A Machine Learning Approach to Improving Occupational Income Scores*

Martin Saavedra
Department of Economics
Oberlin College

Tate Twinam
Interdisciplinary Arts and Sciences
University of Washington Bothell

February 3, 2018

Abstract

Historical studies of labor markets frequently suffer from a lack of data on individual income. The occupational income score (OCCSCORE) is often used as an alternative measure of labor market outcomes. Using modern Census data, we find that the use of OCCSCORE biases results towards zero and can frequently result in statistically significant coefficients of the wrong sign. We use a machine learning approach to construct a new adjusted score based on industry, occupation, and individual demographics. Our alternative score substantially outperforms OCCSCORE in both modern and historical contexts. We illustrate our approach by estimating racial and gender earnings gaps in the 1915 Iowa State Census and intergenerational mobility elasticities using linked data from the 1850-1910 Censuses.

JEL codes: C21, J71, N32
Keywords: OCCSCORE, occupational income score, LIDO score, machine learning, lasso, non-classical measurement error, occupation, earnings gaps

---

*We are grateful for feedback from Brian Beach, Hoyt Bleakley, Melissa Dell, James Feigenbaum, Ekaterina Jardim, Bruce Sacerdote, James Siodla, Randall Walsh, Mark Watson, and Eugene White, as well as seminar participants at the University of Michigan, Harvard University, the College of William and Mary, and Marquette University. We are also grateful for comments from participants at the following conferences: WEAI, the Northeast Ohio Economics Workshop, SEA, and the Pacific Northwest Labor Day conference at UW Seattle. Grant Goehring provided excellent research assistance. All errors are our own.

1 Introduction

Before 1940, data on individual wages and education are not available in the U.S. Census. Consequently, occupation is often the only measure of labor market outcomes available to economic historians. Occupation is a categorical variable; however, many economists use an index of occupational earnings potential as a continuous measure of historical labor market outcomes. One popular example is the 1950 occupational income score (OCCSCORE), which is the median income of an occupation in 1950. Occupational income scores have been used to examine earnings gaps going as far back at 1850, and studies using this approach have been published in numerous top journals in economics and other fields.[1]

Although occupational income scores should be correlated with earnings (for example, physicians and lawyers have higher occupational income scores than laborers), they are obviously an imperfect proxy for true earnings, and it is unclear how much bias this measurement error induces. Additionally, it is unclear if 1950 occupational income scores are good measures of labor market outcomes when examining Censuses several decades before 1950. While this potential bias has been acknowledged in the literature, no attempt has been made to quantify it and diagnose its impact on inferences. In this study, we attempt to measure this bias directly and examine how much it can be mitigated through adjustments to occupational income scores based on demographic and geographic variables universally available to economic historians.

We first develop a formal model of the measurement error problem posed by occupational income scores. The model allows us to determine when attenuation bias will occur and to explicitly quantify its magnitude. We then take this model to the data to estimate the OCCSCORE-induced bias. Because it is difficult to make historical data better, we analyze the performance of occupational income scores by making modern data worse.[2] We generate

[1] See section 2 for examples.

[2] Our approach is in the spirit of Romer (1986), who shows that excess volatility in unemployment time series during the pre-war era is an artifact of the interpolation methods used before the Current Population Survey. Applying the same interpolation methods to unemployment data during the post-war period results in similar levels of volatility.

2000-based occupational income scores and examine how well they predict income in the decades between 1950 and 2000. We then use this index to examine racial and gender earnings gaps from 1950 through 2000 and compare these to the true gaps estimated using actual earnings data. Finally, we use cross-validated lasso regressions to construct new adjusted OCCSCOREs (based on industry, occupation, race, sex, age, state, and region). We compare estimated earnings gaps based on our lasso-adjusted industry, demographic, and occupation (LIDO) scores to those generated using OCCSCORE and true earnings.

We find that although OCCSCORE is correlated with income even for Censuses five decades removed from the base year, racial and gender earnings gaps are significantly attenuated when using OCCSCORE as a proxy for income. The use of OCCSCORE can result in statistically significant coefficients of the wrong sign up to 20 percent of the time in our modern data, particularly for variables indicating state of residency (often used in difference-in-differences analysis exploiting state-level variation in treatments).[3] This is even the case in earnings regressions where the sample is restricted to white males only. We find that adjusting OCCSCORE by race, sex, age, industry, and geography–adjustments rarely made in empirical work–substantially reduces this bias.

To examine the performance of our LIDO scores in a historical context, we exploit a rare source of pre-1950 earnings data: the 1915 Iowa State Census.[4] Estimated race and gender earnings gaps in 1915 Iowa using true earnings data are sizable and negative; however, when using standard OCCSCORE as a proxy for earnings, the racial earnings gap is underestimated by half and the gender earnings gap is attenuated by 95%. Our adjusted OCCSCORE yields race and gender earnings gaps very close to the true values. Finally, we conduct an analysis of OCCSCORE–induced bias in measures of intergenerational income transmission. This analysis is based on father–son pairs linked from the 1880 decennial

[3]This finding is similar to that in Bertrand, Duflo and Mullainathan (2004) who find that difference-in-differences models that do not account for serial correlation in the error terms can result in statistically significant estimates of placebo treatment effects 40% of the time.

[4]The 1915 Iowa State Census data was digitized by Claudia Goldin and Lawrence Katz (Goldin and Katz, 2010).

Census to the 1850, 1860, 1900, and 1910 decennial Censuses. In this setting, we find that standard OCCSCOREs and our alternative scores perform similarly for white males, because measurement errors for fathers and sons are likely to be correlated. However, transmission coefficients are substantially attenuated for black men, suggesting a much higher rate of intergenerational mobility than is likely true. We conclude with recommendations for future research in economic history.

2 Previous Literature

To understand how researchers use the occupational income score variable, we searched for papers containing either "OCCSCORE" or "Occupational Income Score" in top general interest journals and top field journals in labor economics and economic history. This search yielded the 25 papers listed in Table 1.[5] Most of the articles have been published within the last decade, with a median publication year of 2014. Sixteen use the log of occupational income score as a dependent variable, and consequently, we focus our empirical analysis on the log of occupational income score. Of these 25 papers, only four adjust occupational income scores by any demographic variables.[6] Typically, these papers analyze historical Census data for which income or wage data are not available and interpret occupational income score as a proxy for income. Some of the papers, however, present occupational income score along with wage/income data as an alternative measure of socioeconomic status (see Stephens and Yang (2014) or Chin (2005)). Some papers attempt to reduce the bias by limiting the sample to a particular demographic group, typically white males (see Bleakley (2010)). These papers often examine intergenerational mobility, racial and ethnic SES gaps, migrant selection, and the effects of schooling or health interventions.

[5]This search included articles in the following journals: *American Economic Journal: Applied Economics*; *American Economic Journal: Economic Policy*; *American Economic Review*; *Explorations in Economic History*; *Journal of Economic History*; *Journal of Human Resources*; *Journal of Labor Economics*; *Quarterly Journal of Economics*; *Review of Economics and Statistics*, and *The Review of Economic Studies.*

[6]The occupational earnings measure used by Collins and Wanamaker (2014) varies by race and region; Angrist (2002) varies by age and sex; and Collins (2000) presents results using both an unadjusted and a

