3 and 4 mentioned previously. The impact of the political ideology represented by the vote for Lincoln on military enlistment will be demonstrated empirically in the following paragraphs (and has also been shown by Costa and Kahn 2003, 2007; Eli et al. 2018; Zimran 2019), and is unsurprising given that the Civil War was fought over the same issues that defined the 1860 election. The lack of a direct effect of ideology on height is untestable and must be assumed.[29] It is justified (as in Zimran 2019) by the claim that any association between the two would likely be the product of socioeconomic characteristics, which can be included as controls in any regression.

Summary Statistics

Figure 3 presents the distributions of observed heights of Midwesterners and Northeasterners, combining the Fogel et al. (2000) and the Cuff (2005) data with the use of the Gould (1869) weights. A height premium for the Midwest is evident.[30]

Table 2 presents summary statistics for variables observed for both the military and the population at risk for enlistment, as well as height, which is observed only

[29]This assumption could be tested if height were observed in the population, but this would obviate the concern over sample-selection bias.

[30]I do not address issues of truncation arising from minimum height requirements, which existed during the Civil War but do not appear to have been stringently enforced. Ultimately, this is another form of selection on unobservables. Komlos (2004) discusses common methods to address this issue. Zimran (2019, p. 122) explains how it relates to other forms of selection on unobservables.

Table 2. Summary statistics

	Military Only				Census Only				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variable	MW	NE	MW–NE	All	MW	NE	MW–NE	All	Difference
Height (Inches)	68.209	67.481	0.728[a]	67.830					
	(2.515)	(2.613)	[0.089]	(2.592)					
Lincoln Vote Share	0.525	0.591	–0.066[a]	0.559	0.531	0.572	–0.041[a]	0.555	0.004
	(0.109)	(0.107)	[0.018]	(0.113)	(0.113)	(0.104)	[0.104]	(0.110)	[0.006]
Midwestern				0.479				0.412	0.068[b]
				(0.500)				(0.492)	[0.028]
Birthyear	1838.324	1837.949	0.375[b]	1838.129	1835.963	1835.192	0.771[a]	1835.510	2.619[a]
	(6.322)	(6.495)	[0.179]	(6.415)	(7.461)	(7.766)	[0.114]	(7.651)	[0.106]
log(Population Density)	3.829	4.820	–0.991[a]	4.435	3.648	4.810	–1.162[a]	4.332	0.013
	(0.804)	(1.769)	[0.324]	(1.478)	(0.700)	(1.792)	[0.311]	(1.555)	[0.068]
log(Agricultural Value per Capita)	3.885	3.302	0.582[b]	3.581	3.976	3.308	0.668[a]	3.583	–0.001
	(0.527)	(1.292)	[0.229]	(1.043)	(0.428)	(1.307)	[0.219]	(1.090)	[0.047]
log(Manufacturing Value per Capita)	3.469	4.390	–0.921[a]	3.949	3.175	4.334	–1.159[a]	3.860	0.089
	(0.824)	(0.756)	[0.142]	(0.914)	(0.841)	(0.806)	[0.104]	(0.999)	[0.057]
log(Manufacturing Capital per Capita)	2.736	3.801	–1.064[a]	3.291	2.448	3.778	–1.330[a]	3.234	0.057
	(0.856)	(0.702)	[0.134]	(0.944)	(0.850)	(0.765)	[0.095]	(1.034)	[0.059]
log(Agricultural Capital per Capita)	5.562	5.222	0.340[b]	5.385	5.644	5.229	0.415[a]	5.400	–0.015
	(0.451)	(0.997)	[0.168]	(0.803)	(0.432)	(1.004)	[0.157]	(0.843)	[0.035]

(*Continued*)

Table 2. (*Continued*)

	Military Only				Census Only				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variable	MW	NE	MW–NE	All	MW	NE	MW–NE	All	Difference
log(Real and Personal Estate per Capita)	6.174	6.330	–0.156[a]	6.255	6.141	6.327	–0.185[a]	6.251	0.004
	(0.314)	(0.303)	[0.046]	(0.318)	(0.322)	(0.304)	[0.030]	(0.324)	[0.017]
White Collar	0.053	0.071	–0.018	0.062	0.096	0.170	–0.074[a]	0.140	–0.078[a]
	(0.224)	(0.256)	[0.011]	(0.241)	(0.294)	(0.376)	[0.021]	(0.347)	[0.010]
Skilled	0.164	0.318	–0.154[a]	0.244	0.153	0.283	–0.131	0.231	0.013
	(0.370)	(0.466)	[0.026]	(0.430)	(0.360)	(0.451)	[0.016]	(0.421)	[0.013]
Unskilled	0.065	0.208	–0.142[a]	0.139	0.087	0.147	–0.060[a]	0.123	0.017[c]
	(0.247)	(0.406)	[0.016]	(0.346)	(0.282)	(0.354)	[0.008]	(0.328)	[0.009]
Farmer	0.718	0.404	0.314[a]	0.555	0.665	0.400	0.265[a]	0.507	0.048[a]
	(0.450)	(0.491)	[0.032]	(0.497)	(0.472)	(0.490)	[0.031]	(0.500)	[0.016]
Observations	9,873	16,716		26,589	11,611	16,594		28,205	

Significance levels: [a] $p < 0.01$. [b] $p < 0.05$. [c] $p < 0.10$.
Notes: All figures in the enlistments are weighted to match Gould's (1869) distribution of states of enlistment. Standard deviations in parentheses. Standard errors, clustered by county, in square brackets.

for the military sample. Columns 1 and 2 present summary statistics for enlisters in the Midwest and the Northeast; column 3 presents difference-in-means tests comparing columns 1 and 2; column 4 presents summary statistics for all individuals in the military data; columns 5 and 6 present summary statistics for the population of the Midwest and the Northeast from the 1860 census sample; column 7 presents difference-in-means tests comparing columns 5 and 6; column 8 presents summary statistics for all census data; and column 9 presents difference-in-means tests comparing columns 4 and 8. In all cases, the enlistment data are weighted so that the distribution of states of enlistment matches the distribution presented by Gould (1869).

