Disentangling Age, Time, and Cohort Effects in Income Inequality: A Proxy Machine Learning Approach

David Bruns-Smith
Stanford University

Emi Nakamura
UC Berkeley

Jón Steinsson*
UC Berkeley

January 14, 2026

Abstract

A canonical finding from earlier research is that the cross-sectional variance of income increases sharply with age (Deaton and Paxson, 1994). However, the trend in this age profile is not separately identified from time and cohort trends. Conventional methods solve this identification problem by ruling out "time effects." This strong assumption is rejected by the data. We propose a new proxy variable machine learning approach to disentangle age, time and cohort effects. Using this method, we estimate a significantly smaller slope of the age profile of income variance for the US than conventional methods, as well as less erratic slopes for 11 other countries.

JEL Classification: E20, J20

*We thank Avi Feller, Lihua Lei, Jesse Rothstein, and Stefan Wager for incredibly helpful comments and discussions. We thank the Alfred P. Sloan Foundation, and the Smith Richardson Foundation for financial support.

1 Introduction

The amount of uninsurable income risk individuals face over the life-cycle is an important determinant of a wide variety of macroeconomic outcomes such as precautionary savings, the average marginal propensity to consume, and the value of social insurance programs. Seminal work by Deaton and Paxson (1994) argued that the cross-sectional variance of income increases sharply with age in a number of countries. Subsequent work by Storesletten, Telmer, and Yaron (2004) and Guvenen et al. (2021) (among others) constructed similar estimates for the US. A large literature on heterogeneous agent models in macroeconomics has used these statistics to estimate the size and persistence of uninsurable income risk. A steeply upward-sloping profile of income variance as individuals age is interpreted as being caused by an accumulation of large and very persistent uninsurable shocks to income.

Despite their ubiquitous use in the literature, these facts about the age profile of income variance are not as straightforward as they may first appear. Life-cycle models are calibrated to match the evolution of the cross-sectional variance of income as individuals age holding fixed various cohort characteristics as well as the macroeconomic environment. In reality, cohort characteristics and the macroeconomic environment change over time. This makes it difficult to identify whether changes in the cross-sectional variance of income are due to age effects, cohort effect, or time effect.

A typical dataset can tell us how the cross-sectional variance of income of a particular cohort (say the cohort aged 30 in the year 2000) changes as that cohort ages. Suppose we see that the cross-sectional variance of income of this cohort increases between 2000 and 2010. Does this arise from the the accumulation of idiosyncratic income shocks (age effects)? Or does this arise from the fact that 2000 was a business cycle peak, while 2010 was close to a business cycle trough (time effects)?

This is an example of the age-time-cohort identification problem that arises in many contexts in economics and other social sciences where researchers are working with data on the life-cycle of individuals, households, firms, or other entities. Examples include studies of the life-cycle profiles of wages, labor supply, voting, fertility, mortality, health, firm growth, firm innovation, in addition to consumption and income. In all of these cases, researchers are interested in separating age effects, from cohort and time effects. (For example, do voters become more conservative with age?) But since age is equal to time minus cohort, linear age, time, and cohort effects are not separately identified.

Importantly, in a model with additive age effects, time effects, and cohort effects, the three

sets of effects are identified up to a linear rotation (Deaton, 1997; Schulhofer-Wohl, 2018). In other words, it is only the slopes of the age, time, and cohort effects that are unidentified absent further restrictions. This implies that non-linear age-time-cohort effects are identified. It also means that any identifying restriction a researcher makes can be thought of as pinning down the slopes of the age, time, and cohort profiles.

Different parts of the literature address the age-time-cohort problem in different ways. One approach is to introduce seemingly "innocuous" identification assumptions such as fixing two individual years to have the same time effect (i.e., coarsen the time effects) or two (or more) cohorts to have the same cohort effect (coarsen the cohort effects). It is crucial to recognize, however, that these are not truly innocuous assumptions. Rather, they amount to implicitly strong assumptions about the slopes of the age, time, or cohort profiles. Papay and Kraft (2015) and Bell (2020) present examples of how such "innocuous" assumptions can drive the conclusions of empirical studies.[1]

In the literature on the age-profile of cross-sectional income variance, the standard approach to addressing the age-time-cohort problem—dating back to Deaton and Paxson (1994)—has been to drop time effects entirely. This approach yields the strongly upward-sloping age profile of income variance we refer to above. This stark assumption is, however, rejected by the data. The cross-sectional distribution of income responds strongly to recessions (Storesletten, Telmer, and Yaron, 2004; Guvenen, Ozkan, and Song, 2014).

Moreover, the choice of methodology matters for the substantive conclusions. Figure 1 contrasts estimates of the slope of the age profile of income variance for two methods: the conventional approach of dropping time effects (orange squares) versus the alternative of dropping cohort effects (blue dots). The "unconventional" approach—which allows for business cycle factors—yields a slope that is roughly half as large as the conventional approach. This large difference was noted by Heathcote, Storesletten, and Violante (2005), but the subsequent literature has largely ignored this concern.

We propose a new approach to solving the age-time-cohort problem that combines proxy variables with (debiased) machine learning. Our jumping-off point is the "proxy variable" approach in which a subset of the age, time, and cohort effects are modeled as being a function of a set of observable "proxies" (Heckman and Robb, 1985; Winship and Harding, 2008). We choose to model the time effects in this manner. Rather than assuming that time effects are zero, we assume that they are a function of business cycle proxy variables. A key challenge with this approach—which

[1] Schulhofer-Wohl (2018) and Rothstein (2023) address the identification problem by focusing only on non-linear age, time and cohort effects. This is, however, insufficient in our application — the variance of persistent income shocks depends directly on the upward trend of the age profile.

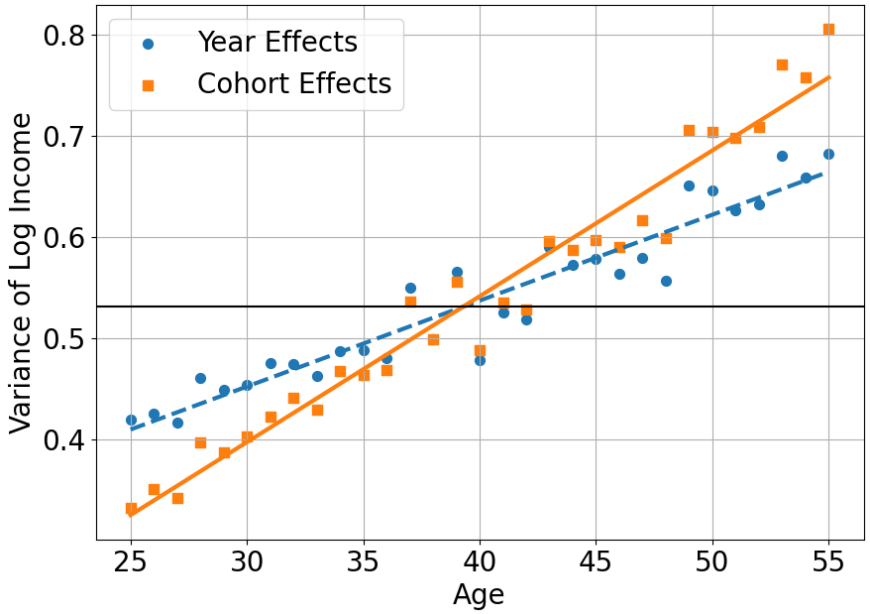


Figure 1: Two Estimates of the Age Profile of Income Variance

Note: The orange squares are age effects from a model that regresses the cross-sectional variance of log-income for a cohort at a points in time on age and cohort effects (but not time effects). The blue circles are age effects from an analogous model with age and time effects (but no cohort effects). We place a black horizontal line where the profiles intersect to aid visualization of the slopes. The data are from the PSID. The sample period is 1978-2019. See sections 2 and 3 for further details.

we demonstrate—is that small and seemingly-reasonable changes to the proxy model generate highly variable predictions for the age profile. The proxy variable approach, therefore, involves a non-trivial model selection problem.

Since machine-learning methods excel at model selection, we adopt the debiased machine learning approach of Chernozhukov, Newey, and Singh (2022) to estimate the slope of the age profile using proxy variables. We choose among a large number of proxy variables using cross-validation—a core idea from the machine learning literature. This approach yields a model that explains the data well, even out of sample, without us having to specify the precise functional form ahead of time. Furthermore, the debiasing procedure of Chernozhukov, Newey, and Singh (2022) yield unbiased point estimates, valid standard errors, and an asymptotically normal confidence interval for the estimated slope of the age profile of cross-sectional income variance.

We apply our method to US data from the Panel Survey of Income Dynamics (PSID) and to 12 countries represented in the Global Repository of Income Dynamics (GRID) database. Our debiased ML method yields much smaller estimates of the slope of the age profile of income variance than conventional methods. For the U.S. using the GRID dataset, our debiased ML estimate of the slope is half as large as the estimate based on conventional methods. This implies a smaller

role for permanent income shocks than is typically assumed in the literature. We also estimate a smaller persistence of income shocks (a half-life of 4 years) using our debiased ML method. Estimates based on the GRID data are generally smaller (and much more precisely estimated) than estimates based on the PSID.

When we apply our method to the 12 countries in the GRID dataset, we find that the conventional approach of dropping time effects yields extreme variability in the estimated slope of the age profile of income variance across countries. For some countries the age profile is hugely positively sloped, while for others it has a very large negative slope. This contrasts strongly with estimates using our debiased ML method. This method yields slopes that are much more homogeneous across countries, clustering around modest positive values (similar to our estimate for the U.S.).

There are two additional novel elements of our procedure that deviate from common practice in debiased machine learning and previous work on proxy variables. First, we implement cross-validation by leaving out whole blocks of contiguous years—a procedure we describe as "blocked cross-validation." Surprisingly, we find that without blocking, proxy variables can fail to achieve identification because machine learning methods can exploit dependence over time to effectively recreate time trends. Previous work on proxy variable methods implicitly ruled out this possibly via linearity assumptions. Second, we must rule out a set of trending proxy variables that flexible machine learning models can use to reconstruct time trends even after blocked cross-validation has been performed. Including this set of "forbidden proxies" resurrects the original identification problem, as above, by effectively reintroducing a time trend. We theoretically substantiate these requirements with a nonparametric analysis of the age profile estimand in the appendix of the paper.