Table 1: Published Studies using Occupational Income Scores

Article	Description	Adjusted	Log
Collins (2000)	Occupational mobility of blacks during the 1940s.	Both	Yes
Minns (2000)	SES growth of immigrants relative to natives in the 1900 and 1910 Censuses.	No	Yes
Angrist (2002)	Effect of sex ratios on marriage markets and labor-force participation.	Yes	Yes
Chin (2005)	Long-run effects of adult incarceration in internment camps on labor-market outcomes.	No	No
Sacerdote (2005)	Intergenerational effects of slavery.	No	No
Bleakley (2007)	Effect of hookworm eradication during school-age on human capital.	No	Yes
Bleakley and Lange (2009)	Quantity-quality childbearing model using the eradication of hookworm as an exogenous shock to the returns for schooling.	No	Yes
Bleakley (2010)	Childhood exposure to malaria and adult SES in Brazil, Colombia, Mexico, and the U.S.	No	Yes
Abramitzky et al. (2012)	Returns to migration and self-selection for Norwegian-U.S. migrants.	No	Yes
Lee (2013)	Repeal of Sunday closing laws affected year of schooling and adult outcomes.	No	Yes
Aaronson et al. (2014)	The Rosenwald School initiative and quality-quantity childbearing model.	No	Yes
Collins and Wanamaker (2014)	The returns to migration and self-selection for blacks during the Great Migration.	Yes	Yes
Cook et al. (2014)	Distinctively black names and socioeconomic status.	No	No
Stephens and Yang (2014)	Sensitivity of prior estimates for the returns to schooling to region-specific birth year effects.	No	No
Collins and Wanamaker (2015)	Self-selection of inter-regional and intra-regional migration.	No	Yes
Lleras-Muney and Shertzer (2015)	The effect of English-only statutes on immigrant children literacy, years of schooling, and occupations.	No	Yes
Olivetti and Paserman (2015)	Creates pseudo-links to estimate father-son and father-daughter elasticities for the intergenerational transmission of SES.	No	Yes
Saavedra (2015)	The effect of school-age incarceration in internment camps on adult outcomes.	No	No
Cook et al. (2016)	19th century blacks with distinctively black names live longer.	No	No
Massey (2016)	U.S. immigrant quota affected the selection of immigrants.	No	Yes
Bleakley and Ferrie (2016)	The effects of a Georgia land lottery on human capital investment. Uses OCCSCORE to measure returns to literacy.	No	No
Lee and Lin (2017)	How natural amenities affect neighborhood income ranks.	No	No
Saavedra (2017)	Early-life exposure to yellow fever affected occupational status.	No	No
Ward (2017)	Self-selection of return migrants.	No	Yes
Carruthers and Wanamaker (2017)	Effect of differential school quality to the black-white income gap.	Yes	Yes

This list underestimates how many researchers use occupational income scores or similar measures. Other papers in these journals may have used average or median income/wages by occupation as a dependent variable but do not refer to the variable as an occupational income score.[7] Occupational income scores are also used in other fields, especially sociology. A Google Scholar search for research articles containing "OCCSCORE" or "occupational income score" currently yields 273 articles. Many of these articles are working papers that will eventually be published in top economics journals. For example, four NBER working papers in 2016 alone contain the phrase "OCCSCORE" or "Occupational Income Score."

3 Econometric Model

3.1 Measurement Error

Consider a linear model with classical measurement error (CME) in the dependent variable. The researcher is interested in $y_i = \alpha + \beta X_i + \epsilon_i$, but instead of observing y_i, the researcher observes $\tilde{y}_i$, which equals y_i plus a measurement error term e_i. In the CME model, y_i and X_i are uncorrelated with e_i, implying that e_i is by definition correlated with the observed value $\tilde{y}_i$. Thus, a regression of the mis-measured $\tilde{y}_i$ on X_i is equivalent to a regression of y_i on X_i with an error term of $\epsilon_i - e_i$. Since e_i is uncorrelated with the true y_i and X_i, regressing the observed value on X_i is equivalent to adding variance to the error term of the regression. For this reason, CME in the dependent variable affects the precision of the regression estimates, but does not lead to bias. This is likely part of the reason that researchers either ignore, or give cursory attention to, measurement error in the dependent variable.

Unfortunately, the CME model is a poor description of occupational income scores. Suppose the true y_i is income. Without knowledge of y_i, the researcher replaces y_i with their best guess of income given occupation. Perhaps this measure is mean or median earnings for

race-adjusted OCCSCORE.

[7] For example, Bailey and Collins (2006) construct average occupational wages across sex-race-industry-region cells in 1940 to analyze the wage gains of black women between 1910 and 1940.

a given occupation. Since the reported y_i is the researcher's best guess of income, the measurement error must be uncorrelated with the reported value, and by definition correlated with the true value. Thus, the opposite of the CME model holds.

Occupational income scores are better described using the Optimal Prediction Error (OPE) model of Hyslop and Imbens (2001). In the OPE model, if researchers use their best guess of y_i (income) given a noisy signal (occupation), then estimates of β are in general biased towards zero. Hyslop and Imbens (2001) refer to this as OPE(1). However, if researchers instead use their best guess of y_i given both the noisy signal and X_i (referred to as an OPE(2) model), then estimates of β are unbiased. Consequently, instead of using occupational income scores, researchers should develop occupational income scores that are conditional on X_i.

In the subsection below, we develop a modified OPE model specifically for occupational income scores. In the model, researchers observe occupation, a vector of relevant covariates X_i, and, from a separate data source, mean earnings by occupation. Changes in X_i can affect earnings through two channels: either by shifting individuals from lower to higher paying occupations, by increasing earnings within a given occupation, or both. We then find conditions equivalent to the OPE(1) and OPE(2) from Hyslop and Imbens (2001). Lastly, it is not always possible to predict income conditional on both occupation and X_i, particularly when X_i is not a demographic variable. We show that so long as demographics are correlated with X_i, then using a demographically-adjusted score will result in estimates that are less biased than the unadjusted occupational income score. We refer to this model as an OPE(3) model.

3.2 An Optimal Prediction Error Model with Occupational Income Scores

A researcher is interested in $y_i = \alpha + \beta X_i + \epsilon_i$, where y_i is the income of individual i, X_i is the policy variable of interest, and $\epsilon_i \perp\!\!\!\perp X_i$. The researcher does not observe y_i, but observes both X_i and occupation j. From a separate source, the researcher also observes the occupational income score of occupation j, which is $\mathrm{E}(y|\mathrm{occ} = j)$.[8] The policy variable could increase income through two channels, either shifting the marginal worker into higher paying occupations or increasing the earnings of workers within a given occupation.

We model the first process by assuming that there are a continuum of occupations and let O_j be a measure of the earnings potential of occupation j. Specifically, define O_j to be the mean income that a random individual would receive if they were to enter occupation j. Note that $\mathrm{E}(y|\mathrm{occ} = j) = \mathrm{E}(y|O_j) \neq O_j$, because high-skilled individuals are more likely to enter high-paying occupations. Occupations are determined through the following data generating process:

$$O_j = \delta_0 + \delta_1 X_i + \eta_i. \tag{1}$$

The parameter δ_1 captures the fact that X_i shifts workers across occupations. The nuisance term η_i captures that some workers enter occupations with higher or lower earnings potential than their X_i would predict, perhaps because of preferences, ability, or luck.

We then model excess earnings within occupation as a separate process. Let the excess earnings of individual i be

$$y_i - O_j = \gamma_0 + \gamma_1 X_i + \nu_i. \tag{2}$$

For each worker i in occupation j, total earnings y_i equals their occupational earnings potential O_j plus their excess earnings. Thus, $\delta_0 + \gamma_0 = \alpha$, $\delta_1 + \gamma_1 = \beta$, and $\eta_i + \nu_i = \epsilon_i$. Further assume that $\eta_i, \nu_i \perp\!\!\!\perp X_i$ and $\eta_i \perp\!\!\!\perp \nu_i$. That X_i is independent of the error terms η_i and ν_i is not a strong assumption given that many econometric models assume that X_i is independent of ϵ_i. The key assumption that separates this model from standard earning regressions models is that the error term affecting occupational sorting (η_i) does not interact with the error term determining excess earnings within an occupation (ν_i).

Lastly, we assume that γ_1 and δ_1 have the same sign. This implies that if X_i increases

[8]The IPUMS OCCSCORE uses median earnings by occupation instead of mean earnings. We use mean earnings to simplify the model.

income across occupations, it also increases income within occupations. This assumption will almost certainly hold for demographic groups that have historically earned less in labor markets (such as women and racial minorities). This assumption likely holds for human capital interventions that increase ability. It is conceivable that there are cases for which γ_1 and δ_1 have opposite signs. For example, suppose an intervention increased the probability that a college graduate continues to law school, but had no other labor market consequences. This intervention would likely increase income across occupations ($\delta_1 > 0$), but not within occupation ($\gamma_1 = 0$) since almost all lawyers have law degrees.