The first row of table 2 confirms the insight given by figure 3—Midwesterners were taller than Northeasterners in the observed sample by 0.73 inches. The second row of the table compares the vote shares for Lincoln. Northeasterners' counties of residence, both in the military and in the population, had a greater vote share for Lincoln than those of Midwesterners. This difference is about 7 percentage points in the military sample and about 4 percentage points in the population. Comparing the enlisters to the population (column 9) reveals virtually no difference between them on the basis of the voting variables, though this is only the unconditional difference.

Except for differences in the regional representation (the enlistment sample statistically significantly overrepresents the Midwest by about 7 percentage points) and in terms of birth year (enlisters are, on average, about 2.6 years younger), none of the other county-specific variables exhibits a large or statistically significant difference between the enlisting population and the census. For the only individual-level variables that are observed, the occupational indicators, there are differences between enlisters and the complete population. In particular, the enlisted sample overrepresents farmers and the unskilled, and underrepresents those with white-collar occupations. These patterns are typical of military enlistment in the nineteenth century (e.g., Margo and Steckel 1983; Zehetmayer 2011; Zimran 2019). But the comparison between the occupations of the military and census samples is complicated by the fact that they are observed up to five years apart and thus may not be directly comparable.[31]

Empirical Exercises

Selection on Observables

Table 2 provides some suggestive evidence pertaining to pattern 1—that selection on observables arises when factors affecting height are over- or underrepresented in the sample. Column 9 of table 2 shows that, for instance, higher skill occupations are underrepresented, indicating selection on observables if occupational skill is correlated with height in the population.

A more formal test for this pattern is given in table 3, which presents two sets of regressions describing the relationship between the observable characteristics described in table 2, on the one hand, and military enlistment and observed height, on the other. Columns 1–4 present the results of probit regressions for the

[31]For the birth cohort of 1847, for instance, the occupations in the census are of 13-year-olds, while the occupations at enlistment are from 18-year-olds in 1865.

Table 3. Relationship of covariates to enlistment probability and observed heights

Variables	(1) Enl	(2) Enl	(3) Enl	(4) Enl	(5) Height	(6) Height	(7) Height	(8) Height	(9) Height
Lincoln Vote Share	1.112[b]	0.903[c]	2.212[a]	1.852[a]					0.720[a]
	(0.530)	(0.489)	(0.696)	(0.586)					(0.334)
Midwest	0.624[a]	0.517[a]			0.538[a]	0.465[a]			
	(0.137)	(0.118)			(0.090)	(0.091)			
log(Population Density)	0.035	0.069	−0.106	−0.066	−0.140[b]	−0.138[b]	−0.217[a]	−0.211[a]	−0.218[a]
	(0.091)	(0.077)	(0.110)	(0.089)	(0.058)	(0.056)	(0.055)	(0.054)	(0.054)
log(Agricultural Value per Capita)	0.241	0.283	0.039	0.126	−0.239[c]	−0.301[b]	−0.262[c]	−0.340[b]	−0.304[b]
	(0.217)	(0.186)	(0.260)	(0.220)	(0.143)	(0.133)	(0.149)	(0.136)	(0.143)
log(Manufacturing Value per Capita)	0.363[b]	0.385[a]	0.679[a]	0.648[a]	−0.083	−0.064	0.012	0.039	0.037
	(0.165)	(0.148)	(0.209)	(0.179)	(0.096)	(0.094)	(0.102)	(0.101)	(0.104)
log(Manufacturing Capital per Capita)	−0.049	−0.141	−0.284	−0.318[c]	0.028	0.029	−0.064	−0.079	−0.084
	(0.160)	(0.145)	(0.191)	(0.168)	(0.090)	(0.089)	(0.090)	(0.089)	(0.092)
log(Agricultural Capital per Capita)	−0.200	−0.227	−0.224	−0.291	0.321[c]	0.368[b]	0.253	0.322[c]	0.263
	(0.204)	(0.173)	(0.258)	(0.223)	(0.181)	(0.168)	(0.206)	(0.181)	(0.201)
log(Real and Personal Estate per Capita)	−0.043	−0.062	0.083	0.025	−0.431[b]	−0.426[b]	−0.255	−0.247	−0.286
	(0.209)	(0.186)	(0.251)	(0.223)	(0.206)	(0.191)	(0.184)	(0.169)	(0.185)
Skilled		0.804[a]		0.837[a]		0.005		0.005	
		(0.066)		(0.065)		(0.081)		(0.080)	

(*Continued*)

https://doi.org/10.1017/ssh.2020.11 Published online by Cambridge University Press

Table 3. (*Continued*)

