ScholarOne - Social Rigidity Across and Within Generations: A Predictive Approach

Haowen Zheng¹ and Siwei Cheng²

¹Cornell University  
²Affiliation not available

March 22, 2024

Abstract

How well can individuals’ parental background and previous life experiences predict their mid-life Socio-Economic Status (SES) attainment? This question is central to stratification research, as a strong power of earlier experiences in predicting later-life outcomes signals substantial intra- or intergenerational status persistence, or put simply, social rigidity. Running machine learning models on panel data to predict outcomes that include hourly wage, total income, family income, and occupational status, we find that a large number (around 4,000) of predictors commonly used in the stratification literature improves the prediction of one’s life chances in middle to late adulthood by about 10 to 50 percent, compared with a null model that uses a simple mean of the outcome variable. The level of predictability depends on the specific outcome being analyzed, with labor market indicators like wages and occupational prestige being more predictable than broader socioeconomic measures such as overall personal and family income. Grouping a comprehensive list of predictors into four unique sets that cover family background, childhood and adolescence development, early labor market experiences, and early adulthood family formation, we find that including income, employment status, and occupational characteristics at early career significantly improves models’ prediction accuracy for mid-life SES attainment. Further, we illustrate the application of predictive models to examine sources of between-group disparity by life stages. Using the Black-white difference as the example, we find that racial differences in early labor market experiences are the most critical in explaining racial inequality in mid-life SES attainment, especially for low-income individuals.

Hosted file

Social Rigidity Across and Within Generations: A Predictive Approach
Abstract

How well can individuals’ parental background and previous life experiences predict their mid-life Socio-Economic Status (SES) attainment? This question is central to stratification research, as a strong power of earlier experiences in predicting later-life outcomes signals substantial intra- or intergenerational status persistence, or put simply, social rigidity. Running machine learning models on panel data to predict outcomes that include hourly wage, total income, family income, and occupational status, we find that a large number (around 4,000) of predictors commonly used in the stratification literature improves the prediction of one’s life chances in middle to late adulthood by about 10 to 50 percent, compared with a null model that uses a simple mean of the outcome variable. The level of predictability depends on the specific outcome being analyzed, with labor market indicators like wages and occupational prestige being more predictable than broader socioeconomic measures such as overall personal and family income. Grouping a comprehensive list of predictors into four unique sets that cover family background, childhood and adolescence development, early labor market experiences, and early adulthood family formation, we find that including income, employment status, and occupational characteristics at early career significantly improves models’ prediction accuracy for mid-life SES attainment. Further, we illustrate the application of predictive models to examine sources of between-group disparity by life stages. Using the Black-white difference as the example, we find that racial differences in early labor market experiences are the most critical in explaining racial inequality in mid-life SES attainment, especially for low-income individuals.
Introduction

In the stratification literature, the concept of “social rigidity” characterizes the persistence in individuals’ socioeconomic status over the life course and across generations. Starting from the classic status attainment model established by Blau and Duncan (1967), to modern-day causal mediation analysis that exploits the temporal structure of longitudinal data sets (Brand et al., 2019; VanderWeele and Tchetgen, 2017; Zhou, 2022a), decades of stratification research has explicated various aspects of social rigidity by analyzing pathways linking individuals’ social origin to their mid-life SES. This literature has considered pathways through intergenerational persistence (e.g., the effects of parental income and childhood experience) as well as pathways through intragenerational persistence (e.g., the effects of early-career attainment). Differentiating between different periods of the status attainment process allows researchers to decompose inequality in mid-life SES into distinct mediating mechanisms (e.g., Bloome and Furey, 2020; Zhou, 2022a).

In current analyses of the status attainment process, researchers typically specify a set of predictor variables and their interrelations, often represented through path diagrams (e.g., Blau and Duncan, 1967) or Directed Acyclic Graphs (DAGs) (e.g., Pearl, 2012; Robins, 2003; VanderWeele and Tchetgen, 2017). These methods are instrumental in unraveling the stratification mechanisms and resolving competing explanations for observed socioeconomic disparities among different groups. However, these analyses often concentrate on a limited set of variables, ignoring the complex reality where a myriad of factors can collectively influence an individual’s status attainment. There is a need for a more comprehensive approach, one that integrates a vast array of variables across various life stages, to more accurately model the long-term status attainment process.

Our study offers an alternative approach to existing literature by shifting to a predictive perspective. We delve into the power of various life course variables in predicting one’s mid-life SES. Historically rooted in statistics and machine learning, the predictive approach is a relatively recent addition to sociological studies (Brand et al., 2023; Garip, 2020;
Molina and Garip, 2019; Salganik et al. 2019, 2020). This approach leverages advanced machine learning algorithms to enhance accuracy in predicting life outcomes. The emergence of strong predictors of life outcomes often indicates a strong continuity of socioeconomic status throughout an individual’s life or across generations, reflecting a greater degree of social rigidity. Conversely, difficulties in predicting life outcomes suggest an element of unpredictability and variability in life chances, pointing towards greater social fluidity and uncertainty of life chances.

Adopting the predictive approach advances stratification research in three ways. Firstly, it expands beyond the scope of traditional mediation analysis, which often limits itself to a small number of predictive and mediating variables. Our method integrates a comprehensive array of variables that are previously identified as influential in determining life opportunities, offering a more complete perspective of the status attainment process. Secondly, while conventional analyses frequently depend on oversimplified model assumptions, the application of machine learning techniques affords us the flexibility to employ models with more complex, nuanced functional forms. Lastly, the machine learning approach enables us to unravel the intricate interactions among a vast set of variables that collectively influence mid-life Socio-Economic Status (SES) attainment.

Despite being a relatively recent methodological approach in sociology, machine learning has been increasingly used by social scientists to predict life outcomes, integrating a comprehensive array of background variables (e.g., Breen and Seltzer, Breen and Seltzer; Salganik et al., 2019, 2020; Savcisens et al., 2023; Vafa et al., 2022). Building on this line of work, our study draws on classic theories from inter- and intra-generational mobility research to categorize these variables into four distinct predictor sets corresponding to sequential life stages: (1) parental background, (2) childhood and adolescence development, (3) early labor market experiences, and (4) early adulthood family formation. We begin with models that exclusively focus on family background and progressively incorporate these predictor sets in line with the chronological progression of these life stages. Our goal is to predict mid-life
SES achieved between the ages of 40 and 50, a life stage where one’s SES tends to stabilize.

We use two empirical examples to illustrate the usefulness of this predictive approach in analyzing social rigidity. First, we assess social rigidity by evaluating prediction accuracy, which is the extent to which observed variables can account for overall variability. This involves analyzing the predictability of various life outcomes, such as hourly wage, total income, family income, and occupational status (Socioeconomic Index, SEI). Second, we use predictive models to examine sources of between-group disparities in mid-life SES, using Black-White difference in mid-life family income and gender difference in mid-life hourly wage as illustrative examples. By sequentially adding prediction sets into the prediction models and simulating predicted distributions, the predictive framework helps us understand the extent to which variation in racial/gender disparity can be explained by racial/gender differences in mechanisms and attributes at distinct life stages.

COMPARING ANALYTICAL FRAMEWORKS FOR STUDYING SOCIAL RIGIDITY

In this section, we compare the predictive framework with existing analytical frameworks and tools. In the following paragraphs and Table 1, we summarize the similarities and differences between the predictive approach and other approaches in terms of analytical tools, quantities of interest, number of variables, and model specification.

[Table 1 about here.]