3.2.1 OPE(1): $DV = \mathrm{E}(y|\mathbf{occ} = j)$

Since y_i is not observable, economic historians often use mean earnings of an occupation as the dependent variable in regressions. In this case,

$$\operatorname{plim} \hat{\beta} = \frac{\mathrm{Cov}(\mathrm{E}(y_i|O_j), X_i)}{\mathrm{Var}(X_i)} = \delta_1 + \gamma_1 \frac{\mathrm{Cov}(\mathrm{E}(X_i|O_j), X_i)}{\mathrm{Var}(X_i)}. \tag{3}$$

Since $\frac{O_j - \delta_0}{\delta_1} = X_i + \frac{\eta_i}{\delta_1}$, we get that

$$\mathrm{E}(X_i|O_j) = \phi_0 \mu_X + \phi_1 \frac{O_j - \delta_0}{\delta_1} \tag{4}$$

where

$$\phi_0 = \frac{\frac{1}{\sigma_X^2}}{\frac{1}{\sigma_X^2} + \frac{\delta_1^2}{\sigma_\eta^2}}, \tag{5}$$

$$\phi_1 = \frac{\frac{\delta_1^2}{\sigma_\eta^2}}{\frac{1}{\sigma_X^2} + \frac{\delta_1^2}{\sigma_\eta^2}}. \tag{6}$$

Some algebra will show that

$$\operatorname{plim} \hat{\beta} = \delta_1 + \gamma_1 \phi_1. \tag{7}$$

Since $\phi_1 \in (0,1)$, $\hat{\beta}$ is biased towards zero.

3.2.2 OPE(2): $DV = \mathrm{E}(y_i|\mathbf{occ} = j, X_i)$

Suppose the researcher from another source can observe $\mathrm{E}(y_i|\text{occ} = j, X_i)$ and uses this as the dependent variable. This is analogous to using a demographically adjusted occupational income score (such as the LIDO score, which we develop below) when the policy variable of interest is one the demographic variables used to construct the adjustment (e.g., a gender and race dummy).

Then we get that

$$\hat{\beta} = \frac{\mathrm{Cov}(\mathrm{E}(y_i|O_j, X_i), X_i)}{\mathrm{Var}(X_i)} \tag{8}$$

$$= \frac{\mathrm{Cov}(\alpha + \beta X_i + \mathrm{E}(\epsilon|O_j, X_i), X_i)}{\mathrm{Var}(X_i)} \tag{9}$$

$$= \beta\frac{\mathrm{Cov}(X_i, X_i)}{\mathrm{Var}(X_i)} = \beta. \tag{10}$$

Thus, $\hat{\beta}$ is unbiased.

3.2.3 OPE(3) $DV = \mathrm{E}(y_i|\mathbf{occ} = j, Z_i)$

Assume that the researcher is interested in $y_i = \alpha + \beta X_i + \epsilon_i$, but she cannot observe $\mathrm{E}(y_i|O_j, X_i)$. However, suppose that the policy variable X_i is correlated with a demographic variable Z_i and the researcher does observe $\mathrm{E}(y_i|\text{occ} = j, Z_i)$. Further assume that Z_i is correlated with X_i so that $Z_i = \lambda_0 + \lambda_1 X_i + \psi_i$, where ψ_i is independent of X_i and η_i. For example, the researcher may be interested in early-life malaria exposure during the late nineteenth or early twentieth century. Because the LIDO score, or any other index of occupational earnings, does not take into account early-life malaria exposure, the researcher does not observe $\mathrm{E}(y_i|\text{occ} = j, X_i)$. But the LIDO score does take into account geographic variables such as state of residency that will be correlated with early-life malaria exposure.

A regression of $\mathrm{E}(y_i|O_j, Z_i)$ on X_i gives us estimates

$$\operatorname{plim}\hat{\beta} = \frac{\mathrm{Cov}(\mathrm{E}(y_i|O_j, Z_i), X_i)}{\mathrm{Var}(X_i)} = \delta_1 + \gamma_1 \frac{\mathrm{Cov}(\mathrm{E}(X_i|O_j, Z_i), X_i)}{\mathrm{Var}(X_i)}. \tag{11}$$

Given O_j and Z_i, we have two noisy measures of X_i since $\frac{O_j - \delta_0}{\delta_1} = X_i + \frac{\eta_i}{\delta_1}$ and $\frac{Z_i - \lambda_0}{\lambda_1} = X_i + \frac{\psi_i}{\lambda_1}$. Therefore, we get that

$$\mathrm{E}(X_i|O_j, Z_i) = \theta_0 \mu_X + \theta_1 \frac{O_j - \delta_0}{\delta_1} + \theta_2 \frac{Z_i - \lambda_0}{\lambda_1} \tag{12}$$

where

$$\theta_0 = \frac{\frac{1}{\sigma_X^2}}{\frac{1}{\sigma_X^2} + \frac{\delta_1^2}{\sigma_\eta^2} + \frac{\lambda_1^2}{\sigma_\psi^2}} \tag{13}$$

$$\theta_1 = \frac{\frac{\delta_1^2}{\sigma_\eta^2}}{\frac{1}{\sigma_X^2} + \frac{\delta_1^2}{\sigma_\eta^2} + \frac{\lambda_1^2}{\sigma_\psi^2}} \tag{14}$$

$$\theta_2 = \frac{\frac{\lambda_1^2}{\sigma_\psi^2}}{\frac{1}{\sigma_X^2} + \frac{\delta_1^2}{\sigma_\eta^2} + \frac{\lambda_1^2}{\sigma_\psi^2}} \tag{15}$$

Some algebra will then reveal that

$$\operatorname{plim}\hat{\beta} = \delta_1 + \gamma_1 \frac{\frac{\delta_1^2}{\sigma_\eta^2} + \frac{\lambda_1^2}{\sigma_\psi^2}}{\frac{1}{\sigma_X^2} + \frac{\delta_1^2}{\sigma_\eta^2} + \frac{\lambda_1^2}{\sigma_\psi^2}}. \tag{16}$$

It follows that $\hat{\beta}$ is biased toward zero, but less biased than the traditional occupational income score. As σ_ψ^2 goes to zero, then Z_i becomes collinear with X_i and the bias goes to zero.

3.3 Constructing an Adjusted Occupational Income Score

The above model suggests that a more useful occupational income score could be derived from average incomes conditioned on both occupation and a suite of relevant demographic variables. There are a number of ways to approach this problem. The most important commonly-available variables influencing labor market outcomes are industry, occupation, sex, race/ethnicity, age, and geographic location.[9] Any demographically-adjusted OCCSCORE should account for differences along these lines.

The simplest and most general approach would be to adjust OCCSCORE in a fully nonparametric manner. For example, one could take an individual's OCCSCORE to be the median (or mean) income in a given base year for that individual's occupation within cells defined by their sex, age, race, and region.[10] The advantage of this measure is that it allows for arbitrary interactions between all of the adjustment variables. For example, the age-earnings profile may differ flexibly between men and women in a given occupation, or the wage gap between races in a given occupation may vary between regions. The disadvantage of this approach is that stratifying on so many variables may result in small or empty cells, leading to excessively variable or missing OCCSCOREs for many individuals. Constructing new OCCSCOREs based on the (relatively small) 1% sample of the 1950 Census exacerbates this problem.

An alternative that avoids this problem involves a less flexible parametric approach. For example, one could regress income in a given base year on a series of occupation, sex, age, race, and geographic state indicator variables. The fitted coefficients could then be used to generate an adjusted OCCSCORE for each possible individual. This strategy is

[9] Another common demographic variable that could be used is an indicator for foreign–born status. However, we caution that this may be misleading. The composition of the foreign–born population in the U.S. in 1950 differs substantially from that in earlier years such as 1900 and 1850 (in both racial/ethnic makeup and human capital terms). Thus, adjusting on this variable in a given base year may lead to inaccurate results when applied to other years.