Variables	(1) Enl	(2) Enl	(3) Enl	(4) Enl	(5) Height	(6) Height	(7) Height	(8) Height	(9) Height
Unskilled		0.797[a]		0.805[a]		−0.049		−0.078	
		(0.088)		(0.093)		(0.098)		(0.097)	
Farmer		0.638[a]		0.667[a]		0.317[a]		0.324[a]	
		(0.061)		(0.060)		(0.086)		(0.084)	
Constant	−2.768[a]	−3.209[a]	−3.346[a]	−3.314[a]	69.905[a]	69.664[a]	69.624[a]	69.365[a]	69.460[a]
	(1.065)	(0.976)	(1.507)	(1.271)	(0.812)	(0.801)	(0.796)	(0.772)	(0.797)
Observations	54,660	45,651	54,660	45,651	26,585	25,470	26,585	25,470	26,585
R-squared					0.088	0.091	0.093	0.097	0.094
State FE	No	No	Yes	Yes	No	No	Yes	Yes	Yes
Pseudo R-squared	0.047	0.139	0.052	0.146					

Significance levels: [a] $p < 0.01$. [b] $p < 0.05$. [c] $p < 0.10$.

Notes: Dependent variable is an indicator for military enlistment in columns with the header "Enl" and height in inches in columns with the header "Height." The sample in columns 1–4 includes all individuals with height data or in the census sample, excluding residents of Missouri, Minnesota, and Rhode Island. Columns 5–9 include only individuals among these who are from the military sample. All specifications include birth year fixed effects and all specifications with height as the outcome also include age-of-measurement fixed effects to standardize age of measurement to age 21. In all specifications, the enlistment data are weighted to match the distribution of states of enlistment. Standard errors in parentheses, clustered at the county level.

probability of military enlistment, with columns 1 and 2 including a Midwest indicator, and columns 3 and 4 including state-specific fixed effects.

Due to the unusual structure of the sample, columns 1–4 are not estimated by an ordinary probit regression, though the interpretation of the coefficients is the same. In the standard setting, the researcher observes a random sample of the population with the military enlistment status of all individuals. In this setting, I observe a sample of military enlisters and their covariates and a sample of the complete population with their covariates but without information on the military enlistment decision.[32] Following Zimran (2019), I use Cosslett's (1981) method to estimate the model; I also use Zimran's (2019) weights, reflecting the general probability of entering the sample, which are necessary for this estimation.[33]

Columns 1–4 of table 3 show that, all else equal, individuals from counties with greater manufacturing production per capita were more likely to enlist. Individuals with skilled or unskilled occupations, or who were farmers, were also more likely to enlist than were individuals with white collar occupations (the excluded group).

Columns 5–8 present OLS regressions for the correlates of height in the military data without any correction for potential bias from selection on unobservables. Columns 5 and 6 include a Midwest indicator and columns 7 and 8 include state-specific fixed effects. All four of these specifications indicate that individuals from counties with greater population density and greater agricultural output per capita tended to be shorter. Moreover, individuals reporting an occupation of farmer at enlistment were, on average, about 0.30 to 0.35 inches taller than individuals reporting other occupations.

These results relate to pattern 1. For instance, the fact that farmers were taller and more likely to enlist than the white-collar workers implies that there was a variable affecting both entrance into the sample and height, and thus there was likely to be selection on observables. It must be noted that this evidence is only suggestive, because the regressions of columns 5–8 do not correct for potential selection on unobservables. The impact of manufacturing value on enlistment also raises suspicions of selection on observables—though columns 5–8 find no relationship of this variable with height in the sample, a population-level relationship is not out of the question. Thus, it is likely that the average observed stature in the sample does not correspond to the actual stature of the population, and that the unconditional Northeast–Midwest height difference in the data is also likely affected.

This analysis requires the use of the sample describing the observable characteristics of the population at risk for military enlistment. If only the selected sample were observed, then the researcher would be able to produce only columns 5–8 and not columns 1–4 of table 3. The height advantage of farmers and of individuals from areas of lower population density and agricultural value would suggest the presence of selection on observables if the researcher had reason to believe that these variables also affected the likelihood of enlisting in the military. The other advantage that comes from the availability of data on the population at risk for military

[32]There are two difficulties to be overcome. The first is that the proportion of military enlisters in the data is a function solely of the sample sizes of the two samples and not of the true population probability of enlistment. The second is that the sample of the population will contain unidentified military enlisters.

[33]Zimran (2018) provides Stata code for this estimation.

enlistment is that, as shown by Zimran (2019), the estimated conditional enlistment probabilities (another term for the conditional selection probabilities in this context) from columns 1–4 can be used to correct for sample-selection bias from selection on observables through the creation of inverse probability weights.

Selection on Unobservables

Table 2 provides suggestive evidence based on pattern 2—that different probabilities of entering the sample across groups suggest different selection on unobservables across these groups. The evidence in column 9 that Midwesterners were more likely to enlist suggests that indeed there may have been differences between regions in selection on unobservables that would affect the use of the military sample in determining population average stature and in testing for a regional difference in stature. Figure 4, which plots the distributions of estimated conditional enlistment probabilities from the estimates of column 3 of table 3, confirms this result, showing a greater mean conditional enlistment probability among Midwesterners. Contemporary reports of negative selection into military enlistment (Coffman 1986; Foner 1970; Weigley 1967), combined with these insights based on pattern 2, suggest that Northeasterners were more negatively selected than Midwesterners on unobservables, and therefore that the Midwest's height premium may have been exaggerated in the data. This will be explored in more detail in the following paragraphs.