The remainder of the paper is organized as follows. Section 2 discusses the data on the life-cycle of income we use. Section 3 describes the conceptual framework we employ and reviews the age-time-cohort identification problem. Section 4 discusses the proxy approach and how it results in a non-trivial model selection problem. Section 5 introduces our debiased machine learning approach to solving this model selection problem. Section 6 presents the results for our empirical application. Finally, section 7 concludes.

2 Data on Income

Our primary variable of interest is the cross-sectional variance of log male labor income by age and year. For the United State, we obtain estimates of this variable from two datasets. First, we use the Global Repository of Income Dynamics (GRID) (Guvenen et al., 2022a), a collection of aggregate statistics constructed using administrative microdata from 12 countries. The data for the US comes from the Longitudinal Employer-Household Dynamics Infrastructure files at the Census Bureau. The sample period is 1998-2019. We estimate our models using data on individuals of age 25-55. When calculating the slope of the age profile, however, we focus on individuals of age 35-50. This limits the influence of education and retirement on our results. All statistics within GRID are constructed using a standardized set of sample selection criteria intended to harmonize the data across countries. For example, income is deflated by the PCE, income observations under a minimum wage-based threshold are removed, and the variance of log male labor income is calculated after residualizing on year and age. We discuss these sample selection criteria in Appendix A.

For comparison with earlier work, we also report results for the US based on data from the Panel Study of Income Dynamics (PSID). Our sample period is 1978-2019. We calculate the variance of log male labor earns adjusted for changes in the PCE and residualized on year and age. We adopt a set of sample selection criteria meant to mimic the harmonized sample selection criteria used in the GRID data.

Beyond the US, GRID also records residual log income variances at the age-year level for 11 other countries: Argentina, Brazil, Canada, Denmark, France, Germany, Italy, Mexico, Norway, Spain, and Sweden. Data for these countries are taken from a range of administrative sources including tax and social security records. The sample period differs slightly across countries. See Appendix A for additional information.

We use a number of other data series as proxy variables for time effects. We discuss those data later in the paper.

3 How Big Is the Permanent Component of Lifecycle Income Risk?

Consider the following simple specification for the log income of individual i at age a:

$$\log Y_{i,a} = g(x_{i,a}) + y_{i,a} \tag{1}$$

$$y_{i,a} = \alpha_i + z_{i,a} + \epsilon_{i,a} \tag{2}$$

$$z_{i,a} = \rho z_{i,a-1} + \eta_{i,a}. \tag{3}$$

Log individual income $\log Y_{i,a}$ consists of a deterministic component $g(x_{i,a})$ and a stochastic component $y_{i,a}$. The deterministic component $g(x_{i,a})$ captures predictable variation in income with age.[2] The stochastic component $y_{i,a}$, consists of three parts: an individual fixed effect α_i, a persistent component $z_{i,a}$, and a transitory component $\epsilon_{i,a}$. The persistent component follows an AR(1) process with persistence ρ and innovations $\eta_{i,a}$. We assume for simplicity that $y_{i,a}$, $z_{i,0}$, α_i, $\epsilon_{i,a}$, and $\eta_{i,a}$ are all mean zero. We denote the variances of α_i, $\epsilon_{i,a}$, and $\eta_{i,a}$ as $\sigma^2_\alpha, \sigma^2_\epsilon$, and σ^2_η, respectively, and write the initial variance of the persistent component as $\sigma^2_{z,0} \coloneqq \text{Var}[z_{i,0}]$.

A key question in models with uninsurable income risk is: what is the relative volatility of the persistent and transitory income shocks $\eta_{i,a}$ and $\epsilon_{i,a}$? Persistent income shocks are much harder to self-insure against. They therefore call for more precautionary savings and imply higher marginal propensities to consume for a given level of liquid assets. Deaton and Paxson (1994) point out that there is a straightforward way to identify the importance of persistent shocks in terms of the age profile of the cross-sectional variance of income. For expositional simplicity, consider the case of $\rho = 1$, i.e., the case where the persistent component is a permanent component. In this case, the variance of the stochastic component of log income is

$$\begin{aligned} \sigma^2_a \coloneqq E[y^2_{i,a}] &= \sigma^2_\alpha + E[z^2_{i,a}] + \sigma^2_\epsilon \\ &= \sigma^2_\alpha + \sigma^2_{z_0} + a\,\sigma^2_\eta + \sigma^2_\epsilon. \end{aligned} \tag{4}$$

Notice that σ^2_a is linear in age with a slope of σ^2_η. In other words, the slope of the age profile of income variance identifies the variance of the persistent component of income σ^2_η when $\rho = 1$. The intuition for this is straightforward: the cross-sectional variance of income will fan out with age since the permanent shocks cumulate as people age (while transitory shocks do not). If $\rho < 1$, the age profile of cross-sectional income will be concave rather than linear. However, the degree to which it rises and the degree to which it is concave will identify σ^2_η and ρ. As we saw in Figure

[2] The GRID data residualize income on age and year. We follow this same procedure when we use the PSID data.

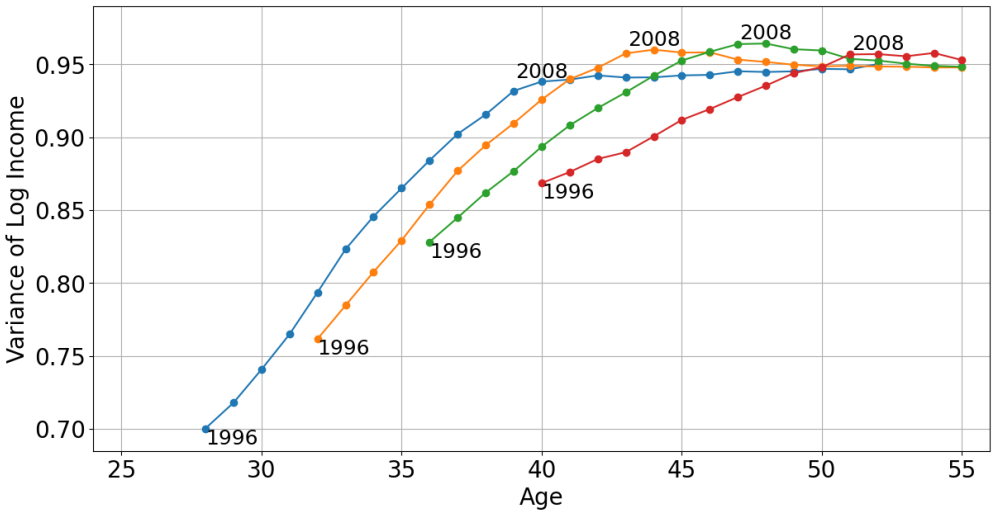


Figure 2: Cross-Sectional Income Variance by Age for Several US Cohorts

Note: Each point is the cross-sectional variance of log income in the US for a particular cohort in a particular year from the GRID data. The points for particular cohorts are connected by lines.

1, the age profile of the cross-sectional variance of income is approximately linear in the PSID. In section 6, we will see that it is somewhat concave in GRID.[3]

The challenge with using the age profile of the cross-sectional variance of income to identify the variance of persistent shocks to income is that the pure age profile is not directly observable in the data. We can compare the cross-sectional variance of income of people of different ages at a given point in time. But differences may then be due to the people belonging to different cohorts. We can follow the cross-sectional variance of income of a particular cohort as it ages. But differences may then be due to the fact that observations at different ages are taken at different times. In the data, we observe log income at age a in year t, and can construct moments $E[(y^t_{i,a})^2] =: \sigma^2_{at}$. Figure 2 illustrates this using data from GRID. We would like to control for the effect of calendar year t and birth cohort $c = t - a$ to get moments σ^2_a that are "timeless" and "cohortless" and can then be used to estimate σ^2_η via equation (4). However, the collinearity of age, time, and cohort makes this tricky to do.

Consider the fixed effects model

$$\sigma^2_{at} = \alpha_a + \beta_t + \gamma_c + \nu_{at}, \tag{5}$$

where α_a is a set of age fixed effects, β_t is a set of time fixed effects, γ_c is a set of cohort fixed effects, and ν_{at} an idiosyncratic component. If the three sets of fixed effects in this model were identified,

[3]The autocovariances of the stochastic component of log income are also informative about σ^2_η. In the $\rho = 1$ case, they are $E[y_{i,a}y_{i,a+h}] = \sigma^2_\alpha + \sigma^2_{z_0} + a\,\sigma^2_\eta$.

α_a would be the pure age effects we seek to use to estimate σ_η^2. Unfortunately, α_a, β_t and γ_c are perfectly collinear since $a = t - c$. In particular, we can replace α_a, β_t and γ_c by

$$\tilde{\alpha}_a = \alpha_a + ka, \qquad \tilde{\beta}_t = \beta_t - kt, \qquad \tilde{\gamma}_c = \gamma_c + kc, \tag{6}$$

where k is any scalar and the predictions of the model will be unchanged. In other words, equation (5) is identified up to a linear rotation. This poses a particularly severe problem in our application since it is the slope of the age-profile that identifies σ_η^2 (when $\rho = 1$). The slope is completely unidentified.

Any restriction that pins down k in equations (6) will render model (5) identified. Even minimal-seeming restrictions—such as setting two time effects equal to each other—will do the trick. It is tempting to view such restrictions as innocuous. In fact they are anything but innocuous since they determine the slope of all three sets of effects α_a, β_t and γ_c.

The standard approach in the literature on the age-profile of the cross-sectional variance of income is to make the strong assumption that all time effects are zero, $\beta_t = 0 \, \forall t$ (e.g., Deaton and Paxson, 1994; Guvenen et al., 2021). It is easy to see from Figure 2 that this assumption is at odds with the data. The cross-sectional variance of all cohorts bends down around 2008. This business cycle fluctuation cannot be captured by a combination of age and cohort effects. The timing of the bend is a non-linear time effect, which *is* identified as emphasized by Schulhofer-Wohl (2018).

4 Proxying for Time Effects

The identification problem discussed above implies that we must make additional assumptions to identify a pure age profile of cross-sectional income variance σ_a^2. The key challenge is how to formulate a compelling set of assumptions. One approach to this is to assume that one set of effects (age, time, or cohort) can be modeled as being a function of observable variables. This approach is referred to as the proxy variable approach in the literature. An early formalization appeared in Heckman and Robb (1985), which was substantially generalized in Winship and Harding (2008).[4]

We propose to model the time effects in our application as being a function of observable variables Z_t. With this assumption (and maintaining linearity for expositional simplicity), equation

[4]Proxy variables have been used in many age-time-cohort applications, including to the level of consumption and income (Gourinchas and Parker, 2002), wealth holdings (Kapteyn et al., 2005), health (Portrait et al., 2010), female labor force participation (Euwals et al., 2011), and happiness (Su et al., 2022).