Comparing with the Classic Status Attainment Model

We begin by contrasting our predictive framework with the traditional status attainment model, a key tool for analyzing social stratification processes. Originating from Blau and Duncan (1967), this approach explicates the relationships between socioeconomic variables using a path diagram. A classic example is Blau and Duncan’s path analysis, which incorporates five variables: father’s education, father’s occupation, respondent’s education, first job, and occupation in 1962. These variables are interconnected by lines indicating correlation.
and directed arrows signifying influences from upstream to downstream variables (e.g., an arrow from father’s education to respondent’s education). In this paradigm, the main quantities of interest include total and partial effects of the upstream variable on the obtained socioeconomic status, the strength of which indicates how strongly socioeconomic status is passed down across and within generations. These estimands are usually estimated as path coefficients through structural equation models, a system of interrelated equations. In model specification, the status attainment model framework often relies on assumptions about the causal ordering of a limited number of variables and employ linear models with no explicit interactions between variables.

Besides direct and indirect effects, another quantity of interest is the in-sample R-squared, which is used as a metric of model fitness, as well as to measure the amount of variation in status attainment that is accounted for by upstream variables. For example, Blau and Duncan (1967: 165-172) were able to account for 33 percent of the variance in (current) occupational (prestige) attainment of a national sample of men in 1960s, by using their father’s occupational prestige, respondent’s own educational attainment and first job. Similarly under the broad umbrella of the status attainment framework, the Wisconsin school of social psychological models showcase the usefulness of social psychological concepts in stratification processes by adding variables like mental ability and influence from significant others into models that further explains away variation in levels of educational and occupational attainment (Sewell et al., 1969). Though informative, the limitation of using R-squared terms stems from the fact that they are merely in-sample statistics, which naturally increase as one adds more variables into the model and can lead to model overfitting. Therefore, in-sample R-squared terms are limited in helping us learn about the predictability of in other datasets or situations beyond the specific conditions of the original data.

Our predictive approach diverges from the classic status attainment models by emphasizing prediction accuracy as a measure of social rigidity. Rather than focusing on the
statistical association among a few selected variables, we reformulate the question of social rigidity as a prediction task: given an individual's information is measured at various life stage points, how well can we predict what will happen to them at a future time point? Our key quantities of interest include out-of-sample R-squared values and predictive distributions for specific groups. While status attainment models often rely on a limited set of variables and use parametric methods like linear regressions, our predictive framework offers greater flexibility in both the number of variables and the model specification. Contemporary machine learning methods, adept at handling a large array of variables through feature selection and cross-validation, effectively mitigate the risk of overfitting. This adaptability is crucial for exploring the intricate intra- and inter-generational dynamics of stratification, as highlighted by existing literature emphasizing the complex interplay of numerous variables throughout the life course. For instance, to accurately reflect one's social origin, a broader range of variables beyond just father’s education and occupation is essential. Additionally, our approach relaxes assumptions about functional forms to account for the potentially complex interrelations among multiple predictors of socioeconomic outcomes. This allows for a more nuanced and comprehensive understanding of the factors influencing socioeconomic trajectories.

Comparing with Traditional and Contemporary Mediation Analysis

Researchers have also used various mediation analysis to characterize the transmission of socioeconomic status across generations and over the life course. We categorize mediation analytical frameworks into three more detailed categories: the conventional mediation analysis, machine-learning-based medication analysis, and sequential mediation analysis (for a more comprehensive review of causal mediation analysis, see Brand et al., 2023).

In general, mediation analysis frameworks primarily aim to understand the direct and indirect effects of a variable on an outcome by identifying the pathways of influence. These frameworks are particularly useful in unpacking the process of status attainment, such as
how family background influences one’s mid-career SES through schooling and employment trajectories, which can span across both inter- and intra-generational contexts (e.g., Bloome and Furey, 2020; Brand et al., 2019). Traditional mediation analysis often relies on linear models supported by directed acyclic graphs (DAG). However, more recent advancements have introduced more versatile machine learning methods. These modern approaches yield model-free estimates and can handle a moderate number of variables (e.g., Brand et al., 2023; Pearl, 2012; VanderWeele and Tchetgen, 2017). Meanwhile, researchers exploring multiple, time-ordered mediators in stratification or mobility processes often turn to sequential mediation analysis. This method, like our predictive approach, is adept at modeling multi-stage processes, such as those aligned with life stages, and frequently employs machine learning models for estimation (e.g., Imai and Yamamoto, 2013; Vansteelandt and Sjolander, 2016; Zhou, 2022b). The key distinction between our predictive framework and sequential mediation analysis lies in their respective focuses. While sequential mediation analysis focuses on stage-specific causal effects of particular treatments or interventions, our predictive framework focuses on the concept of social rigidity, as reflected in the predictive power (measured as the out-of-sample R-squared) of different sets of predictors.

In summary, by focusing on how much previous life stages can predict mid-life SES, our study joins a growing number of studies that aim to understand the predictability of life trajectories (Salganik et al., 2019; Savcisens et al., 2023; Badolato et al., 2023). For example, in the Fragile Families Challenge, social scientists aimed to predict specific outcome variables at the age of 15. Despite the availability of data encompassing thousands of variables collected to enhance the understanding of these families’ lives, researchers were not able to make accurate out-of-sample predictions. Efforts to measure the predictability of life outcomes not only inform our understanding of social rigidity from a social stratification point of view, questions on how to improve prediction accuracy could also lead to further developments in theory and methods. Building on a rich literature in inter- and intra-generational social mobility, our study is the first to investigate the predictability of important life course
Data and Methods

Data and variables

The main source of our data is the National Longitudinal Survey of Youth 1979 (NLSY79). The NLSY79 is a representative sample of 12,686 individuals born between 1957 and 1964, who resided in the United States at the start of the survey. The initial interview took place in 1979 when the respondents were between 14 and 22 years old. Subsequent interviews were conducted annually from 1979 to 1994, and then every two years. We use data from 1979 to 2016, which covers the life experiences of the participants from their early teenage years to the age of 50. We exclude the military sample and a subset that discontinued participation in 1990.

Other data sources include the Occupational Information Network (O*NET), from which we construct occupational skill variables, and the 1990 US Census, from which we construct occupational earnings. As a primary source of occupational information, the O*NET is a comprehensive database of worker attributes and job characteristics. Information is collected from and developed by incumbent workers and occupational analysts.

As shown in Figure 1, we assemble five sets of variables that capture an individual’s life course experiences: family background (A1), early childhood and adolescence development (A2), early labor market experience (B1), early adulthood family formation (B2), and mid-career socioeconomic attainment (C). We group the variables into distinct sets based on age. Figure 1 illustrates representative variables from each set for clarity.

[Figure 1 about here.]

**Set A1:** Family background variables describe one’s basic demographic characteristics and parental background. Basic demographic variables include age, gender, race, residence
(in the US or not, in the South or not, in rural areas or not). Parental background variables include parents’ educational attainment (in years), parents’ labor market attachment (whether they are working for pay, whether they are working more than 35 hours per week), parents’ occupation and SEI (Duncan SEI score), family poverty status, family structure (a dummy indicator for two parent families and a categorical indicator for all family types as specified in the survey \(^1\)), sibship size, and whether the family speaks foreign languages at home. Most of these variables measure respondents’ family background at or before age 14 through retrospective questions about their circumstances at that time.

**Set A2:** The second set of predictors also describes one’s pre-labor market characteristics, measured before age 25. First, it includes a set of cognitive and psychological factors commonly used in the literature, such as the Armed Forces Qualification Test (AFQT 1981), the Rotter Locus of Control Score (1979), and the Rosenberg Self-Esteem Scale (1980). This set also includes measurements on respondents’ juvenile delinquency activities in 1980, and whether they are ever charged by the police for engaging in illegal activities. We include respondents’ educational expectations (i.e., the respondent expects to be in school in the next 5 years) and aspirations (highest grade one would like to obtain), employment and occupational expectations in 5 years, and a friend’s educational aspirations. Second, this set includes educational attainment variables (high school degree, college degree, and advanced degree), and for those with college degree, their college major. Third, we exploit a rich set of high school contextual variables, such as the number of books in the school library, school-level dropout rate, ethnic/racial composition of students and faculty, etc. Lastly, we include a set of early family formation indicators to capture their family status before they formally enter the labor market. These variables include a respondent’s marriage status and whether the respondent has a child by age 18.