[10] This is in the spirit of the adjustments made by some previous authors. For example, Angrist (2002) constructs age- and sex-specific OCCSCOREs based on median income within cells. Collins and Wanamaker (2014) compute income scores by occupation and region specifically for black men.

computationally simple and generates an adjusted OCCSCORE for every type of individual. However, it likely misses many important interactions. There is little reason to believe that early 20$^{\text{th}}$ century earnings gaps between whites and blacks did not differ by region, nor does it seem likely that the age-earnings profile was the same across all occupations.

Our approach bridges these alternatives and aims to balance the need for a rich model of income determinants with the limitations imposed by the small number of observations available for some occupations (particularly for certain groups). For a given base year, we compute a set of lasso-adjusted industry, demographic, and occupation (LIDO) scores as follows. For each Census-classified industry,[11] we regress log income on a set of demographic covariates for all individuals between the ages of 20 and 70 employed with positive earnings in that industry. We use the lasso algorithm, which solves the standard least squares problem subject to a constraint on the sum of the absolute values of the model coefficients (Tibshirani, 1996). This regularization approach controls the complexity of the model based on the importance of the predictors and the size and composition of the sample.

We allow for the following regressors: indicators for all occupations within the given industry, a polynomial for age, indicators for sex, race, and state of residence, and interactions between (1) sex and race, (2) sex and region, (3) occupation and sex, (4) occupation and an indicator for white, (5) Census region and an indicator for white, and (6) Census region and an indicator for black. In 1950, this results in a maximum of 654 possible covariates for the industry with the largest number of represented occupations (educational services). In general, the number of possible covariates is large relative to the sample size for each industry, and in many cases it may exceed the number of observations. The lasso algorithm shrinks coefficients depending on their relative importance, with the constraint forcing the coefficients on the least relevant predictors to zero. The sparsity induced by the lasso depends on the choice of tuning parameter λ for each particular industry (described further below).

The set of potential predictors allows for occupational income to depend on a wide range

[11] Industries follow the 1950 Census Bureau industrial classification system.

of factors. Income for a given occupation can vary depending on the particular industry in which an individual works; it can also vary in flexible ways with race, sex, and geographic region. The age profile of earnings can also differ by industry. This generates income scores that more closely reflect reality than the interaction-free regression approach described above. It also avoids the small-cell overfitting problem that arises from the fully nonparametric approach; the lasso retains only the most relevant predictors of income differences, and the size of the model is scaled depending on the number of observations in each industry.

The extent to which the lasso generates a sparse model depends on the choice of tuning parameter λ, which reflects the stringency of the constraint. Since the importance of different demographic factors likely varies by industry, a one-size-fits-all choice would be inappropriate. We instead use 10-fold cross-validation to select a λ that minimizes out-of-sample mean squared error for each industry.

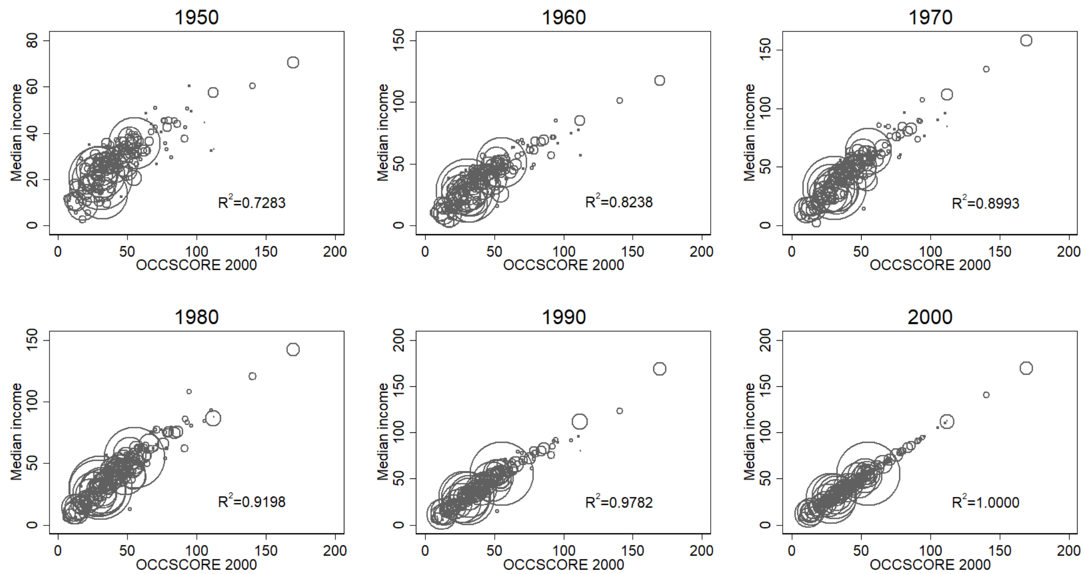
4 Results

4.1 Persistence of Occupational Income

The occupational income score is a weighted average of the median earnings for males and females for each occupational category in 1950. Economic historians will often use this an a proxy for income in the 1850-1940 Censuses. This variable is likely a reasonable proxy for earnings in 1950, but it is unclear whether the relative earnings of occupations are sufficiently stable to remain an accurate proxy for income in the decades before 1950. In this subsection, we test whether median earnings of an occupation accurately predict median earnings in the decades before the base year. We do this by constructing a 2000-based OCCSCORE and test how well the 2000-based OCCSCORE predicts median earnings from past Censuses. If the 2000-based OCCSCORE successfully proxies for median earnings in 1950, then the 1950 OCCSCORE may be a reasonable proxy for median occupational earnings in 1900.

The results on this exercise are in Figure 1. Each circle is an occupation weighted by

Figure 1: 2000 occupational income score and median income



Notes: Median earnings for each occupation are from the 1% sample of the U.S. Census. The 2000-based OCCSCOREs are the median earnings from the 2000 Census for individuals in each occupation. The size of each circle corresponds to the number of individuals in the occupational category during that Census year. Median earnings and OCCSCORE are measured in hundreds of 1950 dollars.

the size of the occupational cell. The 2000 OCCSCORE perfectly predicts median earnings in 2000 by construction. For each decade removed from 2000, the R^2 decreases, implying that OCCSCORE is becoming worse as a proxy for median earnings. Even 50 years removed from the base year, the $R^2 = 0.73$, implying that OCCSCORE remains a strong proxy for median earnings.

To provide further evidence, we examine changes in the rank correlation of median occupational income between 1950 and 2000. Measured by Spearman's rho, the correlation between occupational rankings in 1990 and 2000 is 0.97. While this declines over time, it does so very gradually. Between 1950 and 2000, the correlation is 0.81, which is still fairly high. While we cannot examine how this correlation changes nationally in the decades before 1950, we can examine the correlation between occupational income in 1950 and that in 1915 Iowa using state Census data (described in section 5.1). Using this information, we find that the rank correlation between 1950 and 1915 occupational earnings in Iowa is 0.7.