(5) becomes

$$\sigma_{at}^2 = \alpha_a + \gamma_c + \beta^\top Z_t + \nu_{at}, \tag{7}$$

where Z_t is a vector of variables and β is a vector of coefficients. As long as the proxy variables Z_t are: (1) not perfectly linear in time, (2) capture all time effects, and (3) do not themselves depend on both age and cohort, all of the parameters are identified. We describe the formal criteria for identification in Appendix B.

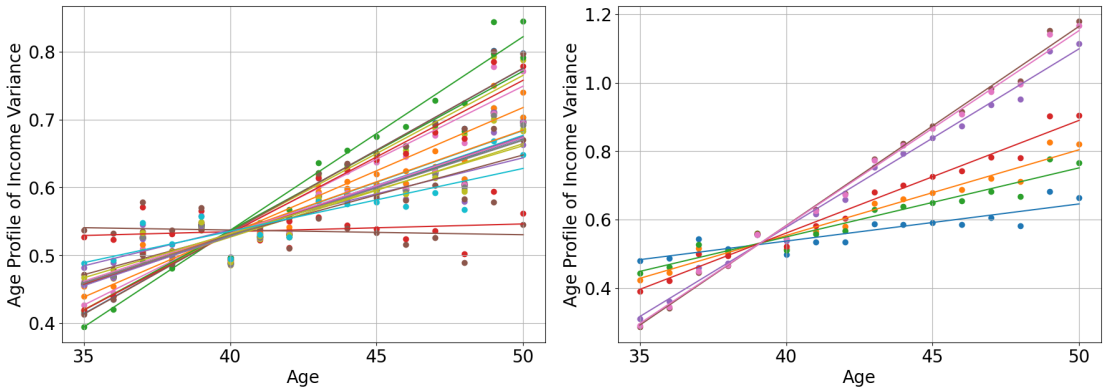
While the proxy variable approach can secure identification, its use raises an important model selection problem. There are many plausible proxy variables and there are many plausible functional forms. How should one choose among these? It turns out that this choice is quite consequential since results from proxy variable modeling of age, time, or cohort effects can be highly sensitive to these choices (see, e.g., Lu and Luo, 2021).

This sensitivity is easy to illustrate in our application. Figure 3 plots the profile of age effects (i.e., the α_a's) and their corresponding best fit lines that results from fitting the specification (7) with different plausible macroeconomic variables in Z_t (left panel) and different functional forms (right panel). Considering first the left panel: different choices of proxy variables can result in a slope of the age profile anywhere from slightly negative to strongly positive. In the right panel, we use a fixed set of proxies — Unemployment, log GDP, and Inflation — but consider polynomial expansions of the proxies of different degrees. As the polynomial degree we include sweeps from one to seven, the estimated slope ranges from relatively flat to strongly positive.

Figure 3 clearly demonstrates that different seemingly-reasonable choices of proxy variables and functional forms yield radically different estimates for the age slope. Clearly, our results will hinge crucially on how we resolve the model selection problem the proxy variable approach gives rise to. We now turn to that issue.

5 Our Methodology: Debiased Proxy Machine Learning

To solve the model selection problem discussed in section 4, we leverage several ideas from the machine learning literature. We consider a large set of candidate models with different proxy variables and different functional forms, including models with flexible interactions between age, cohort, and the proxies for time effects. We employ a blocked cross-validation procedure to choose among these models. Finally, we debias the resulting estimates of the slope of the age profile.



(a) Different Proxies

(b) Different Functional Forms

Figure 3: Sensitivity of Estimates of the Age Profile of Cross-Sectional Income Variance

Note: The left panel presents estimates of the age profile using (7), but using different proxies Z_t for time effects. We present estimates for 30 different proxy variables, used one at a time. The proxy variables are the 10 variables discussed in Section 5, the lag of each of these, and the first difference of each of these. We plot both the estimated age profiles (as dots) and the corresponding best-fit lines. The right panel present estimates of the age profile with different function forms for the model of time effects. In all cases, $Z_t = \{\text{Unemployment, log GDP, Inflation}\}$, but we consider polynomial expansions of these proxies of different degrees from one to nine.

5.1 Proxies, Features, and Models

The specific machine learning methods we use are guided by the relatively modest size of our dataset. Many machine learning methods are designed for large datasets. Our datasets have between 434 and 806 observations (age-time pairs). For datasets of this size, ridge regression, Lasso, and kernel ridge regression perform well, while more flexible models like random forests and neural networks typically fail to generalize. We therefore use ridge regression, lasso, and kernel ridge regression (KRR) as our basic set of models — we consider more complex models in Appendix D.2.

Our baseline set of proxy variables are ten macroeconomic time series: log GDP, log consumption, log investment, export values, log industrial production, unemployment, log CPI, a short term interest rate, the exchange rate, and oil prices. For the US, we also include a measure of the skill premium. We summarize the particular data series we use, and their sources in Appendix A.4. For each variable, we consider eight transformations (in addition to the variable itself): the first difference, lag, square, square root, square of first difference, square root of first difference, square of lag, and square root of lag. All in all, this makes for $9 \times 10 = 90$ "features" (99 for the US).

Flexible machine learning models can theoretically learn the relevant features from the data. In practice, however, preprocessing and trimming of the feature set can make the learning task easier,

substantially improving predictive performance. Given the modest size of our dataset, we are able to train our models with many different feature subsets. However, doing so for every combination of feature sets (2^{90} feature sets) is infeasible. We therefore perform two feature selection steps prior to training our models.

First, we remove proxies that trend too strongly with time as these threaten identification (see Appendix C.3). Specifically, we regress each feature on time and drop features for which this regression has an $R^2 > 0.75$. Second, we filter for proxies that have the strongest univariate relationship with the outcome variable. In particular, we fit kernel ridge regression models using the outcome variable and one proxy at a time.[5] As candidate features, we keep only the ten proxies with the smallest cross-validated mean squared error.

We train our models on all combinations of these ten chosen candidate features. For each sublist, we also include either age and cohort trends or age and cohort dummies. We therefore train our models on $2 \times (2^{10} - 1) = 2046$ different feature matrices for each country. For each method, we consider a range of possible values of the most relevant hyperparameters: the regularization penalties for lasso and ridge regressions, and the kernel and bandwidth for KRR. See Appendix D.1 for details. We normalize each feature (including the dummy variables) by subtracting their mean and dividing by their standard deviation.

5.2 Model Selection with Blocked Cross-Validation

We use cross-validation—i.e., pseudo-out-of-sample fit—to choose among the many models discussed above. Standard cross-validation is performed by *randomly* splitting observations in the dataset into k disjoint "folds" (i.e., subsets). A typical approach is to "train" (i.e., estimate) each model on data from $k - 1$ of the folds, and evaluate performance in terms of mean-squared error (MSE) on the remaining fold. This procedure is repeated using all possible splits into training and test sets. The overall performance of the model is then the average test MSE across splits and the model with the smallest cross-validated MSE is chosen.

This plain-vanilla cross-validation procedure runs into "data leakage" problems when the data are dependent as is the case in our panel data setting. Recall that our data are indexed by age and time. We therefore have many observations from each year (one for those aged 35, another for the those aged 36, etc.). The standard random split of the data into k folds will typically result in data from a given year being divided across folds, and therefore both training and test data can simultaneously contain observations from the same year. As a consequence, the model is never

[5] We use cross-validation to choose the kernel, kernel bandwidth, and penalty parameter for each proxy.

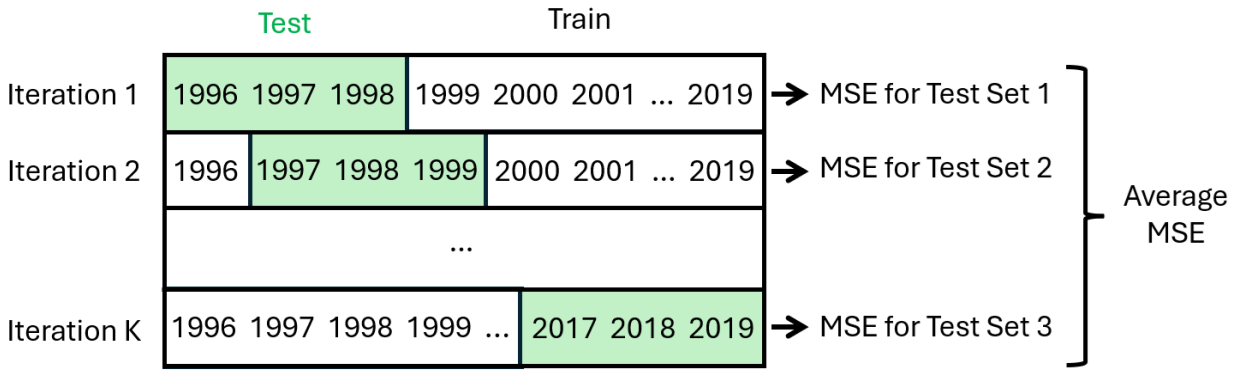


Figure 4: Our Blocked Cross-Validation Procedure with Block Size of Three Years

evaluated out-of-sample when it comes to time. The worry then is that the flexible models we employ can exploit the dependence between training and test sets to spuriously overfit the data. Since business cycles unfold over several years, dependence between observations close in time is also a concern.

To mitigate data leakage between training and test sets due to dependent data, we use *blocked cross-validation*. Rather than splitting the observations randomly, we split observations into blocks of contiguous years.[6] Specifically, we use eight year blocks. To reduce variance as much as possible, we use all possible contiguous year blocks of this length as test sets as illustrated in Figure 4. This is essentially an extension of repeated cross-validation (Burman, 1989; Kim, 2009) to the blocked setting: the average MSE from our procedure is the same as the expectation of uniformly-at-random choosing a contiguous block of years as a test set. The resulting procedure is more computationally intensive than standard k-fold blocked cross-validation, but remains tractable in our application because the number of years in our sample is not too large.