**Set B1:** The third set of predictors includes measures of respondents’ early labor market experience from age 25 to 35. Specifically, the variables include employment status,

---

\(^1\)These categories include e.g., "no man - mother," "father - no woman," "on my own," etc.
hours worked in a week, occupations (in detailed Census categories), occupational skills, occupational earnings, occupational SEI, and individual income variables. We construct occupational skill variables that are commonly used in the literature, including verbal, interactive, computer, science-engineering, creative, quantitative, supervising, and analytical skills. We extract these occupational skills following the standard practice in the literature by conducting factor analysis (Cheng et al., 2019; Liu and Grusky, 2013). For those who report a non-missing occupation, we also merge occupational earnings—measured as the average income within an occupation as reported in the 1990 Census—into the data set. For occupational socioeconomic status, we used gender-specific Hauser-Warren SEI scores and Fredricks Hauser-Warren-equivalent scores updated to the 1980 and 2002 occupational codes (Warren et al., 1998; Stevens and Cho, 1985; Frederick and Hauser, 2010). Individual income variables include log total income, log hourly wage, log total family income measured at each survey round.

**Set B2:** The last set of predictions describes the formation of the family between the ages 25 and 35. These variables include marital status, the number of children, and the age of the children if the respondent has children at each survey round. We include this set as recent research suggests that events occurring in other domains of life, such as the family domain, could impact one’s attainment in the labor market (Budig and England, 2001; Cheng, 2014; Killewald, 2013).

**Set C:** We construct C set as the outcome set to measure one’s mid-life SES attainment. These variables include the average log total income, average log hourly wage, average log family income, and average occupational SEI between ages 40 and 50. We choose this specific age range for constructing outcomes of interest as past research shows that for most individuals, this period is characterized by peak or stabilized income levels, and it also shows the strongest correlation with one’s lifetime earnings (Haider and Solon, 2006; Mazumder, 2005). In the main analysis, we focus on predicting income and SEI. We also present results from an additional analysis, where we predict respondents’ probability of
employment between ages 40 to 50.

To address panel data attrition and missing data, we restrict our analysis sample in several ways. First, to mitigate the impact of sample attrition on the sparsity of life course data, we include respondents in the analysis sample only if they provide a minimum of 5 data points in the B set (out of 8 survey rounds, which dropped 9 percent of the sample) and 2 data points in the C set (out of 6 survey rounds, which dropped another 16 percent of the sample). To maintain a maximum number of observations, we conduct missing data imputation. For predictor sets on family background and early childhood and adolescent experiences, we change the missing value in the original variable to the minimum value in the variable and create a flag variable for each change, following the practices of Killewald and Gough (2013). For predictors that cover one’s early adulthood labor market and family formation experiences, we fill in missing value with non-missing value from the closest survey year between ages 25 to 35. All income variables are inflation-adjusted to 2018 US dollars before log transformation. For those with 0 income, we added 1 before log transformation. The analytic sample contains 7,463 respondents who have complete data for the variables of interest (described below) after missing data imputation and satisfy the sample restriction criteria above.

Machine learning models and evaluation metric

In the analysis, we use various machine learning models to perform the life course outcomes prediction tasks. We start with using set A1 as the prediction set, and then sequentially add set A2, set B1, and set B2 into the model to predict the outcome variables in set C. We partition the analysis sample into two subsets. The training set contains 70 percent of the data while the test set contains 30 percent. We then run models on the training set and report R-squared value from the test set as a common evaluation metric for comparison of fit across models. Before running the machine learning models, we first run a benchmark model that uses the mean of the outcome variable from the training set as the predicted
value (hereafter referred to as the “null model”). In addition, to compare the performance of the more complex machine learning models to that of the traditional methods, we also run a simple ordinary linear model (OLS) and report statistics obtained from it.

We rely on two kinds of machine learning methods to perform life-course prediction tasks. The first kind is shrinkage regressions, namely the Ridge and Lasso regression methods. Ridge regressions are similar to the OLS, except that the coefficients are estimated under a shrinkage penalty \( \lambda \) that shrinks the coefficients toward zero. With this shrinkage penalty parameter, Ridge regression’s advantage over OLS is rooted in the bias-variance trade-off. At \( \lambda = 0 \), the Ridge regression is equivalent to the OLS, where the variance is high but there is no bias. As \( \lambda \) increases, shrinking the Ridge coefficients leads to a substantial reduction in the predictions’ variances but at the expense of an increase in bias. Ridge regression outperforms OLS in situations where there is high variance, where a small change in the training data set can lead to a large change in the least-squares estimates, especially when the number of variables is as large as or even bigger than the number of observations. While the Ridge shrinks the coefficients towards but never reaches zero, the Lasso regression has a variable selection property by shrinking some of the coefficients to zero. This means that the Lasso regression is likely to produce simpler and more interpretable models compared with the Ridge, if the outcome is better characterized as a function of a small set of predictors. However, in terms of prediction accuracy, they can return similar results due to qualitatively similar model behaviors (i.e., as \( \lambda \) increases, the variance decreases and the bias increases). The shrinkage methods are helpful in our case, as one of our research goals is to expand the number of predictors in each life stage. We use cross-validation for selecting the tuning parameter \( \lambda \).

The second set of models consists of tree-based ensemble methods, including random forest, gradient boosting, and Bayesian additive regression trees (BART). These models aggregate multiple regression trees, which stratify or segment the predictor space into smaller and simpler regions in order to achieve the greatest possible reduction in the residual sum of
squares. In each region, the regression tree model then uses the mean of the outcome variable for the training observations as prediction values for the test observations that fall within the same regions. Thus, tree-based methods are likely to outperform models based on linear regression in cases where the relationship between the outcome and predictor variables are highly non-linear and complex. However, individual trees can be sensitive to minor variations in the data, leading to significant changes in the final estimated tree model. To overcome this disadvantage, tree-based ensemble methods combine a large number of “weak learner” trees into a more robust and potentially powerful single model. Here, we focus on three commonly used ensemble methods: the random forest (which averages regression trees that only pick a random sample of predictors through bootstrap sampling), boosting (growing trees sequentially by fitting new trees to the residual of the current fit), and Bayesian additive regression trees (BART; the model grows trees successively but perturbs each tree to avoid local minima). We explore the tuning parameters through cross-validation.\footnote{We fine-tune these models using cross-validation. For the random forest model, we set the tuning parameter to the square root of the total number of predictors. We set it to this typical value, as changing the tuning parameter to other values does not lead to significantly better model performance. For boosting, we picked the number of trees by cross-validation while setting the tree depth and step size of boosting to be relatively small (2 and 0.3 respectively). For BART, we used 200 trees, 1000 iterations, and 100 burn-in iterations. We picked these typical values as BART performs well with minimal tuning.}

In summary, we will benchmark the performance of machine learning models against that of a “null model” and a traditional OLS model. Machine learning models have the advantage over OLS by accommodating a broader array of predictors, while also capturing non-linear relationships and interactions among these predictors. Techniques like Ridge and Lasso regression, known for shrinkage or variable selection, become particularly advantageous as the number of predictors grows. Additionally, tree-based methods prove valuable for their capacity to handle complex model specifications, including non-linearity and interaction terms with flexibility. In our analysis, we will compare the performance of various machine learning models to elucidate which features contribute most significantly to enhancing predictive accuracy.

Following common practice, we report the test $R^2$ as our metric to evaluate predictive
performance:

\[ R^2_{\text{test}} = 1 - \frac{\sum_{i \in \text{test}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{test}} (y_i - \bar{y}_{\text{training}})^2} \]  

(1)

where \( \bar{y}_{\text{training}} \) is the mean of the outcome variable in the training data set obtained from the null model. The test R squared is one of the commonly used evaluation metric in the prediction literature (Salganik et al., 2019). In our case, this measure is particularly helpful for comparison across outcomes, as it does not depend on the variance of the outcome variables.