4.2 Errors of Magnitude

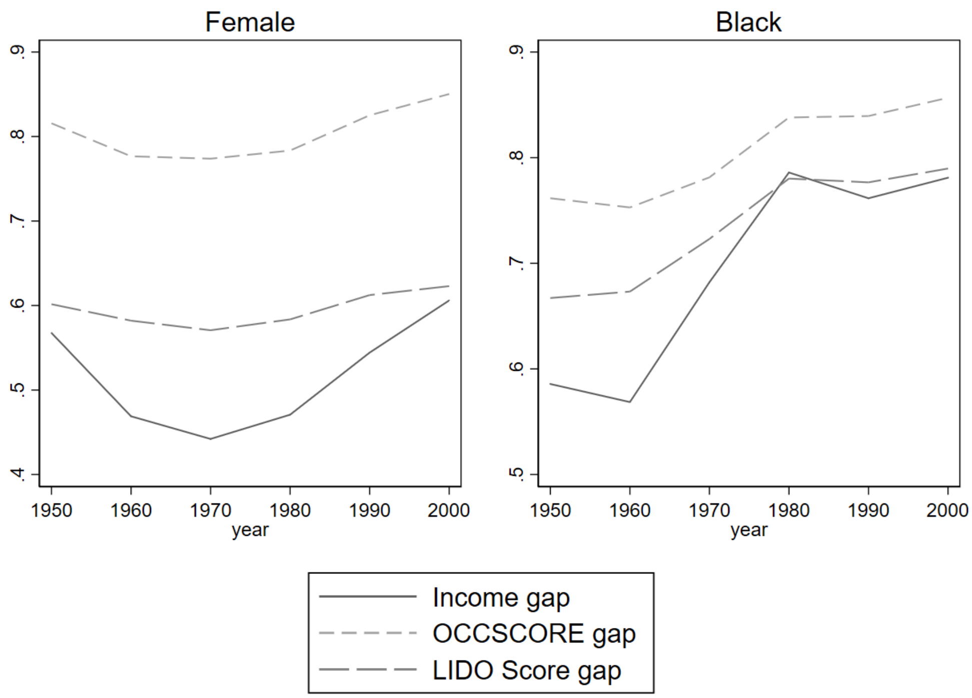
In this section, we analyze the magnitude of bias induced when using OCCSCORE and LIDO score as a proxy for income in an earnings regression. For the moment, we focus on estimating racial and gender earnings gaps. Gelman and Tuerlinckx (2000) and Gelman and Carlin (2014) introduce the Type M error rate as the expected value of an estimate divided by the true parameter value, conditional on the estimate being statistically different from zero. In this context, the true earnings gap is the earnings gap found using actual income data, and the estimated earnings gap is the gap using a proxy for income (either OCCSCORE or LIDO score).[12]

Labor economists and economic historians often use a 1950-based occupational income score to proxy earnings, and then estimate models using data from pre-1950 Census years. We cannot directly test whether such proxies are valid without better historical earnings data. However, we can make modern data worse, so that the modern data suffers from the same problems as historical data. Here, we generate a 2000-based occupational income score and compare estimated racial and gender gaps from 1950-1990 with the true earning gaps a researcher would have obtained by using actual earnings instead of the proxy.

Figure 2 graphs the implied earnings gaps using the three regressions. The first specification regresses the log of earnings on a set of dummies for state of residence, sex, race, and nativity. In addition to these dummy variables, the regression includes age and age squared. We run the regressions separately for every Census year from 1950 to 2000. Because we assume researchers would have used earnings instead of occupational income scores if earnings data were available, we treat these coefficients as the true parameters that researchers would like to estimate. The second regression uses the log of the 2000-based OCCSCORE instead of log earnings as the dependent variable. The dependent variable for the last regression is the log of the 2000-based LIDO score. For each regression, we restrict the sample to adults

[12] Although we do not formally take the expectation of the earnings gaps, the large sample size of the Census ensures that the standard errors are small and the estimated coefficients will be close to their expectations.

Figure 2: Earnings ratios using earnings, OCCSCORE, and adjusted OCCSCORE



Notes: The data are from IPUMS (see Ruggles, Genadek, Goeken, Grover and Sobek (2015)). The graph displays the implied female/male and black/white income ratios from six earnings regressions. Note that the gaps are conditional on age, age squared, a dummy variable for US-born, and state of residency. The female earnings gap is conditional on race, and the black/white earnings gap is conditional on sex. OCCSCORE uses a 2000-based occupational income score, whereas adjusted OCCSCORE is a 2000-based occupational income score conditional on race, sex, age, and region. The sample is restricted to those between ages 25 and 65 who were in the labor force.

ages 25-65 who were in the labor force.

As expected, earnings gaps have declined for blacks since the 1950s and for women since the 1970s, and this is reflected in all three models. However, the magnitude of the gap is highly attenuated when the log of OCCSCORE is used in place of true earnings. For all years, the coefficients on sex and race are of the same sign and statistically significant in all specifications, but the coefficients from the OCCSCORE specification suffer from substantial attenuation bias. Using the LIDO score as the dependent variable greatly reduces this bias. The earnings gap estimated using the LIDO score more closely mirrors the true earnings gap than the OCCSCORE estimates.

Our estimates of the female/male earnings ratio are similar to the extant literature (Goldin, 1990, p. 62). The female/male earnings ratio declined between 1950-1960, after which the gender gap slowly narrowed. Margo (2016) provides Census estimates of the black/white earnings gap that are similar to ours. Black income increased relative to whites during the 1960s and 1970s, but the ratio has not narrowed significantly since the 1980s. Smith (1984) estimates the black/white income gap by assigning each individual the average income of a race by sex by age group cell from the 1970 Census. These estimates, produced at least a decade before the IPUMS OCCSCORE variable was regularly in use, are in essence an adjusted OCCSCORE.

4.3 Errors of Sign

If researchers are primarily concerned with the direction of an effect as opposed to its magnitude, the previous results suggest that qualitative conclusions may not be seriously affected by the use of occupational income scores. The use of OCCSCORE as the dependent variable did not result in sign changes for gender and racial earnings differences. This result does not generalize to regressors with signs less predictable than race and gender indicators. Here, we show that OCCSCORE frequently results in errors of sign, or Type S errors. A Type S error occurs when the true population parameter is non-zero and the estimate is statistically significant and of the wrong sign (see Gelman and Carlin (2014) and Gelman and Tuerlinckx (2000)).

In this section, we consider two models. We regress log income on 176 dummy explanatory variables: a set of dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. Then, we estimate the model with a 2000-based occupational income score as the dependent variable and then again with a 2000-based LIDO score that is adjusted for industry and demographics. Standard errors are clustered at the state level.

We then compare how often these models give conflicting results compared to the "true model" in which the dependent variable is log income. Many researchers are satisfied to use estimators that are biased towards zero since the sign of the estimator is likely to be the same as that of the parameter they are trying to estimate. A more serious problem occurs when researchers find spurious results that are statistically significant and of the wrong sign. For this reason, we say that the two models conflict for a particular coefficient if they produce opposite signs, but the model using occupational income score is statistically significant.

The results from this exercise are shown in Table 2.[13] Estimates that are significant in both the earnings and OCCSCORE regressions are of conflicting signs 4% of the time and the problem worsens as one gets farther removed from the base year. By 1950, 20 percent of statistically significant coefficients have the wrong sign when using OCCSCORE in place of earnings. The variables that are particularly affected are state and age. This finding is especially troubling for difference-in-differences estimates in which the treatment variable is often an explicit function of state of residency and birth cohort. In 1970, 33% of the age coefficients are incorrectly signed; in 1950, 52% of the state coefficients are incorrectly signed. This problem is greatly reduced when using the LIDO score as the dependent variable. From 1980-2000, none of the coefficients are statistically significant and of the wrong sign. The numbers for 1950-1970 are only 4%, 1%, and 2%, respectively.

Table 3 displays the mean ratios of the estimated coefficients to the "true" earnings regression coefficients. Ideally, these ratios would be close to 1 and would never be negative. The results suggest that the OCCSCORE coefficients are typically 28-37% of the earnings regression coefficients. The LIDO score coefficients are closer to being centered around 1 and depending on the year vary between 94-121% of the earnings regression coefficients. Variables that are unlikely to be correlated with the demographic and industry adjusting variables (such as household and family characteristics) produce similar estimates for both the OCCSCORE and LIDO score regressions.

Figure 3 graphs kernel density estimates of the ratio of the estimated and true coefficients.

[13] The results are similar if we drop those in agriculture, an industry in which measuring income is particularly difficult (see Steckel (1991)).