Blocked cross-validation plays an especially important role in our setting. Dependence between observations can degrade the quality of model selection to differing degrees in different settings — compare Bergmeir et al. (2018) and Liu and Zhou (2024). Surprisingly, in our setting, if we use plain-vanilla cross-validation for model selection, the slope of the age profile of σ_a^2 is *completely unidentified* even with proxy variables. We demonstrate this formally in Appendix C.3. In order to identify the contribution of the time effects, the proxy variables must be able to make

[6]There is a fairly extensive literature — both in machine learning and econometrics — going back to the late 80's on cross-validation variants adapted for dependent data. Influential early work includes Snijders (1988); Györfi et al. (1989); Burman et al. (1994); Racine (2000). A modern empirical literature using machine learning methods (Bergmeir and Benítez, 2012; Bergmeir et al., 2018; Cerqueira et al., 2020; Liu and Zhou, 2024) has consistently found that the blocked cross-validation method of Snijders (1988) performs well across a variety of different time-series settings. Racine (2000) additionally introduces "gaps" by dropping observations directly between train and test, but the modern empirical literature finds this does not improve performance in practice.

predictions on unseen years.[7]

5.3 Computing the Slope of the Age Profile

The machine learning procedure described above yields a predictive model for σ^2_{at} that takes as input age a, cohort c, and proxy variables z. We will write $\hat{m}$ to denote our trained model as a function of its inputs. For a given age-time pair (with implied cohort and proxies associated with that year), our model's prediction is $\hat{\sigma}^2_{at} = \hat{m}(a, c, z)$.

Our primary object of interest is not the individual predictions $\hat{m}(a, c, z)$ but rather the slope of the predictions as age varies. In the familiar special case of a linear model (7) we have $\hat{m}(a, c, z) = \hat{\alpha}_a + \hat{\beta}^\top z + \hat{\gamma}_c$. In this case, the age effects are the $\hat{\alpha}_a$'s and one can regress these on age to get the slope of the age profile of $\hat{m}(a, c, z)$.

More generally, $\hat{m}(a, c, z)$ may be non-linear and it is not as obvious how to define age effects. We adopt a definition based on *counterfactually* changing age, but leaving cohort and time fixed. We implement this in practice for the age effect at age 35 (say) by making predictions for σ^2_{35} using $\hat{m}$ for every observation in our dataset but counterfactually setting the age to 35 in all observations. In other words, we calculate the sample analog of $\hat{\sigma}^2_{35} := E[\hat{m}(35, c, z)]$. Doing this for every age yields a set of age effects. We can then calculate the slope of the age effects by regressing these age effects on age. In the special case of the linear model (7), we recover the usual fixed effects profile up to a constant. Specifically, we have $\hat{\sigma}^2_{35} = \hat{\alpha}_{35} + E[\hat{\beta}^\top z + \hat{\gamma}_c]$ where the expectation term is an intercept independent of age. Our non-parametric definition has the advantage that it can be computed for any machine learning model (whether or not the model uses a notion of age fixed effects internally). A similar definition using counterfactual averages was recently independently proposed for cohort profiles in Reynolds (2024).

5.4 Debiased Machine Learning and Standard Errors

Machine learning estimators virtually always use bias to reduce variance. For example, Ridge and Kernel Ridge regressions use regularization to explicitly shift their coefficients towards zero. Regularization can be optimal for minimizing out-of-sample mean square error (by preventing overfitting). But in our context—absent some adjustment—the bias will pass through to our estimate of the slope of the age effects. A key concern for us is therefore: will we estimate a smaller

[7] As a special case, using the AIC or BIC for model selection also yields complete lack of identification. The AIC and BIC are asymptotically equivalent to leave-one-out and leave-k-out cross-validation respectively (Stone, 1977). Winship and Harding (2008) use the AIC for model selection, but they secure identification through a linearity assumption.

slope with our machine learning approach simply due to a mechanical effect of regularization?

We address this issue by using the "Debiased Machine Learning" (DML) approach of Chernozhukov et al. (2018). Recall that $\hat{m}$ is our cross-validated machine learning predictor of σ^2_{at}. Let slope$(\hat{m})$ be our estimate of the slope of the age profile as described in Section 5.3. The key idea behind DML is to estimate a function, $\hat{\alpha}(a, c, z)$, associated with the slope operator, slope$(\cdot)$, called its "Riesz representer." Chernozhukov et al. (2018) show that this function $\hat{\alpha}$ can be used to add the following bias correction term to our baseline estimate of the slope:

$$\text{slope}(\hat{m}) + \hat{E}\big[\hat{\alpha}(a, c, z) \cdot \big(\underbrace{\sigma^2_{at} - \hat{m}(a, c, z)}_{\text{prediction errors}}\big)\big], \tag{8}$$

with the resulting estimate being asymptotically unbiased and having a valid normal confidence interval. This holds even if the estimates $\hat{m}$ and $\hat{\alpha}$ are both biased themselves.

It remains to define and estimate the Riesz representer. Our main strategy for this is to adopt the "automatic" estimation strategy of Chernozhukov, Newey, and Singh (2022). Their approach uses the fact that the true Riesz representer is the unique minimizer of the following optimization problem:

$$\min_{\alpha} \big\{E[\alpha(a, c, z)^2 - 2 \cdot \text{slope}(\alpha)]\big\}. \tag{9}$$

We can estimate $\hat{\alpha}$ using machine learning algorithms by minimizing the loss (9). In doing this, we consider the same large set of candidate models and same set of proxy variables that we used to estimate $\hat{m}$. Likewise, for model selection, we use the same blocked cross-validation procedure. A downside of characterizing the Riesz representer as the solution to the optimization problem (9) is that it provides little intuition. We attempt to provide a more insightful discussion of the Riesz representer in Appendix C.

With estimates of $\hat{m}$ and $\hat{\alpha}$, we can compute the debiased estimate of the slope using (8). Note, that to achieve valid inference, we should estimate $\hat{m}, \hat{\alpha}$ in a separate sample from the one we use to compute (8). In practice, we use a cross-fitting procedure. The notation is cumbersome so we defer a complete description of the cross-fit estimator to Appendix D.5. Because we fit both $\hat{m}$ and $\hat{\alpha}$ using machine learning and then combine them, this procedure is sometimes called "double machine learning" or "doubly-robust estimation".[8]

[8]This approach dates back to at least Robins et al. (1994) for the estimation of average treatment effects. See Kennedy (2022) for a review. For earlier work in the econometrics literature that takes advantage of writing the Riesz representer as the solution to an optimization problem, see Ai and Chen (2003).

Under minimal conditions derived in Chernozhukov et al. (2023), plugging our resulting $\hat{m}$ and $\hat{\alpha}$ into equation (8) results in an unbiased estimator of the slope. This implies that the smaller estimates of the slope our method yields compared to previous estimate are not driven mechanically by regularization bias. To further buttress this point, in Appendix G, we perform a simulation study with semi-synthetic data generation processes calibrated to the US GRID data to assess the validity of our methodology. We design this simulation to act as a sort of placebo, where the ground-truth slope is made to be high, but naive fixed effects regression would result in an incorrectly low slope. We find that the debiasing step reduces bias by a factor of ten and our debiased proxy machine learning estimator accurately recovers the higher value of the slope in simulation.

In addition to ensuring unbiasedness, the debiasing procedure also results in an asymptotically-normal confidence interval with the usual standard errors. We give a complete description of how we compute the point estimate and standard errors in Appendices D.4 and D.5. In our simulation study, we find that naively computing standard errors without debiasing results in a confidence interval that drastically undercovers — the naive 95% confidence interval contains the truth 0% of the time, whereas our debiased confidence interval contains the truth 96% of the time. Thus, debiased machine learning also plays a crucial role in enabling accurate uncertainty quantification.

We provide an outline of our end-to-end procedure in Algorithm 1. We emphasize that our estimator can be applied to other age-time-cohort settings by replacing σ^2_{at} with any other outcome variable.

6 Results

We next employ the model discussed in section 5 to estimate the slope and persistence of the age profile of the cross-sectional variance of income. We begin by presenting results for the United States. We then present an international comparison including 11 other countries available in the GRID dataset.

6.1 Results for the United States

Table 1 reports our estimates of the slope of the age profile of the cross-sectional variance of income for the United States. We present results from the GRID and PSID datasets. For PSID, we present results for two sample periods: 1998-2019 (for comparability with our GRID results) and 1978-2019. We present results for four models: the conventional "age-cohort" model (which includes age and cohort fixed effects, but no time fixed effects), the "age-time" model (which includes age

Algorithm 1 Debiased Proxy Machine Learning

INPUT: block length B, predictive models $\mathcal{M}$, and weighting models $\mathcal{W}$
Let $L = n_{\text{year}} - B + 1$ be the number of blocks.
for year $t = 1, \ldots, L$ **do**
  Let $\mathcal{D}_t^{\text{test}}$ be the observations in years t to $t + B$. Let $\mathcal{D}_t^{\text{train}}$ be the rest.
  for $m_i \in \mathcal{M}$ **do**
    Train model m_i on $\mathcal{D}_t^{\text{train}}$ to predict σ_{at}^2 given a, c, z.
    Compute MSE_{it}, the mean-squared error of m_i on $\mathcal{D}_t^{\text{test}}$.
  end for
  for $w_j \in \mathcal{W}$ **do**
    Train weight model w_j on $\mathcal{D}_t^{\text{train}}$ to minimize the "Riesz loss" (9).
    Compute RL_{jt}, the Riesz loss of w_j on $\mathcal{D}_t^{\text{test}}$.
  end for
end for
Let $\hat{m} = \text{argmin}_{m_i} \frac{1}{L} \sum_{t=1}^{L} \text{MSE}_{it}$.
Let $\hat{w} = \text{argmin}_{w_j} \frac{1}{L} \sum_{t=1}^{L} \text{RL}_{jt}$.
Compute $\hat{\theta}$, the debiased point estimate of the slope, using $\hat{m}$ and $\hat{w}$ as in (20).
Compute $\hat{V}$ using $\hat{m}, \hat{w}, \hat{\theta}$ as in (21), and compute $\widehat{\text{std err}}$ using $\hat{V}$ as in (22).
OUTPUT: $\hat{\theta}$ and $\widehat{\text{std err}}$

and time fixed effects, but no cohort fixed effects), and two versions of our machine learning (ML) model (with and without our debiasing procedure). For the debiased ML estimates, we include estimates of the 95% confidence interval of our estimate.

We estimate these four models on data from individuals of age 25 to 55. However, to limit the influence of schooling and retirement, our main estimates of the slope of the age profile are for age effects over the age range 35 to 50. Later in this section, we present results on the age profile for the full age range.