Predicting model-based distributions

In the second part of the analysis, we illustrate how the predictive framework can be used to study group disparities by race and gender. Here we combine machine learning models with decomposition methods such as those used in Kitagawa (1955), Duncan (1968), Blinder (1973), and Oaxaca (1973). Researchers use such decomposition methods to study group disparities such as the racial/gender wage gap. It proceeds by 1) running linear or generalized regression models on stratified models by group membership (race/gender), and 2) simulating counterfactual statistics by switching the variable distributions or predictive models to an alternative group. In this way, researchers are able to decompose group disparity into a compositional component, the amount of the group disparity driven by differences in groups’ characteristics, and a structural component, the amount driven by differences in how these characteristics are “paid off” in e.g., a labor market. We make three adaptations in our application of such methods. First, instead of using simple linear regressions, we use non-parametric machine learning models, which are highly flexible and releases assumptions in model functional forms. In addition, machine learning models accommodate the large feature space we have in our case. Second, instead of having all the covariates in the single
model (the approach used by Kitagawa, Binder, and Oaxaca), we follow Duncan’s approach by adding the prediction sets sequentially into the model and conduct counterfactual simulations at each step (see also, a recent development and application in Opacic et al. (2023)). This is more suitable for our case and consistent with our analysis in the first part, as we are interested in studying the cumulative influence of life course experiences according the temporal order. Third, besides focusing on the changes in group disparities in expectations (i.e., group means), we predict model-based distributions of the outcome of interest and also highlight changes in specific percentiles of the distribution such as the the 10th and 90th percentiles. We use the regression imputation estimator to simulate the distributions.

The analysis of group disparities follows three steps. To take the analysis of Black-White family income as an example, first, we predict counterfactual outcomes for black individuals, assuming they have the same underlying predictive model as whites. This is implemented by applying the prediction model learned from the white training data to the covariates of the blacks’ sample in the test data. We use $Y^{WMBX}$ to denote this counterfactual outcome, where the notation $WMBX$ stands for “white model from training set, black X from test set.” Second, we predict counterfactual outcomes for black individuals assuming their distribution of covariates ($X = \{X_1, X_2, \ldots, X_K\}$) are the same as whites. We use $Y^{BMWX}$ to denote this counterfactual outcome, where the notation $BMWX$ stands for “black model, white X.” Lastly, we predict counterfactual outcomes ($Y^{WMWX}$) for black individuals, assuming their distribution of covariates are the same as whites and they have the same underlying predictive model as whites.

Formally, let $f_{white}(\cdot)$ and $f_{black}(\cdot)$ denote the prediction model for whites and blacks’ outcomes, respectively. We will use the best performing predictive model (in our case, the BART), as our preferred prediction models. For an individual possessing values from a specified set of covariates $X$, the function $f_{white}(X)$ calculates the anticipated average outcome for this individual, assuming that the impact of these covariates on the outcome is governed by mechanisms observed within the white sample. Further, we use $X_{white}$ and $X_{black}$
to denote covariates distribution drawn from the white subsample and black subsample, respectively. Therefore, $Y^{BM}B$ represent the predicted outcome for blacks using their own covariates distribution and prediction model. Following these definitions, we can write the three counterfactual outcomes described in the previous paragraph as:

$$Y^{WM}B = f_{white}(X_{black}) \quad (2)$$

$$Y^{BM}W = f_{black}(X_{white}) \quad (3)$$

$$Y^{WM}W = f_{white}(X_{white}) \quad (4)$$

In our empirical results, we present the estimated density curves of these predicted counterfactual distributions. We also compute and compare specific features of these distributions, such as the mean and the 90th and 10th percentiles.

Results

Predictability of life course outcomes

How well can mid-life SES be predicted? Which life stage predicts one’s socioeconomic status attainment well? Do results vary by prediction model? Table 1 presents our evaluation metric, test R squared, for two selected outcomes, log hourly wage and log total income. We make three main observations from Table 1. First, the predictability of life course outcomes depends on which outcome we examine. Prediction accuracy is significantly higher for the average log hourly wage between age 40 and 50 than for total income for the same age range. For example, the best R squared statistics in Panel 1, from models that use tree-based assemble methods and sets A1, A2, and B1, is 0.53. However, the best R squared in Panel 2, which comes from the same model and set combination, is only about 60 percent (0.33). Among the best performing machine learning methods, the test R squared from only using set A1 (e.g., Panel 1, 0.25 from BART) exceeds the R squared from using sets A1 and A2 combined (Panel 2, 0.15, Bart). Overall, the predictability for total income ranges
from about 30 to 60 percent of that for hourly wage. Thus, in terms of prediction accuracy, mid-life hourly wage turns out to depend more strongly on previous life experiences than total income.

Second and as expected, predictability varies across prediction sets and increases as we add prediction sets by life course order. Across the two panels for these two different outcomes, a consistent pattern emerges: set B1 appears to be the most powerful set of predictors. This is not surprising, as set B1 characterizes early adulthood labor market experiences, and include variables like employment status and intensity, occupational characteristics (such as skills, earnings, and SEI), as well as individual level income data at each survey round between ages 25 and 35. Take the gradient boosting columns as examples. For log hourly wage, using set A1 alone as predictors explains 24 percent more of the variance in the test data set, compared with a simple baseline model that uses the training set mean as predicted values. This suggests that knowing about individuals’ family background does improve our prediction about their mid-life SES in the later life stage. Adding set A2 into the model, the test R squared increases from 0.24 to 0.35, leading to an improvement of prediction accuracy by 46 percent. We then add set B1 into the model that already contains sets A1 and A2, which covers up to formal labor market entry (here, as defined by age 25). This leads to an increase of the test R squared to 0.53, an even larger improvement in prediction accuracy (about 51 percent improvement in prediction accuracy). Lastly, adding set B2, which contains variables on early adulthood family formation situations, does not lead to further change in test R squared. We observe the same pattern for predicting log total income. Although the level of test R squared is significantly lower, adding set A2 improves prediction accuracy by about 67 percent, while adding set B1 improves it further to 2 times, and adding set B2 does not change the results much further.

The third main observation arises from comparing horizontally across models. Consistent with expectations, the simple OLS model performs similarly to the more complex machine learning models when prediction sets are relatively small. For example, when the
prediction set only includes set A1, which contain fewer than 500 features, its test R squared is only marginally lower than the more complex machine learning methods. However, when the number of features dramatically increase to over 4000 after we add sets A2 and B, the OLS cannot match shrinkage methods like ridge and lasso. This is understandable, given that the shrinkage methods are designed to improve prediction accuracy by giving more weights to predictors that contribute the most to the model, while penalizing or ‘shrinking’ the coefficients of less important variables. This helps with reducing overfitting and improving model performance. In addition, in these cases, the ensemble methods, which allow complex non-linear relationships and interactions—perform only marginally better than shrinkage methods. Comparing across linear shrinkage methods and flexible tree-based methods suggests that the prediction task is improved primarily through expanding the number of variables in the prediction sets, rather than incorporating complex non-linear relationships.

What do these prediction models tell us about the level of social rigidity for this cohort of people overall? We consider a stratification system highly rigid if mid-life outcomes can be accurately predicted using variables commonly referenced in the literature. The scatterplots in Figure 2 presents a more comprehensive illustration across all outcomes and prediction sets, particularly highlighting results by BART, one of the best performing models across outcomes. The data reveals that predicting total income and family income poses the biggest challenge, while hourly wage and occupational socioeconomic status (“SEI”) are relatively more predictable. Overall, the predictability varies widely, ranging from as low as below 0.1, to as high as just over 0.5.