Table 3 also provides evidence informed by pattern 3. Specifically, columns 1–4 show that the vote share for Lincoln enters with a positive and statistically significant coefficient, indicating that individuals from counties that were more supportive of Lincoln were more likely to enlist. The coefficients are not directly interpretable because these are probit coefficients, but it can be shown that the coefficient 2.21 in column 3 indicates that a 10-percentage point increase in Lincoln's vote share was associated with an increase in the probability of enlistment by 7.2 percentage points, relative to a base probability of 44.6 percent. Crucially, column 9 of table 3 shows that there is a positive and strongly statistically significant relationship between the vote share and height in the military sample. Under the assumption that voting patterns are unrelated to height in the population, these results suggest the presence of sample-selection bias induced by selection on unobservables based on the logic of pattern 3. The sample describing the population at risk for military enlistment is crucial to determining that there was in fact a positive relationship between the vote for Lincoln and the probability of entering the military.

The positive and statistically significant coefficient on the Lincoln vote share in column 9 of table 3 also speaks to pattern 4. That the vote for Lincoln is positively associated with enlistment probability (column 3) and observed height implies that the selection on unobservables suggested by pattern 3 is likely negative. That is, individuals from the bottom of the height distribution were likely overrepresented in enlistment. This suggestive finding of negative selection is consistent with the suggestion of Bodenhorn et al. (2017), with the results of Zimran (2019), and with contemporary reports of the characteristics of military enlisters (Coffman 1986; Foner 1970; Weigley 1967).

The combination of the insights in table 3 from patterns 3 and 4 suggests that there likely was sample-selection bias caused by selection on unobservables and that

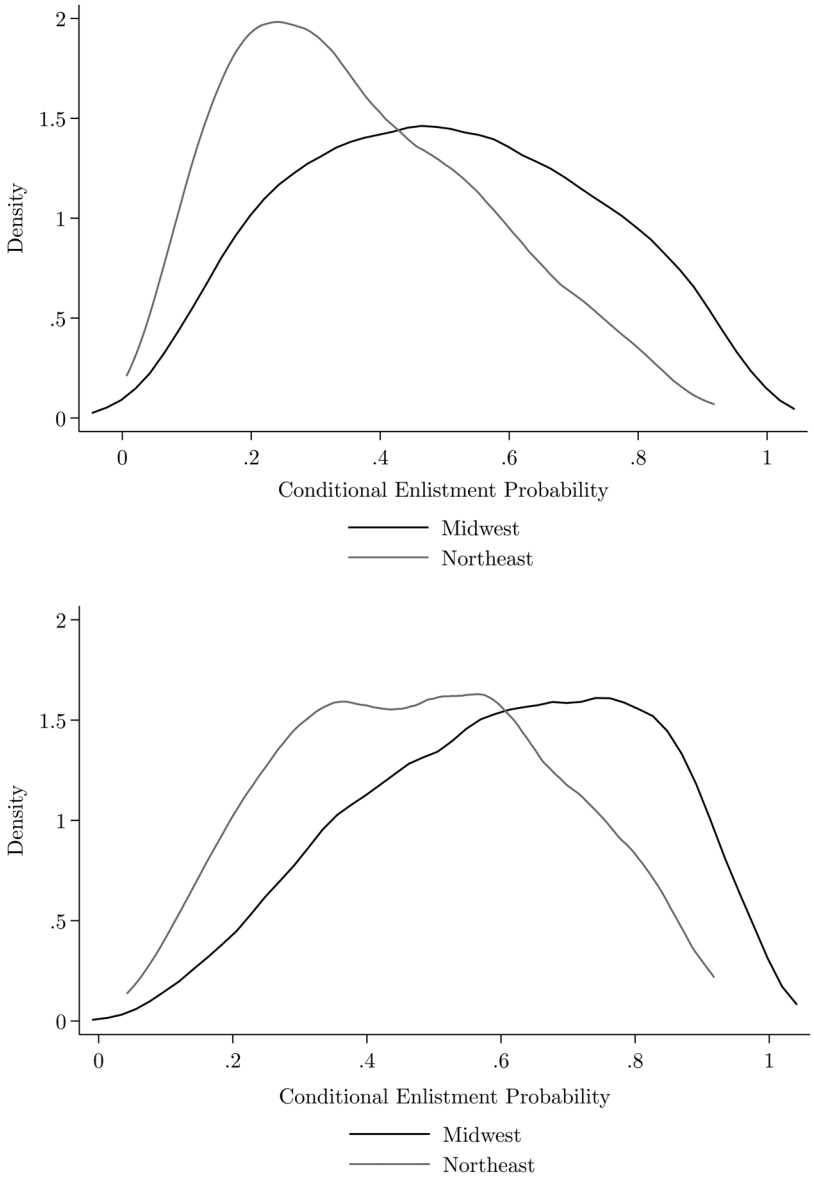

Figure 4. Distribution of estimated enlistment probabilities by region.
Note: Enlistment probability is estimated from the results of the probit regression of column 3 of table 3. The top panel describes the distribution in the sample of census data. The bottom panel describes the distribution in the sample of military enlistment data.

this bias would cause the researcher to understate the average stature of the population from these military data.