Table 2: Percent of significant coefficients with conflicting signs

OCCSCORE						
Year	1950	1960	1970	1980	1990	2000
Age	0.06	0.00	0.33	0.17	0.11	0.14
State	0.52	0.32	0.05	0.00	0.00	0.00
Birth place	0.00	0.00	0.00	0.00	0.00	0.00
Race and sex	0.00	0.00	0.00	0.00	0.00	0.00
Family and household	0.09	0.00	0.00	0.00	0.00	0.00
Mean	0.20	0.10	0.14	0.06	0.03	0.04

LIDO score						
Year	1950	1960	1970	1980	1990	2000
Age	0.06	0.00	0.03	0.00	0.00	0.00
State	0.00	0.00	0.00	0.00	0.00	0.00
Birth place	0.00	0.00	0.00	0.00	0.00	0.00
Race and sex	0.00	0.00	0.00	0.00	0.00	0.00
Family	0.10	0.04	0.04	0.00	0.00	0.00
Mean	0.04	0.01	0.02	0.00	0.00	0.00

Notes: Data are from the 1% samples of the U.S. Census downloaded from IPUMS. We regress the measure of labor market outcomes on 176 dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The "true model" uses log of earnings. Each cell displays the proportion of those estimates that are statistically significant in both models and of the wrong sign.

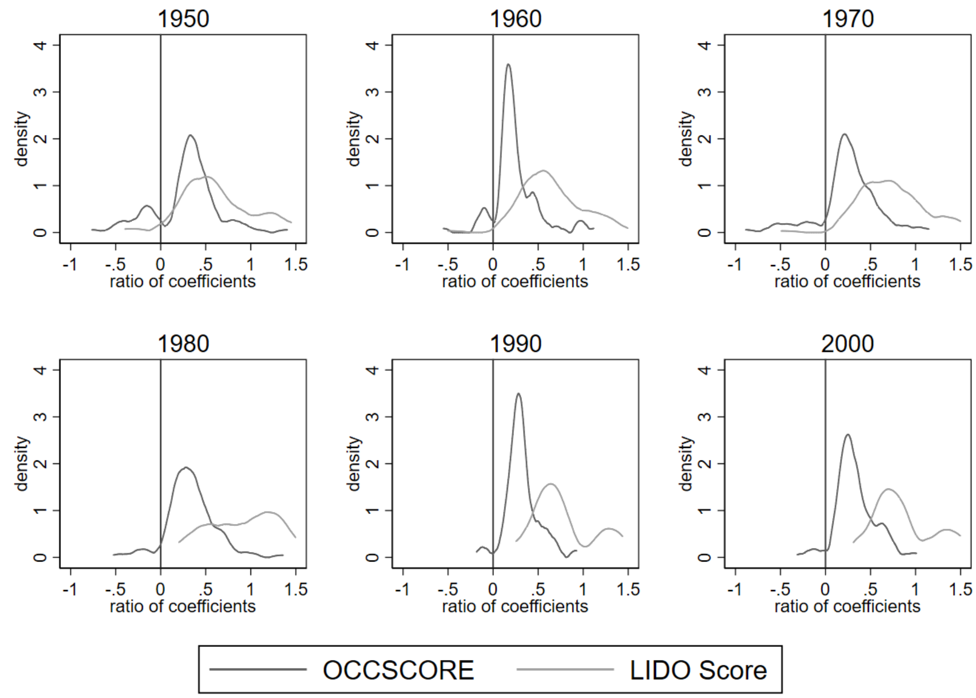
Table 3: Mean ratio of the estimated coefficient to the "true" coefficient

	OCCSCORE					
Year	1950	1960	1970	1980	1990	2000
Age	0.38	0.18	-0.01	0.15	0.20	0.14
State	0.06	0.26	0.29	0.40	0.30	0.33
Birth place	1.07	0.74	0.66	0.73	0.93	0.76
Race and sex	0.56	0.55	0.59	0.49	0.51	0.55
Family and household	0.32	0.39	0.50	0.54	0.44	0.50
Mean	0.28	0.30	0.27	0.37	0.32	0.33

	LIDO OCCSCORE					
Year	1950	1960	1970	1980	1990	2000
Age	1.96	2.39	2.47	1.59	1.59	1.48
State	0.89	0.65	0.76	1.05	0.66	0.75
Birth place	1.06	0.69	0.63	0.74	0.90	0.91
Race and sex	0.94	0.89	1.07	0.90	0.90	0.97
Family and household	0.34	0.37	0.43	0.55	0.50	0.57
Mean	1.12	1.21	1.28	1.09	0.94	0.97

Notes: Data are from the 1% samples of the U.S. Census downloaded from IPUMS. We regress the measure of labor market outcomes on 176 dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The "true model" uses log of earnings. Each cell displays the proportion of those estimates that are statistically significant in both models and of the wrong sign.

Figure 3: Density of Ratios of Estimated to True Coefficients using OCCSCORE and Adjusted OCCSCORE



Notes: Data are from IPUMS (see Ruggles et al. (2015)). The figures display the density of the ratio of the estimated coefficients (using a 2000-based OCCSCORE) and the "true" coefficient using observable income. Each year contains 185 regression coefficients and includes a set of dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The sample is restricted to those between ages 25 and 65 who were in the labor force.

The density for LIDO score is closer to being centered around one (less attenuation bias) and has less weight to the left of zero (conflicting signs). Using OCCSCORE leads to less accurate results the farther away from the base year we get, whereas estimates using the LIDO score are likely to be of the same sign even 50 years prior to the base year. Although the LIDO score is rarely of the wrong sign, it too suffers from some attenuation bias when the data are 50 years removed from the base year.

5 Applications

5.1 Earnings Gaps in the Iowa State Census, 1915

The analysis in section 4 shows that when examining gender and racial earnings gaps, adjusted OCCSCOREs substantially outperform the standard OCCSCORE using modern decennial Census data. To assess the extent to which this conclusion applies in a more historical context, we exploit a rare source of pre-1950 income data, the 1915 Iowa State Census (Goldin and Katz, 2010).[14] This was the first Census in the US to collect data on income prior to 1940. The sample digitized by Goldin and Katz (2010) contains records on 5.5% of the urban population drawn from three of Iowa's largest cities: Des Moines, Dubuque, and Davenport. It also contains 1.8% of the population of counties not containing a major city; the ten counties used span the geography of the state. We compare racial and gender earnings gaps estimated using the standard OCCSCORE and our LIDO score to those estimated using true earnings.

For the estimation, we restrict the sample to those between the ages of 20 and 70 and exclude those with missing occupation data or zero/missing earnings.[15] The Census reports occupation categories according to the 1940 scheme. We cross-walked these with the 1950 scheme to match individuals in 1915 to their 1950 OCCSCORE.[16] The final sample includes 15,201 individuals. We estimate the earnings gap between whites and blacks and men and women; approximately 1% of the sample is black (196 obs) and 14% of the sample is female (2,153 obs).

In column (1) of Table 4, we report the coefficients from a regression of log earnings on an indicator for black and female as well as a quadratic polynomial for age. Women and

[14] This data has also been used to examine intergenerational mobility by Feigenbaum (2017).

[15] We also exclude those whose race is recorded as missing (19 observations) and those whose race is recorded as Mixed or Asian (5 observations); the sample is too small to reliably estimate earnings gaps for these groups.

[16] In some cases, the 1940 scheme aggregated some occupations; for example, bookkeepers, accountants, and cashiers fall into one occupation category in 1940 but are disaggregated into three separate categories in 1950. There are 7 occupation categories in 1940 (out of 194 total) that cannot be matched uniquely to a 1950 occupation; we exclude individuals falling into these categories.

Table 4: Earnings Gaps in the 1915 Iowa State Census

	Log of earnings	Log of 1950 OCCSCORE	LIDO score
	(1)	(2)	(3)
Black	-0.326***	-0.146***	-0.362***
	(0.0469)	(0.0292)	(0.0226)
Female	-0.456***	-0.025**	-0.522***
	(0.0155)	(0.0097)	(0.0075)
Age	0.174***	0.025***	0.077***
	(0.0062)	(0.0039)	(0.0030)
Age^2	-0.133***	-0.018***	-0.075***
	(0.0052)	(0.0032)	(0.0025)
Observations	15,201	15,201	15,201
R^2	0.130	0.006	0.320

Notes: Linear regressions of earnings measures on blacks and female indicators as well as a quadratic polynomial in age. Sample excludes those whose race is recorded as Missing, Mixed, or Asian (24 observations), those who are below the age of 20 or above the age of 70, those with missing occupation data, and those with zero or missing earnings. Sample is further restricted to individuals for which all three OCCSCORE measures could be calculated. *** p<0.01, ** p<0.05, * p<0.1

blacks earn substantially less than white men. As is typical, earnings increase with age but at a diminishing rate. In column (2), we replace log earnings with the log of the standard 1950 OCCSCORE. The results change substantially: the black-white earnings gap coefficient declines by half, and the gender earnings gap declines by 95%. The use of standard 1950 OCCSCOREs as a proxy for earnings leads to substantially misleading results in this sample.