Consistent with Table 1, Figure 2 also suggests that adding set B1, which includes early adulthood labor market experience, into the prediction set, leads to the largest increase in test R squared. As aforementioned, this is not surprising as B1 includes directly many variables
that could be considered as “lag variables” to outcomes in set C. For example, for outcomes that include hourly wage, total income, and family income, the “lag variables” include all individual level wage and income variables. For predicting SEI, the “lag variables” include all “SEI” variables in set B1. Consistent with past studies, we find that these “lag variables” hold the largest predictive power (Kim et al., 2018). But how much of the relatively high predictability of set B1 is attributable to these “lag variables?” In Panel B of Figure 2, we drop those variables from the prediction sets and redo the prediction tasks. The transparent dots and dashed lines present the updated results after dropping those “lag variables.” We see that dropping those variables does decrease the test R squared for models that include set B1. When using all variable sets to predict log hourly wage, dropping all individual level income variables leads to a 15 percent decrease in test R squared (from .53 to .46), and for log family income, it leads to a 12 percent decrease (from .38 to .34). For predicting mid-career occupational SEI, dropping all SEI variables from set B1 does not decrease the test R squared as dramatically. This suggests that the majority of variation in mid-life occupational status is predominantly accounted for by other variables, such as individual income levels and occupational characteristics (e.g. occupational earnings and occupational skills). Interestingly, omitting all individual income variables from the models has minimal impact on the predictions for mid-life log total income. This could be attributed to the inherent difficulty in predicting total income, even when these ‘lag variables’ are included in the analysis. Notably, Figure 2 suggests that there is considerable level of social fluidity as the test R squared ranges from 0.3 to 0.53 across the four outcomes, when using all the prediction sets combined together to do the prediction task, even after including all the “lag variables.”

The high social fluidity among this cohort may be partially attributed to their experiencing the 2008 economic recession during mid-life. We consider this as a significant aspect of the cohort’s life experience. To further investigate, we also ran similar tasks to predict the same average outcomes between ages 36 to 44, covering pre-recession periods. The test R squared increased by about 10 percentage points. This could be due to the removal of "recession effect" or the fact that the outcomes were measured earlier thus closer to the prediction sets in terms of time.
Using the predictive approach to study between-group disparities

In this section, we use two examples to demonstrate the application of our predictive framework in analyzing sources of between-group disparities in mid-life SEI. We begin by considering race (specifically, Black and White) as the group membership variable, with a focus on the average log family income between age 40 and 50 as the primary outcome of interest. Following the approach of classic decomposition analysis, we use counterfactual simulations to assess the amount of group disparity that is explained by the composition of covariates (i.e., switching the covariate distribution among different groups) and the amount explained by the influence of covariates on the outcome (i.e., switching the prediction model between groups).

Is the Black-White difference in the distributions of mid-life income attainment mainly driven by the racial differences in the composition of characteristics, or the racial differences in the effects of these characteristics? Figure 3 presents results obtained from the BART model, with prediction sets added sequentially across the panels, moving from left to right. For example, the first panel shows results from a model incorporating only set A1, and the second panel shows results from a model that includes both sets A1 and A2. On the X axis, “BMBX” stands for distributions from prediction models trained on Blacks’ training set, and predicted from using Blacks’ test set. Similarly, “WMBX” stands for distributions from models trained on Whites’ training set, and predicted from using Blacks’ test set; “BMWX” stands for distributions from models trained on Blacks’ training set, and predicted from using Whites’ test set; and “WMWX” stands for distributions from models trained on Whites’ train set, and predicted from using Whites’ test set. The horizontal white lines represents the 10th percentile distribution, mean, and the 90th percentile distribution, from top to bottom, and they are connected by dashed red lines. In summary, we focus on not only the changes in expected values (group means), but also changes at the lower and upper ends of the distributions. Comparing the panels in this figure enables us to evaluate the distinct

4Additional analyses on other outcome variables are detailed in the Appendix.
roles that various life stages play in predicting the distributions of mid-life socioeconomic status.

A few patterns emerge from Figure 3. First, the relative contributions of the covariates composition and effects depend on the specific life stage under examination. The first two panels in Figure 3 suggest that when we only use pre-labor-market variables in the prediction, the two factors play equally important roles. This is evident regardless of whether we focus on means (the middle line) or the 10th and 90th percentile points (top and bottom lines), as both "WMBX" and "BMWX" columns show similar distributions. However, when set B variables on labor market attachment, occupational characteristics, income, and adulthood family formation are included, Black-White differences in covariate composition turn out to play a bigger role in explaining the racial disparity in mid-life racial disparity in family income distributions. Compared with a model that switches Blacks’ prediction model with Whites’ model (while maintaining Blacks’ covariates), a model that switches Blacks’ covariates with Whites’ covariates (while keeping Black’s prediction model) leads to a much more dramatic upwards shift in the simulated distribution. In other words, “BMWX” model brings the simulated distribution closer to that of the last distribution (“WMWX”). This indicates that racial differences in early adulthood characteristics, especially in terms of labor market experiences, are critical for explaining the racial differences in family income between ages 40 and 50. Moreover, this pattern is especially pronounced among low-income Blacks. In the last panel, which includes all variables, the steep slope of the red, dashed line connecting the 10th percentile points across distributions highlights this finding.

In summary, family background and early life experience and how they are translated into mid-life SES attainment are key in predicting the gap in Blacks’ and Whites’ average log family income between age 40 and 50. However, by early adulthood, the racial gap in characteristic distributions characteristics plays a more significant role in predicting mid-life socioeconomic status. This implies that the differential impact of family background and early life experiences has already manifested in measurable racial inequality in their early
adulthood situation, such as employment attachment, occupations, early career income, marital status, number of children, etc., which further contributes to racial inequality in mid-life SES attainment. This finding is particularly relevant for understanding the life course production of inequality for low-income Blacks.

[Figure 3 about here.]

In our second example, we apply the same analytical framework to decompose the gender gap in log hourly wage at mid-life into portions attributable to the composition and effects of covariates at different life stages. The first two panels of Figure 4 show that when the model only includes family background and early childhood and adolescence variables, switching women’s characteristics with men’s do not change the simulated distributions much from the original distribution (“WMWX”), and that the gender disparity in mid-life hourly wage is primarily driven by how these personal characteristics are translated into the gender differentiated stratification process. This finding underscores the substantial influence of pre-labor-market mechanisms in determining mid-career gender disparities. However, once we take into account early adulthood experiences, the gender difference in the distributions of the covariates appears to be as important as how women and men get different “returns” to these covariates. Switching either the covariates by gender or the model leads to a similar level of upward shift in the simulated distributions, while changing both at the same time leads to the largest jump (“MMMX”).

[Figure 4 about here.]

**Additional analysis 1: Predicting the probability of employment**

In the main analysis, we focused on predicting mid-life income and SEI attainments. This means that the analytical sample was restricted to respondents with at least one non-missing income and occupation value in survey waves between ages 40 and 50. This criterion may lead to the exclusion of respondents who are consistently unemployed. To the extent that the
life chances of employed individuals tend to be less volatile and hence more predictable, the results from main findings can be considered as an upper bound for our ability to predict the cohort’s life chances. To supplement the main analysis, we conduct an additional analysis on predicting the probability of employment during age 40 and 50, where we expand the main analytical sample to also include those who are consistently unemployed and create an indicator for the probability of employment. For each respondent, we calculate the probability of employment as a continuous variable, defined as the number of survey waves in which they were employed divided by the total number of survey waves with non-missing data\textsuperscript{5}. We then conduct similar analysis using both the OLS and more complex machine learning models. We evaluate the model performance using the same metric, test R squared.