A more common concern than whether there is bias induced by selection on unobservables is whether such a bias would affect conclusions regarding trends or differences in the outcome over time or space. For example, researchers might not seek to use height data to describe the average stature of the population as a whole—though this is done (e.g., Fogel 1986; Floud et al. 2011)—but instead to describe trends in average stature over time or differences over space. In this case, it is not the presence of sample-selection bias that is important, but whether it varies over time or space. Fortunately, a more detailed analysis based on patterns 3–5 can shed light on whether the Northeast–Midwest height difference in the sample can be taken as informative of a true Northeast–Midwest difference in average stature.

The top panel of figure 5 plots a nonparametric regression of height on the estimated conditional enlistment probabilities separately by region. The key feature in this graph, building on pattern 5, is that the height premium for the Midwest is present among those with conditional enlistment probabilities close to one. Pattern 5 concluded that individuals so predisposed to enlist on the basis of their observable characteristics that they do so almost regardless of their unobservables have no selection on unobservables. Patterns among these individuals can thus be taken as unaffected by sample-selection bias. The presence of a Midwestern height premium at the right extreme of the top panel of figure 5 indicates that even though there is sample-selection bias (as shown previously) it is unlikely to have produced a spurious Midwestern height premium. Indeed, the presence of a Midwestern height premium at all levels of the conditional enlistment probability, at which the level of selection on unobservables is constant,[34] provides validation to the existence of a true Midwestern height premium notwithstanding the presence of selection on unobservables.

The bottom panel of figure 5 changes the *y*-axis of the figure to be the residuals of height after a regression on all observable characteristics except region.[35] This adjustment reverses the direction of the slope of the relationship of height and the conditional enlistment probability, indicating that the negative relationship in the top panel is the product of observable characteristics that drive enlistment also being associated with lower stature (i.e., of negative selection on observables). When controlling for these observables, however, the upward slope, following patterns 3 and 4, indicates negative selection on unobservables into the military in both regions. That is, those with a greater probability of enlistment (analogous to hawks in the example) were taller than those with a lower probability of enlistment (analogous to doves in the example), just as in patterns 3 and 4.

Figure 6 uses this graph to investigate in more detail how the presence of sample-selection bias induced by selection on unobservables would affect the comparison of

[34]That selection is constant for individuals with the same conditional enlistment probability is one of the main results of Heckman (1979).

[35]I do not use residuals of enlistment probability because, as Heckman (1979) shows, it is the enlistment probability that determines the degree of selection on unobservables. Residual variation in the enlistment probability is used only for identification. Zimran (2019, p. 110) discusses this distinction in detail.