Moving to column (3), we replace the unadjusted OCCSCORE with 1950 LIDO score.[17] Using this approach, the earnings gap for blacks is now almost the same as that estimated using true earnings. The estimated earnings gap for women is slightly larger than the true value but much closer in magnitude than the standard OCCSCORE estimate. In this context, it seems clear that the LIDO score is a substantially better proxy for earnings than OCCSCORE.

[17] Because industry was not recorded in the Iowa State Census, we instead estimate our lasso-adjusted score by occupation (retaining all of the other predictors listed in section 3.3).

5.2 Estimates of Intergenerational Mobility

Labor economists often measure intergenerational mobility by regressing a son's socioeconomic status on his father's socioeconomic status:

$$I_i^{\text{son}} = \beta_0 + \beta_1 I_i^{\text{father}} + \delta_i \tag{17}$$

where I_i^{son} is the log income of a son observed during adulthood, and I_i^{father} is the log income of a father observed while the son was a child. The transmission coefficient β_1 is an elasticity typically between 0 and 1, with 1 representing perfect immobility between generations and 0 representing perfect mobility. Historical evidence on occupational mobility across generations relies heavily on occupational income scores instead of income for two reasons. First, to obtain data on fathers' and sons' labor market outcomes in the Census, one needs to link across Census years, which is typically only possible using given and surnames. Names do not become publicly available in the Census until 72 years after the Census year, meaning occupations are the only available labor market outcomes for both fathers and sons. Second, estimates of how intergenerational mobility have changed over time require data spanning at least three generations, implying that such estimates must make use of historical data.

Let $e_i^{\text{son}} = I_i^{\text{son}} - \tilde{y}_i^{\text{son}}$ and $e_i^{\text{father}} = I_i^{\text{father}} - \tilde{y}_i^{\text{father}}$ be the measurement error from using an occupational index (either OCCSCORE or LIDO score) for the son and father, respectively. Then researchers estimate:

$$\tilde{y}_i^{\text{son}} = \beta_0 + \beta_1 \tilde{y}_i^{\text{father}} + \underbrace{\beta_1 e_i^{\text{father}} - e_i^{\text{son}} + \delta_i}_{\epsilon_i}. \tag{18}$$

This regression differs from the model in Section 3 since OCCSCORE appears on both the left-hand and right-hand side of the regression. The measurement errors e_i are likely to be smaller if one uses a demographically-adjusted LIDO score instead of OCCSCORE, since racial, age, and regional differences in occupational earnings will not be captured in e_i. However, when using OCCSCORE, some of the measurement error is likely to cancel out since e_i^{son} is positively correlated with e_i^{father}. Our estimate of the transmission coefficient will be biased if $\text{Cov}(\tilde{y}_i^{\text{father}}, \beta_1 e_i^{\text{father}} - e_i^{\text{son}}) \neq 0$. If there is little intergenerational mobility, in which case β_1 is close to 1, and if the son's measurement error is highly correlated with the father's measurement error, then the second term of covariance is close to zero. Alternatively, suppose $e_i^{\text{son}} = \tilde{\beta}_0 + \tilde{\beta}_1 e_i^{father} + \nu_i$, where ν_i is independent of all other variables. The transmission coefficient $\tilde{\beta}_1$ reflects that fathers who earn above average within their occupations are likely to have sons who earn above average within occupations. Then, $\text{Cov}(\tilde{y}_i^{\text{father}}, \beta_1 e_i^{\text{father}} - e_i^{\text{son}}) = \text{Cov}(\tilde{y}_i^{\text{father}}, \beta_1 e_i^{\text{father}} - \tilde{\beta}_1 e_i^{\text{father}})$. Thus, the bias from estimating equation 18 using OCCSCOREs will be small so long as the transmission in overall income from father to son is similar to the transmission of excess income within occupation. For these reasons, using occupational income scores instead of income should lead to a smaller amount bias in this particular context.

In Table 5, we provide estimates of intergenerational mobility using the IPUMS linked data sets. These data link the 1% samples of the 1850, 1860, 1900, and 1910 Censuses to the 1880 complete count. The sample is restricted to those who, during the first Census year, were children of the household head, male, and no older than 15 years old. We regress the log of a son's OCCSCORE (during the second Census year) on the log of the father's OCCSCORE (during the first Census year), and then repeat the regression using LIDO score. To make the estimates comparable, we restrict the sample father-son pairs in which neither the father nor the son has a missing LIDO score. The resulting coefficients are elasticities, with higher coefficients implying occupational immobility. Row 1 of columns (2) and (4) of Table 5 are replications of the estimates in row 7 of Table 3 of Olivetti and Paserman (2015), but restricting the sample to those with non-missing LIDO scores. We present estimates for whites using all samples, and for blacks using samples in which both the father and son are observed in the postbellum period.

The OCCSCORE estimates suggest that blacks had twice the intergenerational mobility of whites (β_1 closer to zero). The LIDO score estimates suggest that black intergenerational mobility was much closer to white intergenerational mobility than previously thought. In fact, between 1880-1900 and 1880-1910, blacks had less intergenerational mobility.

6 Conclusion

Using modern Census data, we find that even though occupational income is highly correlated over time, occupational income scores systematically underestimate racial and gender income gaps by a substantial margin. Furthermore, other standard earnings regression covariates, like state of residency and state of birth indicators, are strongly attenuated and can be of the wrong sign up to 20 percent of the time. We construct a new lasso-adjusted industry, demographic, and occupation (LIDO) score which flexibly accounts for differences in earnings across race, gender, age, region, occupation, and industry (variables available in every Census going back to 1850). Our alternative scores significantly reduce attenuation bias and limit the probability that the estimates are of the wrong sign.

To examine the performance of the LIDO score in a historical context, we exploit the 1915 Iowa State Census, which collected data on both occupation and earnings. We find that estimated race and gender earnings gaps in 1915 Iowa using true earnings are sizable; however, when using standard OCCSCORE as a proxy, the racial earnings gap is attenuated by half and the gender earnings gap is attenuated by 95%. Our LIDO score yields race and gender earnings gaps very close to their true values. We also conduct an analysis of OCCSCORE-induced bias in measures of intergenerational income transmission. This analysis is based on father-son pairs linked across the 1850, 1860, 1880, 1900, and 1910 decennial Censuses. In this setting, we find that standard OCCSCOREs and LIDO scores perform similarly for white males, because measurement errors for fathers and sons are likely to be correlated. However, transmission coefficients are substantially attenuated for black

Table 5: Estimates of Intergenerational Mobility

Panel A: Mobility among whites

Dependent variable: log of son's OCCSCORE

	(1)	(2)	(3)	(4)	(5)	(6)
	1850-1880	1860-1880	1880-1900	1880-1910	1880-1920	1880-1930
Log of father's OCCSCORE	0.402***	0.460***	0.544***	0.428***	0.403***	0.379***
	(0.0219)	(0.0172)	(0.0127)	(0.0134)	(0.0150)	(0.0146)
N	2804	3945	8354	7345	5687	5524

Dependent variable: log of son's LIDO score

	(1)	(2)	(3)	(4)	(5)	(6)
	1850-1880	1860-1880	1880-1900	1880-1910	1880-1920	1880-1930
Log of father's LIDO score	0.457***	0.439***	0.408***	0.367***	0.394***	0.387***
	(0.0193)	(0.0154)	(0.0107)	(0.0105)	(0.0132)	(0.0138)
N	2804	3945	8354	7345	5687	5524

Panel B: Mobility among blacks

Dependent variable: log of son's OCCSCORE

	(1)	(2)	(3)	(4)	(5)	(6)
	1850-1880	1860-1880	1880-1900	1880-1910	1880-1920	1880-1930
Log of father's OCCSCORE			0.174***	0.161**	0.0825	0.184*
			(0.0514)	(0.0581)	(0.0698)	(0.0786)
N			520	375	250	204

Dependent variable: log of son's LIDO score

	(1)	(2)	(3)	(4)	(5)	(6)
	1850-1880	1860-1880	1880-1900	1880-1910	1880-1920	1880-1930
Log of father's LIDO score			0.600***	0.541***	0.562***	0.506***
			(0.0436)	(0.0600)	(0.0954)	(0.0969)
N			520	375	250	204

Notes: Data are from the IPUMS linked data files. These data are from 1% samples of the 1850, 1860, 1900, and 1910 Censuses linked to the 1880 complete count. The sample is restricted to those who during the first Census year were children of the household head, male, and no older than 15 years old. Standard errors are in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

men, suggesting a much higher rate of intergenerational mobility than is likely true.