Table 3 presents results from this supplementary analysis. Overall, the results are consistent with our main findings with regard to patterns across prediction sets and models. While only using family background and early childhood and adolescent variables leads to a test R squared of below 0.1, adding labor market variables significantly improves prediction accuracy by almost doubling the test R squared. However, the general level of predictability is lower than all the previous outcomes. This suggests that employment status among the general population is harder to predict than income or occupational status among the employed workforce. While the OLS model is widely off when the feature space expands significantly after including set B1, the ML model test R squared hovers right below 0.3. Overall, results from this supplementary analysis align with our main conclusion that the predictability of life course outcomes is generally low to moderate.

\textsuperscript{5}The maximum number of non-missing survey waves is 6, as NLSY79 switched to biennial survey schedule after 1994.
Additional analysis 2: A closer look into important variables

We conduct another supplementary analysis to identify the most influential variables for predicting mid-life SES attainment, using the random forest method. The importance of each predictor is computed based on changes in the outcome variables' variances due to splits over the given predictor, averaged over all trees. Figure 5 presents the top 20 most important variables across prediction sets, scaled relative to the maximum value within each panel. The top left panel shows results from a model that only includes the prediction set on family background. Out of 99 predictors (the relatively large number of features is due to the transformation of some variables into dummy variables, such as the 12-category classification of parents’ occupational attainments), the variables that contribute most significantly to the reduction in the outcome’s variance include parents’ educational attainment, SEI, age, gender, sibship size, region, family poverty status, and race. This is consistent with the stratification literature on intergenerational social mobility, as these variables are among the most commonly used variables in this field of study. Similarly, the top right panel presents results from the model that uses all pre-labor-market predictors. When the model includes both A1 and A2, it identifies the AFQT score, which measures respondents’ cognitive abilities, as the most powerful predictor. Variables that are ranked next on the list, such as the Rosenberg score, educational aspirations and attainments, have only about 30 to 40 percent of the importance value, relative to AFQT's value on the index. Besides the commonly used stratification variables, a few high school indices that measure the school contexts (such as number of library books, teachers' salary) also emerge among the top 20 most important variables out of over 500 features. Lastly, the last bottom left panel presents results from a model that uses all predictors across all life course stages. Consistent with the previous results indicating the set B1 (on labor market experiences) as the most important predictor set for predicting mid-life SES, the model predominantly selects set B1 variables as the most influential (from out of over 4000 features). In addition, results suggest that

---

6The results are similar whether we include or exclude set B2 (adult family formation).
income/wage variables from later waves within the set B time range (ages 25-35) appear more predictive for mid-life SES attainment. The bottom left panel also reaffirms the significance of AFQT score as an influential predictor in the model.

[Figure 5 about here.]

Although it is valuable to pinpoint important predictors, such as wage and income variables in B1 and the AFQT cognitive score in A2, we must be cautious when interpreting these as indicators of strong causal effects of these variables themselves. In fact, it is plausible that these variables themselves are determined by upstream variables, such as family background in A1. Consequently, they may primarily serve as mediators, channeling the causal effects of preceding variables, rather than being direct causal agents themselves. How well can we predict these important intermediate variables? To answer this question and get a sense of the endogeneity of these important intermediate variables, we conduct a similar exercise aimed at predicting these variables. Table 4 presents the results. As anticipated, a considerable amount of variation in these important intermediate variables can be explained by variables that measure one’s family background and other pre-labor market characteristics, with the test R squared ranging from 0.24 to 0.45.

[Table 4 about here.]

**DISCUSSION**

How persistent is socioeconomic status across generations and over the life course? In this project, we set out to answer this question by reconsidering “social rigidity” under a predictive framework. Our research extends beyond the conventional approach of examining variations within samples. By using machine learning techniques, we focus on the ability to predict life outcomes in a new sample. Higher prediction accuracy implies a robust persistence of socioeconomic status, indicating a greater level of social rigidity. Compared to previous
approaches to studying social rigidity, this predictive methodology offers three primary adv-
antages: it accommodates the integration of an extensive array of mediating variables; it
grants versatility in non-linear forms; and it facilitates the inclusion of interactions across a
broad range of variables.

We apply our predictive framework to the NLSY79 data, a widely recognized longitudi-
dinal dataset in the stratification literature. The data set encompasses a rich array of social
factors spanning the life course trajectories of late Baby Boomers. We select variables com-
monly used in previous stratification research and group them into four distinct prediction
sets that characterize both intergenerational and intragenerational aspects of social rigidity:
family background, childhood and adolescence experiences, early adulthood labor market
experiences, and early adulthood family formation processes. We employ machine learn-
ing models to predict one’s mid-life SES by incrementally adding prediction sets following
the chronological order of the life course. We use out-of-sample test statistics, such as the
test R squared statistics, to quantify the predictive power of different models. Our analysis
contributes to the literature in several ways, as we discuss below.

First, how predictable is mid-life SES? Using state-of-art machine learning models to
explore the complex relationship between the predictors and outcome variables, we find that
a predictors set that includes as comprehensive as 4,000 variables can explain about 10 to
50 percent more variance in mid-life SES than a baseline model that uses the mean of the
outcome variable in the training set as a simple prediction baseline. The level of predictability
depends on the specific outcome being analyzed, with labor market indicators like wages and
occupational prestige being more predictable than broader socioeconomic measures such as
overall personal and family income. Predictability increases as we sequentially incorporate
prediction sets that align with the life course’s temporal sequence. Overall, the results reflect
a moderate degree of predictability of mid-life socioeconomic status in the United States,
with the best performing models predicting just above 50% of the variations in hourly wage.

On the one hand, the test R squared from our study’s test R squared value surpasses
those documented in earlier research. For instance, the Fragile Families Challenge reports the highest test R squared at approximately 0.2 for predicting material hardship and GPA at age 15, and this metric drops to around 0.05 for other outcomes (Salganik et al., 2020). The fact that prior life experience can explain as high as half of the variations of socioeconomic status at mid-life implies substantial persistence of socioeconomic indicators both over the life course and across generations. In addition, our exercise reinforces the credibility of the long-standing social stratification theories as most important variables selected in this data-driven way are consistent with the central mechanisms in the field of study. However, when interpreting our results in comparison with other studies, one should note the important differences in research design. First, we examine outcomes at a different life stage. Mid-life socioeconomic outcomes may be easier to be predicted based on sociological theories, compared with outcomes at adolescence, the target life stage in the Fragile Families Challenge. In contrast to younger ages, individuals’ positions in mid-life tend to exhibit greater structure and stability. This phase of life is influenced by various social institutions that contribute to shaping one’s trajectory, including acquiring educational qualifications, forming family bonds through assortative mating, and embarking on distinct career paths in the labor market. Second, we examine different target populations. Our study is based on a cohort-specific survey that best characterizes the life trajectory of the late Baby Boomers, while the Fragile Families Challenge draws from data that over-sample births to unmarried mothers, resulting in the inclusion of a large number of Black, Hispanic, and low-income families (Reichman et al., 2001). Therefore, difference in the out-of-sample R squared can also stem from the inherent variation in predictability of life course outcomes across social groups. Lastly, we examine different outcomes. Focusing on social rigidity, we select traditional indicators from the stratification literature. As our results imply, the level of predictability depends on which outcome we examine. This is likely the case when comparing across studies. Specifically, income and occupational outcomes might be easier to predict

\[\text{More accurately, the need to examine income and occupational outcomes entails a sample exclusion of those who were consistently unemployed during age 40 to 50.}\]
than outcomes such as student GPA, material hardship, layoff, etc.

On the other hand, our results corroborate that unpredictability might be the norm in life course prediction tasks, considering that the best model can only achieve a moderate level of predictability (Salganik et al., 2020; Badolato et al., 2023). However, one should not rush to dismiss prediction tasks as unproductive, even if the level of predictability is low. The inability of prediction raises many meaningful questions, such as how social science researchers as a collective, should understand the lack of predictivity of important social outcomes; and whether we can improve our ability of prediction by more rigorously theorize and measure the social factors that matter for forecasting important life course outcomes such as a person’s mid-life SES (Garip and Macy, 2023). For example, as our results show, much of the improvement in prediction accuracy stems from the expansion of the predictor sets rather than allowing for complex interactions between variables. This underscores the necessity of incorporating a diverse range of measurements for social concepts and mechanisms in the initial stages of data collection. Such an approach should be informed by the development of theoretical frameworks and the identification of pertinent mechanisms in the existing literature.