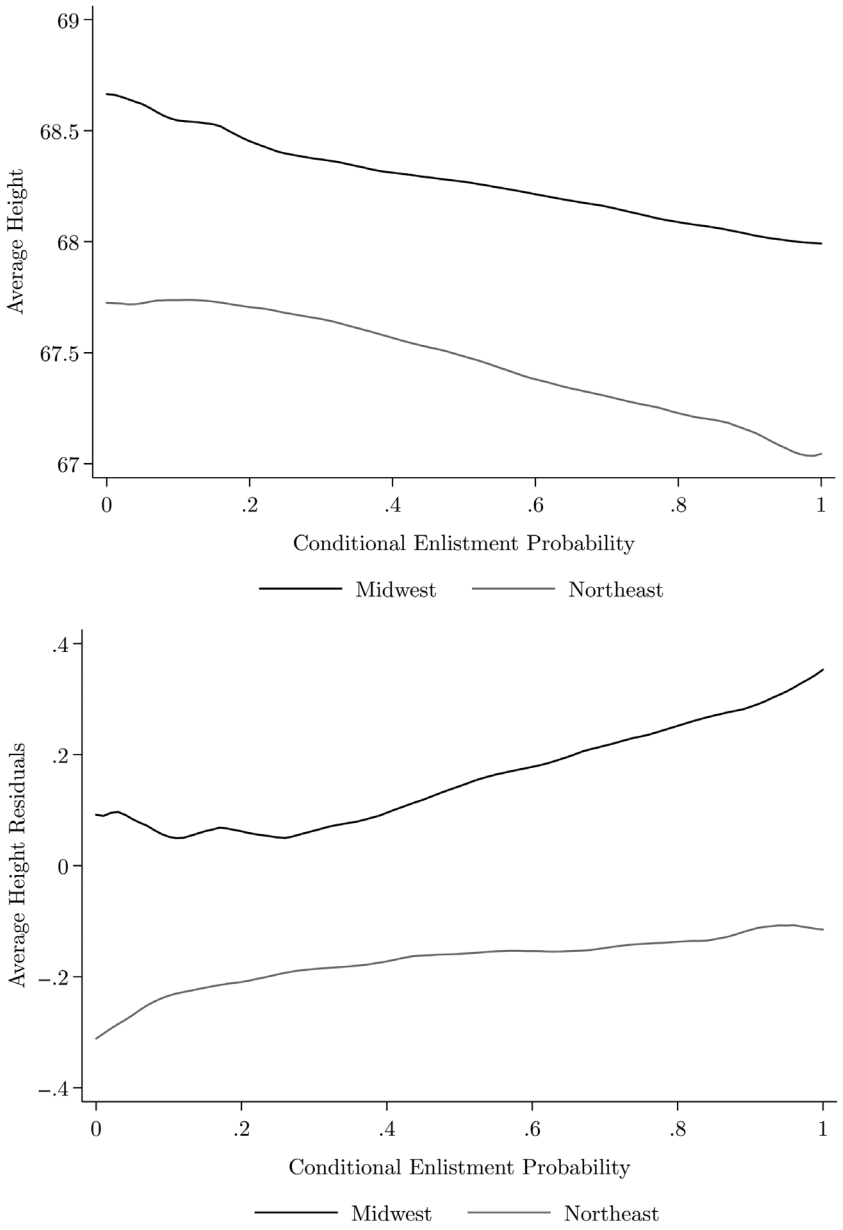

Figure 5. Height and residuals by region and estimated enlistment probability.
Note: The *x*-axis is the enlistment probability estimated in column 3 of table 3. The *y*-axis in the top panel is height. The *y*-axis in the bottom panel is residuals of height from a regression of heights on all variables in column 5 of table 3 except for state fixed effects. Each graph presents results of a separate nonparametric regression for each region.

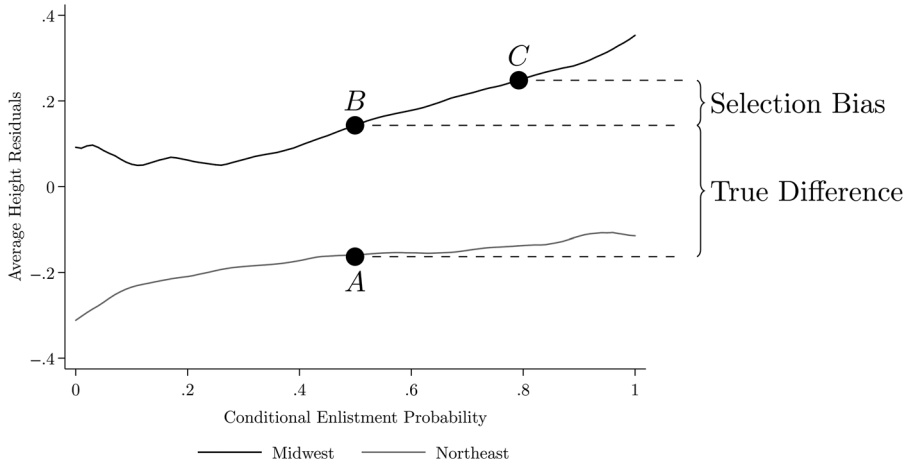

Figure 6. Residuals of height and enlistment probability, annotated.
Note: This figure repeats the bottom panel of figure 5 but is annotated to show the danger of comparing individuals with different enlistment probabilities instead of making comparisons only for individuals with the same enlistment probabilities. Points A and B correspond to the modal enlistment probability of Northeasterners, while point C corresponds to the modal enlistment probability of Midwesterners, each from the bottom panel of figure 4.

average heights of the Northeast and the Midwest. It repeats the bottom panel of figure 5, but indicates approximately the points at which the bulk of Northeasterners and Midwesterners are located in the distribution of enlistment probability, as shown in the bottom panel of figure 4—points *A* and *C*, respectively. The effect of selection on unobservables on estimation of the height difference between the regions can be illustrated by comparing these points. Point *A* is (loosely) the average observed height of Northeasterners, while point *C* is (again loosely) the average observed height of Midwesterners. A comparison of these two points yields the Midwest's observed height advantage. But this comparison conflates two differences—the true Northeast–Midwest difference and the difference in sample-selection bias between the regions, which is greater for the Northeast at point *A* than for the Midwest at point *C* because of the greater enlistment probability of the Midwest. A better comparison would be of points *A* and *B*, which compares individuals with the same enlistment probability, and thus the same degree of sample-selection bias. More generally, rather than computing the difference in heights between the Midwest and the Northeast using the distribution of enlistment probabilities in the data (figure 4), a correct comparison would be a weighted average of differences between individuals across regions with the same enlistment probability.

On the whole, then, the patterns of selection on unobservables revealed by this analysis suggest that the Midwest–Northeast height premium is likely overstated, but that there truly was a premium. This is consistent with the conclusions of Zimran (2019).

Conclusion

Sample-selection bias generated by selection on observables and selection on unobservables poses a central challenge to the use of historical data to draw conclusions about broader populations of interest. Though this issue arises throughout social science history, it has recently been especially salient in anthropometric history, where a new literature (e.g., Bodenhorn et al. 2017; Zimran 2019) has focused on understanding how sample-selection bias might affect inference from historical data.

This article develops a simple theoretical example to identify five patterns that sample-selection bias creates in a potentially selected sample. It then uses these patterns to motivate and execute some empirical exercises that are informative regarding the potential presence and impact of sample-selection bias in a sample of military stature from the antebellum United States, especially on the determination of the Northeast–Midwest height difference from these data. These exercises are simple and intuitively grounded, and can be applied in other empirical settings to guide social science historians in their engagement with sources whose use might be confounded by the presence of sample-selection bias.

The insight that can be gained from these exercises increases in the data available to the researcher. With the potentially selected sample alone, it is not possible to determine whether any observed patterns are true population patterns or the product of sample-selection bias. But if the researcher is able to determine whether certain groups are over- or underrepresented in the sample relative to the population (perhaps from external data on population shares), it is possible to use pattern 2 to suggest whether concern over sample-selection bias is in order. An excluded variable enables the researcher to gain insights from patterns 3 and 4. But if only the potentially selected sample is available, the researcher must make assumptions about whether and how the excluded variable affects entry into the sample. Finally, the strongest conclusions are possible if the researcher also has access to a supplemental sample describing the observable characteristics for the population of interest. Such a data set enables the researcher to formally test whether and how the excluded variable affects entry into the sample and to compute conditional selection probabilities.