A result that emerges from Table 1 is that our debiased ML estimate of the slope of the age profile is considerably smaller than the age-cohort estimate that has been widely used in the literature. For the GRID dataset, the debiased ML estimate of the slope is 0.33% per year, while the age-cohort estimate is 0.67% per year.[9] For the PSID with sample period 1978-2019, the debiased ML estimate is 0.84% per year, while the age-cohort estimate is 1.39% per year. We conclude from this that the widely-used age-cohort estimate of the slope of the age profile of cross-sectional income variance likely yields estimates that are considerably upward biased. The age-cohort estimate is outside the 95% confidence interval for the debiased ML estimate in both of the cases just discussed. In these cases, our debiased ML estimate is much closer to the age-time estimate than

[9] A slope of 0.33% per year means that we estimate that the cross-sectional variance of log income rises by 0.0033 per year of age. This then also means that the cross-sectional variance rises by 0.33% per year to a first order approximation.

Table 1: Slope of Age Effects of the Cross-Sectional Variance of Income

Dataset	Years	Age-Cohort	Age-Year	Biased ML	Debiased ML
GRID	1998-2019	0.67	0.40	0.36	0.33 ± 0.13
PSID	1998-2019	1.34	0.64	0.56	1.08 ± 0.69
PSID	1978-2019	1.39	0.90	0.84	0.84 ± 0.35

Note: The table presents results on the slope of the age effects of the cross-section variance of income for the United States. We present the slopes multiplied by 100 to ease readability. The "Age-Cohort" column presents results from a model with age and cohort effects, but no time effects. The "Age-Time" column presents results from a model with age and time effects, but no cohort effects. The "Biased ML" column presents results from our model without debiasing. The "Debiased ML" column presents results for our model including our debiasing procedure. This last column also presents estimates of the 95% confidence interval of the estimate. The PSID results use sample selection criteria similar to those for GRID.

the age-cohort estimate.

A second result that emerges from Table 1 is that we estimate a considerably smaller slope of the age profile with GRID data than with PSID data. The debiased ML estimate for GRID is 0.33% per year, while it is 0.93% per year and 0.84% per year for the two different sample periods we use for PSID. Our estimate for the GRID dataset is much more precise than our estimate for PSID. The 95% confidence interval for our GRID estimate is $[0.20, 0.46]$, while even for the longer sample period in the PSID our estimate is much wider at $[0.49, 1.19]$.

Table 2 presents estimates for an alternative measure of the growth in the cross-sectional variance of income over the life cycle: the change in the cross-sectional variance of income between age 35 and age 50. These estimates tell the same story as the slope estimates in Table 1. The debiased ML estimates are substantially smaller than the age-cohort estimates and the estimates from GRID are substantially smaller than the estimates from the PSID.

Table 2: Change in Cross-Sectional Variance of Income from Age 35 to 50

Dataset	Years	Age-Cohort	Age-Year	Biased ML	Debiased ML
GRID	1998-2019	10.4	6.5	5.7	5.3
PSID	1998-2019	18.9	8.4	9.6	13.1
PSID	1978-2019	24.0	15.8	14.7	14.7

Note: The table presents results on the change in the cross-sectional variance of income from age 35 to age 50. We multiply these estimates by 100 for readability. See the note for Table 1 for more detail.

Figure 5 plots our estimated age profiles for the GRID dataset. This estimated age profile is very smooth, reflecting the fact that the GRID dataset is based on a large administrative dataset. The estimated age profile for the PSID is estimated with vastly more sampling error (see Figure H.1 in the appendix). We can clearly see the difference in slope between the profile for the age-cohort model and the profile for the debiased ML model in Figure 5.

Another notable feature of the age profile in Figure 5 is that it is concave. This suggests that

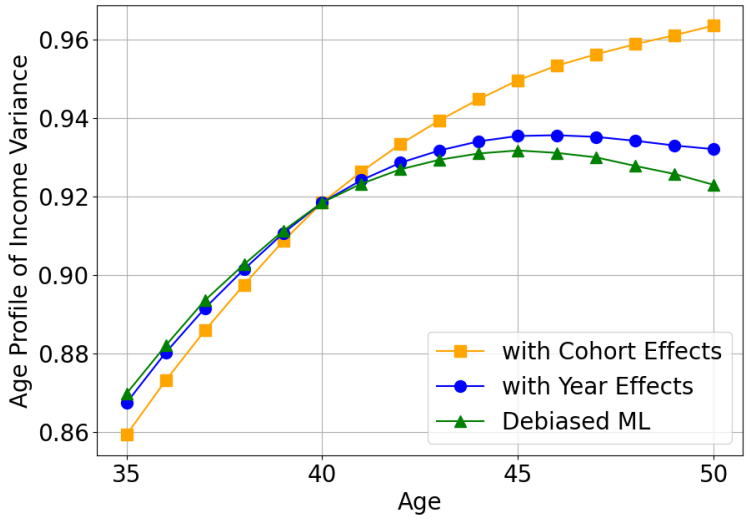


Figure 5: Age Profile of Cross-Sectional Income Variance in the US

Note: The figure plots the estimated age profile of the cross-sectional variance of income for three models using GRID data.

income shocks are less than fully permanent. Table 3 presents GMM estimates of the persistence parameter ρ for the income process presented in Section 3. (We describe the details of our GMM procedure in Appendix F.) For the GRID dataset, our debiased ML estimate of ρ is 0.854. This estimate implies that the half-life of income shocks in GRID is about 4 years. As with our estimates of the slope of the age profile, our debiased ML estimate for ρ is smaller than the age-cohort estimate. The age-cohort estimate for the GRID dataset is 0.934 implying a half-life of 10 years. Our debiased ML estimate therefore implies not only a smaller slope of the age profile, but also less persistence of income shocks.

Table 3: Persistence of Income Shocks (ρ)

Dataset	Years	Age-Cohort	Age-Year	Biased ML	Debiased ML
GRID	1998-2019	0.934	0.878	0.882	0.854 ± 0.075
PSID	1998-2019	0.974	0.950	0.939	0.944 ± 0.072
PSID	1978-2019	1.049	1.056	1.060	1.050 ± 0.045

Note: The table presents results on the persistence of income shocks—ρ in Section 3. See the note for Table 1 for more detail.

For the PSID, we estimate considerably larger values for ρ. For our longer PSID sample, we estimate a value for ρ that is larger than one. This lines up with the linear visual appearance of the (noisy) income profile for the PSID in Figure H.1. This estimate implies that income shocks are permanent. It matches the age profiles reported by Deaton and Paxson (1994) using CEX survey data, which similarly look noisy but linear.

Table 4: Slope of Age Effects Across Countries

Country	Years	Age-Cohort	Age-Year	Biased ML	Debiased ML
Argentina	1996-2015	-2.16	0.80	-0.13	0.33 ± 1.95
Brazil	1993-2018	-1.37	1.33	-0.25	0.24 ± 1.63
Canada	1996-2016	0.41	0.27	0.28	0.27 ± 0.10
Denmark	1996-2016	0.50	0.08	0.28	0.30 ± 0.26
France	1996-2016	0.44	0.28	0.38	0.43 ± 0.29
Germany	2001-2016	0.93	0.26	0.67	0.73 ± 0.33
Italy	1996-2016	1.09	-0.29	0.15	0.20 ± 0.92
Mexico	2005-2019	0.13	0.82	0.20	0.43 ± 0.40
Norway	1996-2017	0.91	0.48	0.71	0.63 ± 0.19
Spain	2005-2018	1.67	0.70	0.14	0.00 ± 0.43
Sweden	1996-2016	-0.23	0.18	0.09	0.02 ± 0.12
USA	1998-2019	0.67	0.40	0.36	0.33 ± 0.13

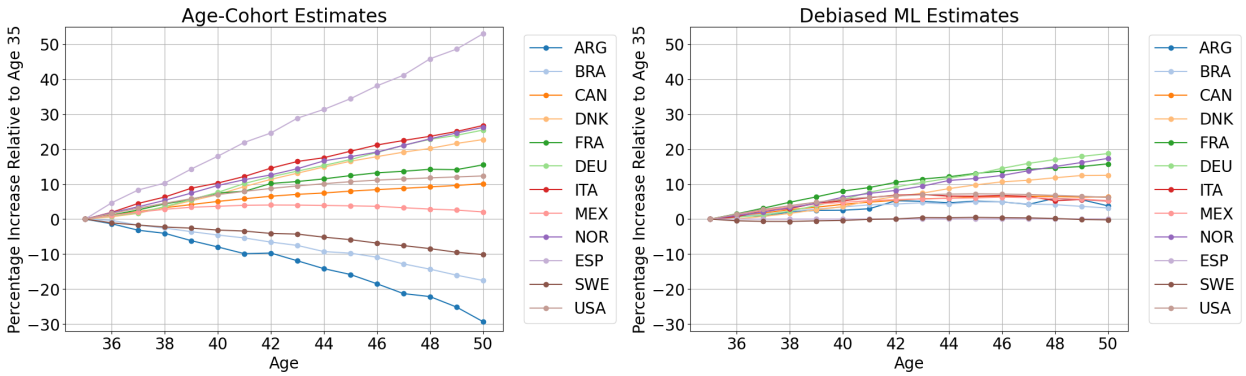
Note: The table presents results on the slope of the age effects of the cross-section variance of income for twelve countries using data from GRID. We present the slopes multiplied by 100 to ease readability. See the note for Table 1 for more detail.

6.2 Results for Twelve Countries

Table 4 presents estimates of the slope of the age profile of cross-sectional income variance for twelve countries using GRID data. We present estimates for the same four models as we did for the US above: the conventional "age-cohort" model, the "age-time" model, and two versions of our machine learning (ML) model (with and without our debiasing procedure). As with the US estimates, we estimate the age profile on the full range of ages, but then estimate its slope over the age range 35-50 years.

Consider first the estimates for the conventional age-cohort model. These vary wildly across the twelve countries. For the US, our age-cohort estimate is 0.67. For several countries we estimate much higher slopes: for Germany the slope is 0.93, for Italy it is 1.09, and for Spain it is 1.67. For other countries we estimate much lower slopes. For three countries we actually estimate negative slopes: -0.23 for Sweden, -1.37 for Brazil, and -2.16 for Argentina. Taken at face value, these results suggest enormous heterogeneity in the nature of income risk over the life-cycle across different countries.