Second, we explore variation in out-of-sample prediction accuracy by drawing different sets of predictors from one’s family background, and various progressive life stages. High levels of predictability for mid-life outcomes indicate high levels of social rigidity, i.e., strong continuity of socioeconomic status across generations and across the life course. By sequentially adding predictor sets into the model, we find that while family background and childhood and teenage experiences have limited predictive power for one’s mid-life SES attainment, incorporating early career experiences significantly improves models’ prediction accuracy. Among those predictors, family and individual income and wages, employment status, occupational skills, and hours worked, especially those in one’s late 20s, are particularly strong predictors for mid-life SES. In an additional exercise, we replaced those family and individual income indicators with an average occupational income constructed from the
1990 US Census, and we found that the models are almost equally predictive for the same mid-life SES outcome. This re-affirms our conclusion that early labor market experiences are the most strong predictors for one’s mid-life SES. This suggests that individuals’ early-career job sequences, such as those obtained from large resume data sets, can be useful for predicting later-life outcomes, even if those data sets often do not contain or only include a limited number of pre-labor-market characteristics (e.g. Vafa et al., 2022).

Third, we apply the predictive framework to study sources of group disparity in mid-life SES attainment and decompose it by life stages. We use two examples to illustrate how to use such an approach. For example, we simulate distributions under counterfactual situations where we switch Blacks’ and Whites’ models or characteristics by life stages. We find different patterns before and after including set B. When the model only includes pre-labor-market prediction sets, the racial difference in the distribution and returns of family background and early life experiences similarly drive the racial disparity in mid-life average family income. In contrast, when prediction sets on early adulthood are added into the model, compositional differences in Blacks’ and Whites’ early labor market characteristics appear more important than the returns to these characteristics. These results indicate that social policies that target childhood and adolescence life stages to reduce radical inequalities in early labor market participation are likely crucial for reducing racial inequality at mid-life.

Our study is not without limitations. First, as we draw from a single-cohort survey, our results are necessarily limited to the social rigidity in the life course progression of the late Baby Boomers. Future research should investigate whether the reported level of predictability can be generalized to other birth cohorts, especially as the NLSY97 cohort enters mid-life ages. Second, as we use secondary data that is already collected, we are limited by using variables measured in the survey. There might be important variables that are observed by the individual but not observed by the researcher. Lastly, we are constrained by the relatively small sample size of a longitudinal survey data set (Salganik et al., 2019). Fully leveraging the power of complex machine learning models often requires that the model
learns from a large number of observations, especially when the feature space is big, too. Therefore, it is possible that the lack of improvement from shrinkage methods to complex tree-based methods is due to limited sample size rather than the true lacking of complex non-linear relationships among those predictors. However, this can be remedied by making use of the large online or administrative/registry data sets that are becoming increasingly available to researchers, which also affords the opportunity to combine information across multiple data sources through transfer learning (Savcisens et al., 2023; Vafa et al., 2022). Altogether, applying predictive models to a variety of outcomes and samples can collectively improve our understanding of the intricate process of social stratification.
References


Blau, P. M. and O. D. Duncan (1967). The american occupational structure.


Breen, C. F. and N. Seltzer. Sociodemographic characteristics alone cannot predict individual-level longevity.


Figure 1: Chronological Overview of Prediction and Outcome Sets with Representative Variables
<table>
<thead>
<tr>
<th>Outcome</th>
<th>Family income</th>
<th>Hourly wage</th>
<th>Total income</th>
<th>SEI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aset1</th>
<th>Aset1 + Aset2</th>
<th>Aset1 + Aset2 + Bset1</th>
<th>Aset1 + Aset2 + Bset1 + Bset2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Predictability across Outcomes and Sets

Note: The left panel shows results before dropping “lag” variables, and the right panel shows results after dropping “lag variables” by adding transparent dots and dashed lines. For outcomes that include hourly wage, total income, and family income, the “lag variables” include all individual level wage and income variables in set B1 and thus subsequent modeling. For predicting SEI, the “lag variables” include all “SEI” variables in set B1 and subsequent modeling.
Figure 3: Predicting Log Family Income Distributions by Race

Notes: 1) Results are obtained from the BART model. 2) We add prediction sets sequentially in the panels, from left to right. For example, the first panel shows results from a model that only contains set A1, and the second panel shows results from a model that contains both sets A1 and A2, so on and so forth. 3) In the X axis, BMBX stands for distributions from models trained on Blacks’ training set, and predicted from using Blacks’ test set. Similarly, WMBX stands for distributions from models trained on Whites’ training set, and predicted from using Blacks’ test set; BMWX stands for distributions from models trained on Blacks’ training set, and predicted from using Whites’ test set; and WMWX stands for distributions from models trained on Whites’ train set, and predicted from using Whites’ test set. 4) The horizontal white lines represent the 10th percentile distribution, mean, and the 90th percentile distribution, from top to bottom. They are connected by the red dashed lines within each panel.
Figure 4: Predicting Log Hourly Wage Distributions by Gender

Notes: 1) Results are obtained from the BART model. 2) We add prediction sets sequentially in the panels, from left to right. For example, the first panel shows results from a model that only contains set A1, and the second panel shows results from a model that contains both sets A1 and A2, so on and so forth. 3) In the X axis, WMWX stands for distributions from models trained on women’s training set, and predicted from using men’s test set. Similarly, MMMX stands for distributions from models trained on men’s training set, and predicted from using women’s test set; WMMX stands for distributions from models trained on women’s training set, and predicted from using men’s test set; and MMMX stands for distributions from models trained on men’s train set, and predicted from using men’s test set. 4) The horizontal white lines represent the 10th percentile distribution, mean, and the 90th percentile distribution, from top to bottom. They are connected by the red dashed lines within each panel.
Figure 5: Top 20 Most Important Predictors Used in Random Forest to Predict Log Hourly Wage, across Prediction Sets

Notes: Variable importance is computed using the mean decrease in the variance of the responses, and expressed relative to the maximum within each panel.
<table>
<thead>
<tr>
<th>Analytic Framework</th>
<th>Analytic Tools (examples)</th>
<th>Quantities of interest</th>
<th>Number of variables</th>
<th>Model Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Classic status attainment model</td>
<td>path diagram; structural equation models; mobility table methods</td>
<td>total effects; partial effects; direct and indirect effects; in-sample $R^2$</td>
<td>Limited</td>
<td>Typically linear models, no explicit interaction</td>
</tr>
<tr>
<td>(b) Conventional mediation analysis</td>
<td>Directed acyclic graph; linear regression models</td>
<td>Natural direct effect; natural indirect effect</td>
<td>Limited</td>
<td>Typically linear models with two-way interactions</td>
</tr>
<tr>
<td>(c) Machine-learning-based mediation analysis</td>
<td>Directed acyclic graph; machine learning methods</td>
<td>Natural direct effect; natural indirect effect</td>
<td>Moderate</td>
<td>Flexible non-linear models with complex interactions</td>
</tr>
<tr>
<td>(d) Sequential mediation analysis</td>
<td>Multi-stage directed acyclic graph of variables; dynamic treatment regimes and weighting methods; machine learning methods</td>
<td>Stage-specific effects</td>
<td>Spanning across multiple periods</td>
<td>Flexible non-linear models with complex interactions</td>
</tr>
<tr>
<td>(e) Our predictive framework</td>
<td>Multi-stage directed acyclic graph of wholistic sets of variables; machine learning methods</td>
<td>Prediction accuracy (out-of-sample $R^2$); predictive group-specific distributions</td>
<td>Large number of variables, variable sets spanning across multiple periods</td>
<td>Flexible non-linear models with complex interactions</td>
</tr>
</tbody>
</table>
Table 2: Comparison of Test R-squared for Log Hourly Wage and Log Total Income across Models.