It is important to emphasize that these exercises are not a substitute for a direct and formal correction as performed by Zimran (2019) on the basis of Heckman's (1979) method. The goal of this article is instead to develop a better understanding of what it is that this method does, and to provide scholars with a simple, but incomplete and informal, method to check for the presence and likely impact of sample-selection bias and to decide on this basis whether a formal correction is necessary.

It is also important to note that regardless of how researchers confront problems of bias in their data, no statistical exercise is a substitute for serious consideration of the limitations of a data source. Even if the exercises proposed in this article reveal no evidence of sample-selection bias affecting conclusions, ultimately the exercises are able to go only as far as statistical and economic theory allow. As Bodenhorn et al. (2017) argue, data sources created by voluntary choice and the conclusions that they produce must always be confronted with skepticism.

These exercises are also useful in cases other than trying to determine whether conclusions are affected by sample-selection bias. For instance, selection on unobservables may be an outcome of interest in some cases, such as in Ferrie's (1997) and Stewart's

(2006) studies of migration to the frontier in the nineteenth-century United States. In such cases, although the role played by sample selection is different, the intuition to recognize its presence and to understand its role, and the possible exercises that can be used to uncover it, is the same as in the case discussed in this article.[36]

Acknowledgments. I am grateful to William Collins for comments on several drafts of this paper; to Ran Abramitzky, Richard Steckel, and Marlous van Waijenburg for helpful discussions; to an anonymous reader for comments; and to the editors Anne McCants, Kris Inwood, Hamish Maxwell-Stewart, and Ewout Depauw. Thanks are also due to Timothy Cuff for sharing data on Pennsylvania recruits to the Union Army. This project, by virtue of its use of the Union Army Project data, was supported by Award Number P01 AG10120 from the National Institute on Aging. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health. This paper previously circulated under the title "Intuition to Recognize and Address Sample-Selection Bias in Historical Sources, with Illustrations from the Historical Heights Literature."

References

Abramitzky, Ran (2015) "Economics and the modern economic historian." Journal of Economic History **75** (4): 1240–51.

Biavaschi, Costanza, Corrado Guilietti, and Zahra Siddique (2017) "The economic payoff of name Americanization." Journal of Labor Economics **35** (4): 1089–116.

Bodenhorn, Howard, Timothy W. Guinnane, and Thomas A. Mroz (2017) "Sample-selection biases and the industrialization puzzle." Journal of Economic History **77** (1): 171–207.

—— (2019) "Diagnosing sample-selection bias in historical heights: A reply to Komlos and A'Hearn." Journal of Economic History **79** (4): 1154–75.

Bushway, Shawn, Brian D. Johnson, and Lee Ann Slocum (2007) "Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology." Journal of Quantitative Criminology **23** (2): 151–78.

Coffman, Edward M. (1986) The Old Army: A Portrait of the American Army in Peacetime, 1784–1898. New York: Oxford University Press.

Collins, William J. (2015) "Looking forward: Positive and normative views of economic history's future." Journal of Economic History **75** (4): 1228–33.

Cosslett, Stephen R. (1981) "Efficient estimation of discrete-choice models," in Charles F. Manski and Daniel McFadden (eds.) Structural Analysis of Discrete Data with Econometric Applications. Cambridge, MA: MIT Press: 51–111.

Costa, Dora L., and Matthew E. Kahn (2003) "Cowards and heroes: Group loyalty in the American Civil War." Quarterly Journal of Economics **118** (2): 519–48.

—— (2007) "Deserters, social norms, and migration." Journal of Law and Economics **50** (2): 323–53.

Cuff, Timothy (2005) The Hidden Cost of Economic Development: The Biological Standard of Living in Antebellum Pennsylvania. Burlington, VT: Ashgate.

Easterlin, Richard (1960) "Interregional differences in per capita income, population, and total income, 1840–1950," in National Bureau of Economic Research Council (eds.) Trends in the American Economy in the Nineteenth Century. Conference on Research in Income and Wealth. Princeton, NJ: Princeton University Press: 73–140.

Eli, Shari, Laura Salisbury, and Allison Shertzer (2018) "Ideology and migration after the American Civil War." Journal of Economic History **78** (3): 822–61.

Ferrie, Joseph P. (1997) "Migration to the frontier in mid-nineteenth century America: A re-examination of Turner's 'safety valve.'" Mimeograph, Northwestern University.

[36]In the case of these studies, the outcome of interest is wealth accumulation, and the selection issue is the choice of where to live (on the frontier or not). The selection problem is that, for example, frontier wealth accumulation is observed only for those who chose to move to the frontier and not for the whole population.

Floud, Roderick, Robert W. Fogel, Bernard Harris, and Sok Chul Hong (2011) The Changing Body: Health, Nutrition, and Human Development in the Western World since 1700. New York: Cambridge University Press.

Fogel, Robert W. (1986) "Nutrition and the decline in mortality since 1700: Some preliminary findings," in Stanley L. Engerman and Robert E. Gallman (eds.) Long-Term Factors in American Economic Growth. Chicago: University of Chicago Press: 439–556.