Our results strongly suggest that future research in economic history and other fields should utilize LIDO scores whenever researchers are interested in true earnings. When should researchers continue to use standard occupational income scores? It is possible, although not obvious, that unadjusted OCCSCOREs are a reasonable proxy for total earnings over the life cycle. A recent college graduate may have a low adjusted OCCSCORE since young professionals are below their peak occupational earnings, whereas the standard OCCSCORE would assign workers of all ages within the occupation the same earnings. Race and gender earnings gaps may also shrink over the life-cycle.[18] Without linked Census data, it is impossible to know whether OCCSCOREs or LIDO scores provide a better proxy for lifetime wealth. Although linked Census data does exist, it does not cover the 1950-2000 periods in which earnings data are available.

Studies that are primarily interested in occupational status, rather than earnings, may wish to retain the standard OCCSCORE. The 1950 OCCSCORE provides a ranking of occupations by earnings, and we find substantial persistence in these occupational rankings. The logical jump that researchers have made, but is not supported by the data, is that if a treatment increased occupational status, it also increased income. To make such statements credible, researchers should use LIDO scores. Even for studies that focus on life-time earnings or occupational status, there is another reason to use LIDO scores as a complement to standard OCCSCOREs. Labor economists using modern data almost invariably use earnings instead of using OCCSCORE, and using LIDO scores would make the historical literature more comparable to the modern literature.

References

Aaronson, Daniel, Fabian Lange, and Bhashkar Mazumder (2014). Fertility transitions along the extensive and intensive margins. *American Economic Review*, **104**(11), pp. 3701–3724.

[18]Although most studies use a base-year that occurs after the study period; thus, the racial and gender gaps that exist during the base-year are unlikely to disappear.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, **102**(5), pp. 1832–1856.

Angrist, Josh (2002). How do sex ratios affect marriage and labor markets? Evidence from America's second generation. *Quarterly Journal of Economics*, pp. 997–1038.

Bailey, Martha J and William J Collins (2006). The wage gains of African-American women in the 1940s. *Journal of Economic History*, **66**(03), pp. 737–777.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, **119**(1), pp. 249–275.

Bleakley, Hoyt (2007). Disease and development: evidence from hookworm eradication in the American South. *Quarterly Journal of Economics*, **122**(1), p. 73.

——— (2010). Malaria eradication in the Americas: A retrospective analysis of childhood exposure. *American Economic Journal: Applied Economics*, **2**(2), pp. 1–45.

Bleakley, Hoyt and Joseph Ferrie (2016). Shocking behavior: Random wealth in antebellum Georgia and human capital across generations. *Quarterly Journal of Economics*, **131**(3), pp. 1455–1495.

Bleakley, Hoyt and Fabian Lange (2009). Chronic disease burden and the interaction of education, fertility, and growth. *Review of Economics and Statistics*, **91**(1), pp. 52–65.

Carruthers, Celeste K and Marianne H Wanamaker (2017). Separate and unequal in the labor market: human capital and the jim crow wage gap. *Journal of Labor Economics*, **35**(3), pp. 000–000.

Chin, Aimee (2005). Long-Run Labor Market Effects of Japanese American Internment during World War II on Working-Age Male Internees. *Journal of Labor Economics*, **23**(3), pp. 491–525.

Collins, William J (2000). African-American economic mobility in the 1940s: a portrait from the Palmer Survey. *Journal of Economic History*, **60**(03), pp. 756–781.

Collins, William J and Marianne H Wanamaker (2014). Selection and economic gains in the great migration of african americans: New evidence from linked census data. *American Economic Journal: Applied Economics*, **6**(1), pp. 220–252.

——— (2015). The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants. *Journal of Economic History*, **75**(04), pp. 947–992.

Cook, Lisa D, Trevon D Logan, and John M Parman (2014). Distinctively black names in the American past. *Explorations in Economic History*, **53**, pp. 64–82.

——— (2016). The mortality consequences of distinctively black names. *Explorations in Economic History*, **59**, pp. 114–125.

Feigenbaum, James J. (2017). Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940. *Economic Journal*.

Gelman, Andrew and John Carlin (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, **9**(6), pp. 641–651.

Gelman, Andrew and Francis Tuerlinckx (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, **15**(3), pp. 373–390.

Goldin, Claudia (1990). *The gender gap: An economic history of American women*. New York: Cambridge University Press.

Goldin, Claudia and Lawrence Katz (2010). The 1915 Iowa State Census Project: ICPSR28501-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, `http://doi.org/10.3886/ICPSR28501.v1`.

Hyslop, Dean R and Guido W Imbens (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics*, **19**(4), pp. 475–481.

Lee, Dara N (2013). The impact of repealing Sunday closing laws on educational attainment. *Journal of Human Resources*, **48**(2), pp. 286–310.

Lee, Sanghoon and Jeffrey Lin (2017). Natural Amenities, Neighbourhood Dynamics, and Persistence in the Spatial Distribution of Income. *Review of Economic Studies*, p. rdx018.

Lleras-Muney, Adriana and Allison Shertzer (2015). Did the Americanization Movement Succeed? An Evaluation of the Effect of English-Only and Compulsory Schooling Laws on Immigrants. *American Economic Journal: Economic Policy*, **7**(3), pp. 258–290.

Margo, Robert A (2016). Obama, Katrina, and the Persistence of Racial Inequality. *Journal of Economic History*, **76**(02), pp. 301–341.

Massey, Catherine G (2016). Immigration quotas and immigrant selection. *Explorations in Economic History*, **60**, pp. 21–40.

Minns, Chris (2000). Income, cohort effects, and occupational mobility: a new look at immigration to the United States at the turn of the 20th century. *Explorations in Economic History*, **37**(4), pp. 326–350.

Olivetti, Claudia and M Daniele Paserman (2015). In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850–1940. *American Economic Review*, **105**(8), pp. 2695–2724.

Romer, Christina (1986). Spurious volatility in historical unemployment data. *Journal of Political Economy*, pp. 1–37.

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.

Saavedra, Martin (2015). School quality and educational attainment: Japanese American internment as a natural experiment. *Explorations in Economic History*, **57**, pp. 59–78.

——— (2017). Early-life disease exposure and occupational status: The impact of yellow fever during the 19th century. *Explorations in Economic History*, **64**, pp. 62–81.

Sacerdote, Bruce (2005). Slavery and the intergenerational transmission of human capital. *Review of Economics and Statistics*, **87**(2), pp. 217–234.

Smith, James P (1984). Race and human capital. *American Economic Review*, **74**(4), pp. 685–698.

Steckel, Richard H (1991). The quality of census data for historical inquiry: A research agenda. *Social Science History*, **15**(4), pp. 579–599.

Stephens, Melvin and Dou-Yan Yang (2014). Compulsory education and the benefits of schooling. *American Economic Review*, **104**(6), pp. 1777–1792.

Tibshirani, Robert (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**(1), pp. 267–288.

Ward, Zachary (2017). Birds of passage: Return migration, self-selection and immigration quotas. *Explorations in Economic History*, **64**, pp. 37–52.