<table>
<thead>
<tr>
<th>Methods</th>
<th>OLS</th>
<th>Ridge</th>
<th>Lasso</th>
<th>Random Forest</th>
<th>Gradient Boosting</th>
<th>BART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel 1: Log hourly wage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A1</strong></td>
<td>0.22</td>
<td>0.23</td>
<td>0.23</td>
<td>0.24</td>
<td>0.24</td>
<td>0.25</td>
</tr>
<tr>
<td>+ <strong>A2</strong></td>
<td>0.31</td>
<td>0.34</td>
<td>0.35</td>
<td>0.33</td>
<td>0.35</td>
<td>0.36</td>
</tr>
<tr>
<td>+ <strong>B1</strong></td>
<td>−1.21</td>
<td>0.47</td>
<td>0.51</td>
<td>0.52</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>+ <strong>B2</strong></td>
<td>−1.34</td>
<td>0.47</td>
<td>0.51</td>
<td>0.52</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>Panel 2: Log total income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A1</strong></td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>+ <strong>A2</strong></td>
<td>0.08</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>+ <strong>B1</strong></td>
<td>−1.84</td>
<td>0.30</td>
<td>0.31</td>
<td>0.31</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td>+ <strong>B2</strong></td>
<td>−1.87</td>
<td>0.31</td>
<td>0.32</td>
<td>0.32</td>
<td>0.33</td>
<td>0.33</td>
</tr>
</tbody>
</table>

*Note:* Prediction accuracy is measured by test set R-squared.
Table 3: Test R-squared for Probability of Employment across Models.

<table>
<thead>
<tr>
<th>Methods</th>
<th>OLS</th>
<th>Ridge</th>
<th>Lasso</th>
<th>Random Forest</th>
<th>Gradient Boosting</th>
<th>BART</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1</strong></td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>+ <strong>A2</strong></td>
<td>0.07</td>
<td>0.12</td>
<td>0.12</td>
<td>0.11</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>+ <strong>B1</strong></td>
<td>-2.14</td>
<td>0.26</td>
<td>0.28</td>
<td>0.27</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>+ <strong>B2</strong></td>
<td>-2.18</td>
<td>0.26</td>
<td>0.28</td>
<td>0.28</td>
<td>0.29</td>
<td>0.29</td>
</tr>
</tbody>
</table>

*Note:* Prediction accuracy is measured by test set R-squared.

Table 4: Predicting Intermediate Variables Using Upstream Predictors

<table>
<thead>
<tr>
<th>ImpVar</th>
<th>Prediction set</th>
<th>Test R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFQT (A2)</td>
<td>A1</td>
<td>0.43</td>
</tr>
<tr>
<td>Average log hourly wage (B1)</td>
<td>A1 + A2</td>
<td>0.35</td>
</tr>
<tr>
<td>Average log total income (B1)</td>
<td>A1 + A2</td>
<td>0.24</td>
</tr>
<tr>
<td>Average log family income (B1)</td>
<td>A1 + A2</td>
<td>0.26</td>
</tr>
<tr>
<td>Average SEI (B1)</td>
<td>A1 + A2</td>
<td>0.45</td>
</tr>
</tbody>
</table>

*Note:* Prediction accuracy is measured by test set R-squared.  
Appendix: Additional Decomposition Results

Table 1: Decomposition of Racial Disparity in Mid-life Log Hourly Wage and Log Total Income.

<table>
<thead>
<tr>
<th></th>
<th>Aset1 Mean</th>
<th>P10</th>
<th>P90</th>
<th>+Aset2 Mean</th>
<th>P10</th>
<th>P90</th>
<th>+Bset1 Mean</th>
<th>P10</th>
<th>P90</th>
<th>+Bset2 Mean</th>
<th>P10</th>
<th>P90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel 1: Log hourly wage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMBX</td>
<td>2.32</td>
<td>2.02</td>
<td>2.64</td>
<td>2.32</td>
<td>1.92</td>
<td>2.78</td>
<td>2.30</td>
<td>1.83</td>
<td>2.84</td>
<td>2.29</td>
<td>1.84</td>
<td>2.88</td>
</tr>
<tr>
<td>BMWX</td>
<td>2.57</td>
<td>2.22</td>
<td>2.94</td>
<td>2.57</td>
<td>2.10</td>
<td>3.03</td>
<td>2.58</td>
<td>2.02</td>
<td>3.16</td>
<td>2.59</td>
<td>2.01</td>
<td>3.18</td>
</tr>
<tr>
<td>WMBX</td>
<td>2.39</td>
<td>1.99</td>
<td>2.79</td>
<td>2.41</td>
<td>1.97</td>
<td>2.86</td>
<td>2.33</td>
<td>1.91</td>
<td>2.83</td>
<td>2.34</td>
<td>1.90</td>
<td>2.90</td>
</tr>
<tr>
<td>WMWX</td>
<td>2.62</td>
<td>2.20</td>
<td>3.02</td>
<td>2.59</td>
<td>2.10</td>
<td>3.08</td>
<td>2.59</td>
<td>2.03</td>
<td>3.20</td>
<td>2.59</td>
<td>2.02</td>
<td>3.21</td>
</tr>
<tr>
<td>Panel 2: Log total income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMBX</td>
<td>7.74</td>
<td>6.36</td>
<td>9.01</td>
<td>7.70</td>
<td>5.91</td>
<td>9.40</td>
<td>7.64</td>
<td>4.43</td>
<td>9.97</td>
<td>7.67</td>
<td>4.36</td>
<td>9.92</td>
</tr>
<tr>
<td>WMWX</td>
<td>8.62</td>
<td>7.18</td>
<td>9.96</td>
<td>8.54</td>
<td>6.90</td>
<td>10.10</td>
<td>8.49</td>
<td>5.74</td>
<td>10.31</td>
<td>8.50</td>
<td>5.67</td>
<td>10.40</td>
</tr>
</tbody>
</table>

*Note:* Prediction accuracy is measured by test set R-squared.

Table 2: Decomposition of Gender Disparity in Mid-life Log Family Income and Log Total Income.

<table>
<thead>
<tr>
<th></th>
<th>Aset1</th>
<th>+Aset2</th>
<th>+Bset1</th>
<th>+Bset2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>P10</td>
<td>P90</td>
<td>Mean</td>
</tr>
<tr>
<td>Panel 1: Log family income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMMX</td>
<td>10.13</td>
<td>9.11</td>
<td>10.99</td>
<td>10.08</td>
</tr>
<tr>
<td>MMWX</td>
<td>10.09</td>
<td>8.96</td>
<td>10.99</td>
<td>10.23</td>
</tr>
<tr>
<td>WMMX</td>
<td>10.19</td>
<td>9.17</td>
<td>11.02</td>
<td>9.98</td>
</tr>
<tr>
<td>WMWX</td>
<td>10.14</td>
<td>9.05</td>
<td>11.00</td>
<td>10.11</td>
</tr>
<tr>
<td>Panel 2: Log total income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMMX</td>
<td>8.67</td>
<td>7.07</td>
<td>10.00</td>
<td>8.63</td>
</tr>
<tr>
<td>MMWX</td>
<td>8.58</td>
<td>6.94</td>
<td>9.95</td>
<td>8.73</td>
</tr>
<tr>
<td>WMMX</td>
<td>7.80</td>
<td>6.46</td>
<td>8.86</td>
<td>7.47</td>
</tr>
<tr>
<td>WMWX</td>
<td>7.75</td>
<td>6.39</td>
<td>8.86</td>
<td>7.69</td>
</tr>
</tbody>
</table>

*Note:* Prediction accuracy is measured by test set R-squared.