Fogel, Robert W., Dora L. Costa, Michael R. Haines, Chulhee Lee, Louis Nguyen, Clayne Pope, Irvin Rosenberg, Nevin Scrimshaw, James Trussell, Sven Wilson, Larry T. Wimmer, John Kim, Julene Bassett, Joseph Burton, and Noelle Yetter (2000) Aging of Veterans of the Union Army: Version M-5. Chicago: Center for Population Economics, University of Chicago Graduate School of Business, Department of Economics, Brigham Young University, and the National Bureau of Economic Research.

Fogel, Robert W., and Stanley L. Engerman (1974) Time on the Cross: The Economics of American Negro Slavery. Boston: Little, Brown, and Co.

Fogel, Robert W., Stanley L. Engerman, Roderick Floud, Gerald Friedman, Robert A. Margo, Kenneth Sokoloff, Richard H. Steckel, T. James Trussell, Georgia Villaflor, and Kenneth W. Wachter (1983) "Secular changes in American and British Stature and Nutrition." Journal of Interdisciplinary History **14** (2): 445–81.

Foner, Jack D. (1970) The United States Soldier between Two Wars: Army Life and Reforms, 1865–1898. New York: Humanities Press.

Gallman, Robert E. (1996) "Dietary change in antebellum America." Journal of Economic History **56** (1): 193–201.

Gould, Benjamin Apthorp (1869) Investigations in the Military and Anthropological Statistics of American Soldiers. Sanitary Memoirs of the War of the Rebellion. Collected and Published by the United States Sanitary Commission. New York: Hurd and Houghton.

Heckman, James J. (1979) "Sample selection bias as a specification error." Econometrica **47** (1): 153–61.

ICPSR (1999) United States Historical Election Returns, 1824–1968 (ICPSR 1) [machine-readable database]. Ann Arbor, MI.

Komlos, John (2004) "How to (and how not to) analyze deficient height samples: An introduction." Historical Methods **37** (4): 160–73.

—— (2012) "A three-decade history of the antebellum puzzle: Explaining the shrinking of the US Population at the Onset of Modern Economic Growth." Journal of the Historical Society **12** (4): 395–445.

—— (2019) "Shrinking in a growing economy is not so puzzling after all." Economics and Human Biology **32**: 40–55.

—— (2020) "Multicollinearity in the presence of errors-in-variables can increase the probability of type-I error." Journal of Economics and Econometrics **63** (1): 1–17.

Komlos, John, and Brian A'Hearn (2019) "Clarifications of a puzzle: The decline in nutritional status at the onset of modern economic growth in the United States." Journal of Economic History **79** (4): 1129–53.

Kosack, Edward, and Zachary Ward (2014) "Who crossed the border? Self-selection of Mexican migrants in the early twentieth century." Journal of Economic History **74** (4): 1015–44.

Logan, Trevon D., and Jonathan B. Pritchett (2018) "On the marital status of US slaves: Evidence from Touro Infirmary, New Orleans, Louisiana." Explorations in Economic History **69**: 50–63.

Manson, Steven, Jonathan Schroeder, David Van Riper, and Steven Ruggles (2017) IPUMS National Historical Geographic Information System: Version 12.0 [Database]. Minneapolis: University of Minnesota.

Margo, Robert A., and Richard H. Steckel (1983) "Heights of native-born whites during the antebellum period." Journal of Economic History **43** (1): 167–74.

McKeown, Thomas (1976) The Modern Rise of Population. London: Arnold.

Mitch, David (1993) "The role of human capital in the first Industrial Revolution," in Joel Mokyr (ed.) The British Industrial Revolution: An Economic Perspective. Boulder, CO: Westview: 267–307.

Mokyr, Joel, and Cormac Ó Gráda (1996) "Height and health in the United Kingdom 1815–1860: Evidence from the East India Company Army." Explorations in Economic History **33**: 141–68.

Records of the Adjutant General's Office (1861–1865) Regimental records, including descriptive rolls, order, and letter books, and morning reports, of volunteer organizations, Civil War, 1861–65. Records relating to volunteers and volunteer organizations. Record Group 94.2.4. Washington, DC: National Archives Building.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek (2015) Integrated Public Use Microdata Series: Version 6.0 [machine-readable database]. Minneapolis: University of Minnesota.

Steckel, Richard H., and Nicolas Ziebarth (2016) "Selectivity and measured catch-up growth of American Slaves." Journal of Economic History **76** (1): 104–38.

Stewart, James I. (2006) "Migration to the agricultural frontier and wealth accumulation." Explorations in Economic History **43**: 547–77.

Vella, Francis (1998) "Estimating models with sample selection bias: A survey." Journal of Human Resources **33** (1): 127–69.

Weigley, Russell F. (1967) History of the United States Army. New York: The Macmillan Company.

Zehetmayer, Matthias (2011) "The continuation of the antebellum puzzle: Stature in the US, 1847–1894." European Review of Economic History **15**: 313–27.

Zimran, Ariell (2018) "Replication: Sample-selection bias and height trends in the nineteenth-century United States." Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor]. http://doi.org/10.3886/E107742V1

—— (2019) "Sample-selection bias and height trends in the nineteenth-century United States." Journal of Economic History **79** (1): 99–138.

—— (2020) "Transportation and health in the antebellum United States, 1820–1847." Journal of Economic History, Forthcoming.

Cite this article: Zimran, Ariell (2020) "Recognizing Sample-Selection Bias in Historical Data," *Social Science History* **44**:525–554. doi:10.1017/ssh.2020.11