Our estimates for the debiased ML model are much more homogeneous across countries. The estimate for the US is 0.33. The largest two slopes we estimate are 0.63 for Norway and 0.73 for Germany. None of the estimates are negative. The smallest two estimates are 0.0 for Spain and 0.02 for Sweden. Brazil and Argentina's estimates are close to the median country at 0.24 and 0.33, respectively. This contrasts sharply with the large negative estimate for these countries for the



(a) Age-Cohort Estimates

(b) Debiased ML Estimates

Figure 6: Normalized Age Profiles Across Countries

Note: The figure presents estimated age profiles for the twelve countries with have data on from GRID. The age profiles are normalized to zero at age 35 and plot the percentage changes in the cross-sectional variance of income at different ages relative to age 35.

age-cohort model.[10]

The confidence intervals we report for the debiased ML model in Table 4 are clustered by year (allowing for correlation across age within year). These confidence intervals reveal that our procedure yields quite precise estimates for a number of countries including the United States and Canada. However, for other countries, the confidence intervals we estimate are quite wide. This is particularly the case for Argentina and Brazil. For these countries, the confidence intervals include all plausible values for the slope of the age profile. These large confidence intervals arise because the proxy variables do not capture the time effects for these countries well in the earliest part of the sample, especially in 1996-1998 in Argentina and 1994-1995 in Brazil. Developing better proxy variables for time effects for these countries is an important area for further research.

Figure 6 visualizes the large difference between our estimates from the conventional age-cohort model and our debiased ML model. We plot the age profiles for all twelve countries normalized to zero at age 35 for each country. The left panel plots the estimated age profiles for the age-cohort model, while the right panel plots the age profiles for the debiased ML model. The difference is very stark.

Figure 7 plots our debiased ML estimates of the age profiles for the twelve countries in levels over the full range of ages we use to estimate these profiles: ages 25 through 55 years. Several results stand out. First, there is a large degree of heterogeneity in the level of income variance across countries. The three Latin American countries in our dataset (Argentina, Brazil, and Mexico) have the highest income variance. Below them is the United States. Considerably below that

[10]Table H.1 in Appendix H.2 presents results on the persistence of income shocks across countries.

come Canada and the eight European countries in our dataset.

For the three Scandinavian countries in our dataset (Denmark, Norway, and Sweden) we see a substantial drop in income variance from age 25 to age 30. This is the case to a lesser extent in France. In the other countries, the income profile is flat or upward sloping at young ages. Income variance in Norway is actually larger than in the U.S. for 25 year olds. But by age 30 it is nearly 30% lower. The extremely low income variance in Sweden and Denmark at older ages arises to a very large extent "due to" a substantial fall from ages 25 to 35 relative to countries such as Canada and Italy.

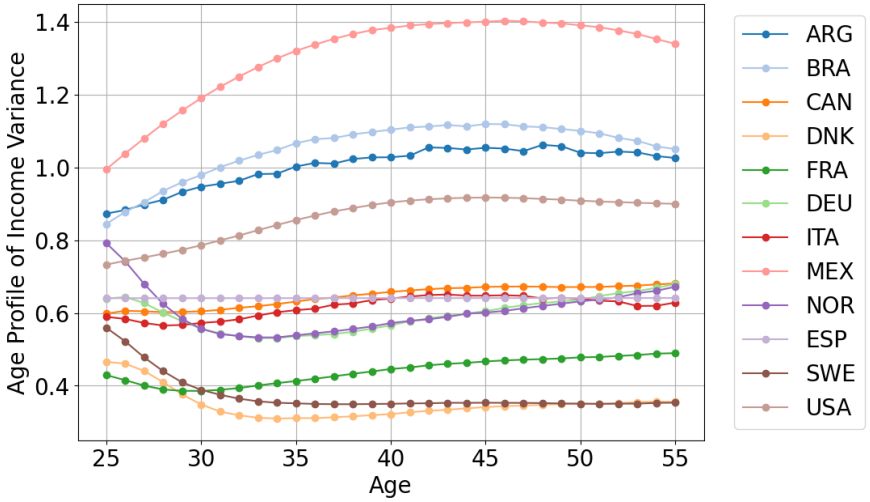


Figure 7: Age Profiles of Income Variance for an Extended Age Range

Finally, the degree to which the age profile is concave or convex is correlated with its level across countries. The countries with large income variance also have concave age profiles of income variance, while the countries with small income variance tend to have convex income profiles. The countries with middle of the road income variance levels tend to have income profiles that are flat and neither convex nor concave.

7 Conclusion

We propose a new proxy variable machine learning approach to disentangle age, time and cohort effects. We apply this methodology to the classic problem of estimating how the cross-sectional variance of income changes over the lifecycle. Our benchmark empirical analysis applies our methodology to the newly developed Global Repository of Income Dynamics (GRID) data, a collection of aggregate statistics constructed using administrative microdata for 11 countries: Argentina, Brazil, Canada, Denmark, France, Germany, Italy, Mexico, Norway, Spain, Sweden and

the United States.

Our methodology allows us to make substantially weaker identifying assumptions than the standard approach in the literature of ruling out time effects—which is clearly rejected by fluctuations in inequality over the business cycle. We estimate a significantly smaller slope of the age profile of income variance for the US than conventional methods. The implications of our methodology for countries other than the US are even more striking. In emerging economies, ruling out time effects—which embody macroeconomics shocks—is highly problematic and implies erratic and variable profiles for the income variance over the lifecycle. Our new methodology yields age profiles that are much more consistent across countries.

Our application demonstrates a number of methodological points relevant for other applications. First, in panel data settings with substantial macroeconomic effects, it is crucial to use "blocked" cross-validation as opposed to off-the-shelf cross-fitting techniques from the machnine-learning literature that simply leave out a random fraction of the data. Second, in weakening the assumptions about time effects, we must restrict attention to proxy variables that differ sufficiently from time trends. Finally, we find that kernel ridge regression performs well in our application (with its relatively modestly-sized dataset) relative to more complex machine learning models such as random forests or neural networks. We believe that our methodology has the potential to provide a structured alternative to existing methods for addressing the identification challenges arising from the potential existence of age, time and cohort effects in a wide variety of applications in the social sciences ranging from understanding the age profile of political preferences to the age profile of household wealth.

References

AI, C. AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71, 1795–1843.

BACH, P., O. SCHACHT, V. CHERNOZHUKOV, S. KLAASSEN, AND M. SPINDLER (2024): "Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study," in *Causal Learning and Reasoning*, PMLR, 1065–1117.

BELL, A. (2020): "Age period cohort analysis: a review of what we should and shouldn't do," *Annals of human biology*, 47, 208–217.

BEN-MICHAEL, E., A. FELLER, D. A. HIRSHBERG, AND J. R. ZUBIZARRETA (2021): "The balancing act in causal inference," *arXiv preprint arXiv:2110.14831*.

BERGMEIR, C. AND J. M. BENÍTEZ (2012): "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, 191, 192–213.

BERGMEIR, C., R. J. HYNDMAN, AND B. KOO (2018): "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Computational Statistics & Data Analysis*, 120, 70–83.

BRUNS-SMITH, D., O. DUKES, A. FELLER, AND E. L. OGBURN (2023): "Augmented balancing weights as linear regression," *arXiv preprint arXiv:2304.14545*.

BURMAN, P. (1989): "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika*, 76, 503–514.

BURMAN, P., E. CHOW, AND D. NOLAN (1994): "A cross-validatory method for dependent data," *Biometrika*, 81, 351–358.

CAMERON, A. C. AND P. K. TRIVEDI (2005): "Microeconometrics: Methods and Applications," .

CERQUEIRA, V., L. TORGO, AND I. MOZETIČ (2020): "Evaluating time series forecasting models: An empirical study on performance estimation methods," *Machine Learning*, 109, 1997–2028.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/debiased machine learning for treatment and structural parameters," .

CHERNOZHUKOV, V., W. NEWEY, R. SINGH, AND V. SYRGKANIS (2020): "Adversarial estimation of riesz representers," *arXiv preprint arXiv:2101.00009*.

CHERNOZHUKOV, V., W. K. NEWEY, AND R. SINGH (2022): "Automatic debiased machine learning of causal and structural effects," *Econometrica*, 90, 967–1027.

——— (2023): "A simple and general debiased machine learning theorem with finite-sample guarantees," *Biometrika*, 110, 257–264.

DARVAS, Z. M. (2021): "Timely measurement of real effective exchange rates," Tech. rep., Bruegel working paper.

DEATON, A. (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore, MD: Johns Hopkins University Press.

DEATON, A. AND C. PAXSON (1994): "Intertemporal choice and inequality," *Journal of political economy*, 102, 437–467.

EUWALS, R., M. KNOEF, AND D. VAN VUUREN (2011): "The trend in female labour force partici-

pation: what can be expected for the future?" *Empirical Economics*, 40, 729–753.

FARKAS, G. (1977): "Cohort, age, and period effects upon the employment of white females: Evidence for 1957–1968," *Demography*, 14, 33–42.

GOURINCHAS, P.-O. AND J. A. PARKER (2002): "Consumption over the Life-Cycle," *Econometrica*, 70, 47–89.

GUVENEN, F., F. KARAHAN, S. OZKAN, AND J. SONG (2021): "What Do Data on Millions of US Workers Reveal about Lifecycle Earnings Dynamics?" *Econometrica*, 89, 2303–2339.

GUVENEN, F., S. OZKAN, AND J. SONG (2014): "The nature of countercyclical income risk," *Journal of Political Economy*, 122, 621–660.

GUVENEN, F., L. PISTAFERRI, AND G. L. VIOLANTE (2022a): "The Global Repository of Income Dynamics," https://www.grid-database.org/. Accessed DD.MM.YYYY.

——— (2022b): "Global trends in income inequality and income dynamics: New insights from GRID," *Quantitative Economics*, 13, 1321–1360.

GYÖRFI, L., W. HÄRDLE, P. SARDA, AND P. VIEU (1989): "Regression Estimation and Time Series Analysis," *Nonparametric Curve Estimation from Time Series*, 15–51.

HEATHCOTE, J., K. STORESLETTEN, AND G. L. VIOLANTE (2005): "Two views of inequality over the life cycle," *Journal of the European Economic Association*, 3, 765–775.

HECKMAN, J. AND R. ROBB (1985): "Using longitudinal data to estimate age, period and cohort effects in earnings equations," in *Cohort analysis in social research: Beyond the identification problem*, Springer, 137–150.

HIRSHBERG, D. A., A. MALEKI, AND J. R. ZUBIZARRETA (2019): "Minimax linear estimation of the retargeted mean," *arXiv preprint arXiv:1901.10296*.

HOLZMÜLLER, D., L. GRINSZTAJN, AND I. STEINWART (2024): "Better by default: Strong pre-tuned mlps and boosted trees on tabular data," *Advances in Neural Information Processing Systems*, 37, 26577–26658.

IMBENS, G. W. AND D. B. RUBIN (2015): *Causal inference in statistics, social, and biomedical sciences*, Cambridge university press.

KAPTEYN, A., R. ALESSIE, AND A. LUSARDI (2005): "Explaining the wealth holdings of different cohorts: Productivity growth and social security," *European Economic Review*, 49, 1361–1391.

KENNEDY, E. H. (2022): "Semiparametric doubly robust targeted double machine learning: a review," *arXiv preprint arXiv:2203.06469*.

KIM, J.-H. (2009): "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap," *Computational statistics & data analysis*, 53, 3735–3745.

LIU, S. AND D. J. ZHOU (2024): "Using cross-validation methods to select time series models: Promises and pitfalls," *British Journal of Mathematical and Statistical Psychology*, 77, 337–355.

LU, Y. AND L. LUO (2021): "Cohort variation in US violent crime patterns from 1960 to 2014: An age–period–cohort-interaction approach," *Journal of quantitative criminology*, 37, 1047–1081.

PAPAY, J. P. AND M. A. KRAFT (2015): "Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement," *Journal of Public Economics*, 130, 105–119.

PEARL, J. (2009): *Causality*, Cambridge university press.

PORTRAIT, F., R. ALESSIE, AND D. DEEG (2010): "Do early life and contemporaneous macroconditions explain health at older ages? An application to functional limitations of Dutch older individuals," *Journal of Population Economics*, 23, 617–642.

RACINE, J. (2000): "Consistent cross-validatory model-selection for dependent data: hv-block cross-validation," *Journal of econometrics*, 99, 39–61.

REYNOLDS, N. (2024): "Local Average Cohort Effects," Working Paper, University of Essex.

ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American statistical Association*, 89, 846–866.

ROTHSTEIN, J. (2023): "The Lost Generation?: Labor Market Outcomes for Post-Great Recession Entrants," *Journal of Human Resources*, 58, 1452–1479.

SCHOLKOPF, B. AND A. J. SMOLA (2018): *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press.

SCHULHOFER-WOHL, S. (2018): "The age-time-cohort problem and the identification of structural parameters in life-cycle models," *Quantitative Economics*, 9, 643–658.

SINGH, R. (2021): "Kernel ridge Riesz representers: Generalization, mis-specification, and the counterfactual effective dimension," *arXiv preprint arXiv:2102.11076*.

SNIJDERS, T. A. (1988): "On cross-validation for predictor evaluation in time series," in *On Model Uncertainty and its Statistical Implications: Proceedings of a Workshop, Held in Groningen, The Netherlands, September 25–26, 1986*, Springer, 56–69.

STONE, M. (1977): "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 44–47.

STORESLETTEN, K., C. I. TELMER, AND A. YARON (2004): "Cyclical dynamics in idiosyncratic labor market risk," *Journal of political Economy*, 112, 695–717.

SU, Y.-S., D. LIEN, AND Y. YAO (2022): "Economic growth and happiness in China: A Bayesian multilevel age-period-cohort analysis based on the CGSS data 2005–2015," *International Review of Economics & Finance*, 77, 191–205.

VAN DER VAART, A. AND J. A. WELLNER (2023): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Nature.

WILLIAMS, N. T., O. J. HINES, AND K. E. RUDOLPH (2025): "Riesz representers for the rest of us," *arXiv preprint arXiv:2507.19413*.

WINSHIP, C. AND D. J. HARDING (2008): "A mechanism-based approach to the identification of age–period–cohort models," *Sociological methods & research*, 36, 362–401.

A Data

A.1 GRID US

For the results reported in section Section 6.1, we use data for the United States released as part of the Global Repository of Income Dynamics (GRID) (Guvenen et al., 2022a). The data come from the Longitudinal Employer-Household Dynamics Infrastructure files at the Census Bureau. We now describe the sample selection criteria used by GRID to construct the age-year moments. Note that we do not have access to the underlying microdata. Observations are at the individual level, and we consider only males for our main specification. The income measure is PCE-adjusted log labor earnings including wages, salaries, tips, bonuses, but not self-employed or business income. Observations are kept if they are above a floor of 260 times the real federal hourly minimum wage in that year. Income is winsorized at the 99.999999th quantile. Finally, log income is residualized by a regression on year and age dummies. See Guvenen et al. (2022b) and the GRID documentation for more details.

A.2 PSID following GRID

In section Section 6.1, we also present results based on data from the PSID. We develop sample selection criteria meant to mimic those used by the GRID dataset discussed above. We use the individual file to track all male individuals in the PSID over time, regardless of their head or marital status. We consider only individuals from the representative SRC sample, i.e. we drop the poverty and immigrant oversamples. We use PCE-deflated log labor income as the income measure, keeping only observations above the 260 times the real federal hourly minimum wage floor. We winsorize observations above the 99.99 quantile. Finally, log income is residualized by a regression on year and age dummies.

A.3 Grid International

In section Section 6.2, we also present results for 11 additional countries available in the GRID database. GRID tries as much as possible to harmonize the data across countries, but the precise sample selection criteria and variable definitions differ slightly. The GRID website provides details on sample section and variable definitions for each country — https://www.grid-database.org/documentation. We provide three illustrative examples here that highlight some of the divergences across countries.

The data for Germany combines records from social security data and personal income tax records. GRID uses random 10% subsample of the social security observations, and a random 25% subsample of the tax observations. Annual earnings includes overtime pay, bonuses, 13th month pay, paid sick leave, severance pay, and vacation allowance. Self-employment is excluded. A wage floor is imposed based on the Germany national minimum wage.

By contrast, the data for Norway comes from annual tax records, which cover the entire population. Labor income includes salaries and hourly wages; fees received by board members, bonuses, commissions; overtime, piecework, performance, caregiver, severance, and holiday payments; fixed wage and irregular supplements. Self-employment is excluded. Norway does not have a minimum wage, so a wage floor is imposed based on the US minimum wage.

Finally, the data for Argentina comes from employer-employee matched social security records. GRID uses a random 3% subsample, which is representative of formal employment at private firms — however private formal employment is only 30-40% of total employment. Labor income includes base salary, overtime compensation, performance and seasonal bonuses, paid vacations, paid sick leaves, and severance payments. Self-employment is excluded. A wage floor is imposed based on the Brazilian national minimum wage.

The sample period available for different countries in GRID also varies somewhat. Table A.1 provides the sample period for each country.

Table A.1: GRID Sample Years

Country	Full Sample Years	Sample Years with Proxies
Argentina	1996-2015	1996-2015
Brazil	1985-2018	1993-2018
Canada	1983-2016	1996-2016
Denmark	1987-2016	1996-2016
France	1991-2016	1996-2016
Germany	2001-2016	2001-2016
Italy	1985-2016	1996-2016
Mexico	2005-2019	2005-2019
Norway	1993-2017	1996-2017
Spain	2005-2018	2005-2018
Sweden	1985-2016	1996-2016
USA	1998-2019	1998-2019

A.4 Proxy Variables

Our baseline set of proxy variables are ten macroeconomic time series: log GDP, log consumption, log investment, export values, log industrial production, unemployment, log CPI, a short term interest rate, the exchange rate, and oil prices. We have sought to obtain harmonized definitions for these time series across countries. We source GDP, consumption, investment, exports, unemployment, and CPI from the World Bank's World Development Indicators (WDI) database. GDP, consumption, investment are measured in constant 2015 US dollars. For exports, we use the export value index (2015 = 100). For unemployment, we use the percent unemployed out of the male labor force because our income data is restricted to men. We source the industrial production index from the International Monetary Fund's (IMF) International Financial Statistics (IFS) database. Industrial production data were not available for Argentina and Denmark. We use a single measure of the short term nominal interest rate per country. We choose among the T-bill rate, the money market rate, and the policy rate (whichever series has the longest sample for each country). We source the T-bill and money market rates from Bloomberg, and the policy rate from the IMF IFS database. For the exchange rate, we use the real effective exchange rate series from Darvas (2021). We use their broad measure, REER 170. Finally, we use the same oil price series across countries. We adopt the West Texas Intermediate (WTI) spot crude oil price from FRED, Federal Reserve Bank of St. Louis.

Finally, for the United States, we include a measure of the skills premium. From the Current Population Survey, we source median usual weekly earnings of full-time wage and salary workers age 25 years and over with a Bachelor's degree and higher, and with a high school diploma, but no college. We use the ratio of these two series as our skills premium measurement.

B Proxy Assumptions

In this appendix, we provide formal criteria for when a vector of proxy variables Z is sufficient for replacing time in the age-time-cohort problem.

B.1 The Mechanism-Based Approach

We adopt the general framework introduced in Winship and Harding (2008). This framework assumes that the effects of age, time, and cohort on some outcome are *mediated* by a set of mechanistic explanatory variables. By Pearl's "front door criterion" (Pearl, 2009), as long as we observe *all* of the relevant mediators for at least one of time or cohort, then the age profile is identified.

One example of a causal diagram that would satisfy these requirements is illustrated in Figure B.1. A variety of assumptions are encoded in this diagram: We are assuming that the time effect acts *only* through unemployment, and the cohort effects act *only* through education. We are also assuming that age and cohort have no effect on unemployment (except through time), and age and time have no effect on education (except through cohort).

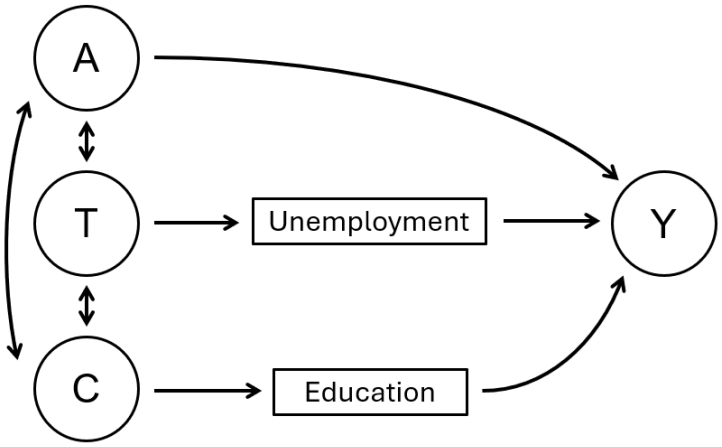


Figure B.1: Simple Example with Mediators

Figure B.2 presents a more complicated causal diagram, with a mechanism that is influenced by both time and cohort (cohort size), and with multiple levels of mediators. In this setting, the age profile would still be identified, although the estimation strategy would be more complicated. See Winship and Harding (2008) for details.

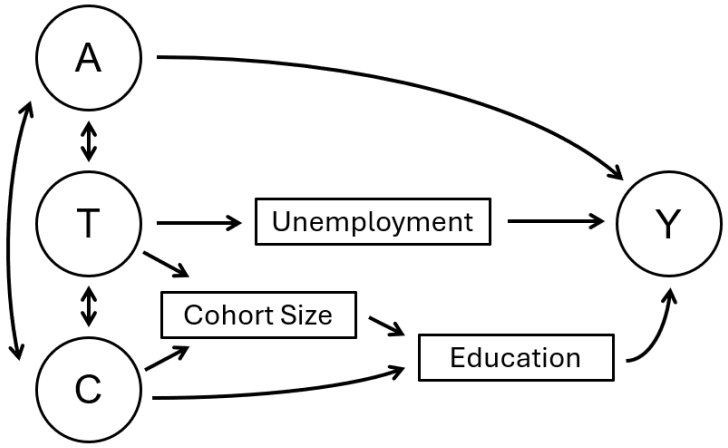


Figure B.2: More Complex Example with Mediators

B.2 Our Approach: A Simple Proxy Assumption

Our approach is a special case that Winship and Harding (2008) call the "proxy variable" approach. The original idea of proxying for time effects goes back to at least Farkas (1977) who replace calendar year with the unemployment rate but without giving any formal justification. Heckman and Robb (1985) provides a formal justification from a structural perspective, and introduces the term "proxy variable."

Figure B.3 presents a diagram for the structure we assume. Notice that we do not do any modeling of the mechanisms that sit between age and the outcome, or between cohort and the outcome. This is essentially without loss of generality, because we only need to cut-off a single one of A, T, or C from Y in order to identify the age profile. We choose T because we think the macroeconomic mediators for T are easier to identify than those for A or C. Importantly, this structure still allows a hypothetical additional mechanism like education to effect Y, we just assume that it is captured by a combination of age and cohort effect.

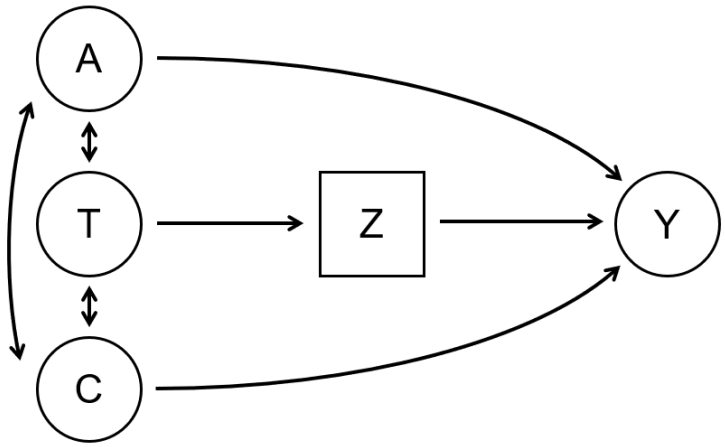


Figure B.3: Proxy Variable Structure

Ultimately, this causal diagram boils down to two main substantive assumptions. First, an exclusion restriction: we assume that, conditional on A and C, T has no effect on Y except through Z. For example, this means there are no additional unobserved macroeconomic variables that cause Y. Second, we assume that neither A nor C effect Z. Intuitively, we usually think of a macro-variable like the unemployment rate as being a pure time effect. The same assumption is made in the main application in Winship and Harding (2008). But while this may not be a particularly strong assumption, it does still rule out, e.g. an age composition effect on the unemployment rate. This also rules out diagrams like Figure B.2, where T affects Y through education via cohort size, but both cohort size and education are also affected by C.

We could relax this second assumption slightly: the proxies Z can be affected by *at most* two of A, T, and C. This would introduce additional complexity in the form of a first stage regression, as discussed in Winship and Harding (2008). As a consequence, we make the stronger assumption embedded in Figure B.3. Extending our machine learning estimator to work for any causal diagram compatible with Winship and Harding (2008) — such as Figure B.2 — is one possible direction for future work. Of course, if any of the proxies Z are affected by all three (A, C, T) the model is no longer identified. In this case, we would need to find intermediate proxies that are plausibly affected by only two of the three.

Finally, the front-door criterion requires stochasticity of the mediators Z.[11] In other words, the proxies Z cannot be perfectly deterministic functions of T—otherwise we'd back to some form of multicollinearity. This connects directly to much of the discussion in our technical companion paper. For the current paper, this motivates our requirement that any proxy variables must have an R^2 of less than 0.75 in the regression of T on Z.

While the approach we adopt relies on non-trivial assumptions, these assumptions are much weaker than the assumptions made in most existing work that seeks to solve the age-time-cohort problem in economics.

Note that there are other conceptual frameworks for reasoning about counterfactuals. We could have formulated our requirements for identification in terms of the potential outcomes framework following Imbens and Rubin (2015). This would take the form of a SUTVA assumption on potential outcomes in terms of age combined with a "no unobserved confounders" assumption conditional on c and z. We explore this connection in the technical companion paper.

C Intuition for Riesz Representers and Double Machine Learning

C.1 Double/Debiased Machine Learning

The machine learning prediction task (by minimizing the mean squared error) tries to estimate the conditional expectation:

$$m_0(a, c, z) \coloneqq \mathbb{E}[\sigma^2_{at}|a, c, z].$$

Typically, our final machine learning model $\hat{m}(a, c, z)$ introduces bias to reduce the variance of the estimator. Mechanically, this means that we do not make the in-sample prediction residuals as small as we possibly could, given a highly flexible machine learning function class. This is usually

[11] We thank Lihua Lei for pointing this out.

necessary to control overfitting and successfully generalize out of sample.

The fundamental idea behind double/debiased machine learning is to exploit the fact that our actual estimand of interest is the slope, slope(m_0)—a single scalar—which is much simpler than the whole function $m_0(a, c, z)$. We do not need a completely unbiased estimate of the entire function m_0 (which would have prohibitively large variance). We only need an unbiased estimate of slope(m_0). This should be easier to achieve.

Above, we characterize the bias of machine learning estimators as being due to not making the prediction residuals as small as possible. For example, while OLS produces the smallest possible residuals among linear models, Lasso and Ridge are regularized, creating bias toward a simpler model and resulting in larger residuals. Double/debiased machine learning cleverly addresses this regularization bias by correcting the prediction residuals *but only if they matter for the slope*. This works by adding a bias correction term to the initial estimate of the slope, slope$(\hat{m})$:

$$\text{slope}(\hat{m}) + \mathbb{E}\Big[\nabla\text{slope}(a, c, z) \cdot (\underbrace{\sigma^2_{at} - \hat{m}(a, c, z)}_{\text{prediction errors}})\Big]. \tag{10}$$

What is ∇slope(a, c, z)? It is the gradient of the slope when viewed as an operator, slope$(\cdot)$, that takes a *function* of a, c, z such as $\hat{m}$ as its input. Note that because the input to slope$(\cdot)$ is a function of a, c, z, the gradient ∇slope is also a function of a, c, z. We will define this gradient more formally in the next section. For now think of this gradient intuitively: it measures how much changing $\hat{m}(a, c, z)$ at a particular value of a, c, z impacts the final estimate slope$(\hat{m})$. For example, we would expect that the observations where a is at the extremes of the age range would be more impactful for the slope than observations in the middle of the age range. Using this interpretation, when we do debiasing in Equation (10), we correct the residuals more if the prediction at that point is very influential for the final slope estimate.

To understand why (10) corrects bias, notice that it takes the form of a first-order Taylor expansion. The familiar Taylor approximation from elementary calculus is: $f(x) \approx f(a) + f'(a)(x - a)$. In expression (10), slope$(\cdot)$ corresponds to $f(\cdot)$, and $\hat{m}$ corresponds to a. But what about x? Here we can apply the law of iterated expectations: $\mathbb{E}[\sigma^2_{at}] = \mathbb{E}[\mathbb{E}[\sigma^2_{at}|a, c, z]] = \mathbb{E}[m_0(a, c, z)]$. When we plug this expression back into (10), we get:

$$\text{slope}(\hat{m}) + \mathbb{E}\Big[\nabla\text{slope}(a, c, z) \cdot (m_0(a, c, z) - \hat{m}(a, c, z))\Big].$$

So the debiased estimate is a first-order Taylor approximation of the true population estimand

slope(m_0), where the Taylor approximation is taken around our initial, biased machine learning estimate, slope$(\hat{m})$. Furthermore, we can compute this bias correction term even though we do not know the true m_0 in advance. There are deep reasons from empirical process theory why this formally works and provides a "semiparametrically efficient" estimate (Kennedy, 2022). However, this description is a simple intuitive way to think about the whole process:

1. We want to correct the residuals but only if they matter for estimating the slope.
2. One way to do this is to multiply the residuals by the "gradient of the slope with respect to the observation".
3. We end up with a Taylor series approximation for the true estimand, slope(m_0).

C.2 "Riesz Representers" Are Gradients with Respect to Function-Valued Inputs.

In the previous section, we describe a Taylor approximation using ∇slope(a, c, z). This is the gradient of the slope operator, slope$(\cdot)$, with respect to its input, a function of a, c, z. The crucial observation is that the slope is a *linear* functional. For any two functions $m_1(a, c, z)$ and $m_2(a, c, z)$, and any $a, b \in \mathbb{R}$,

$$\text{slope}(am_1 + bm_2) = a \cdot \text{slope}(m_1) + b \cdot \text{slope}(m_2).$$

The gradient of a linear functional with respect to its function-valued input is called the "Riesz representer."

Linear functionals have a special property. Just like a linear function defined on $\mathbb{R}^d$ can be written in terms of coefficients in $\mathbb{R}^d$, a linear functional defined over functions $m(a, c, z)$ can be written in terms of "coefficients" that are themselves a function of a, c, z. This isn't true for completely arbitrary spaces of functions. In particular, we need the following restriction on the functions m that we consider: $\mathbb{E}[m(a, c, z)^2] < \infty$. We denote the set of functions that satisfy this condition L_2. Functions in L_2 have an inner product just like vectors in $\mathbb{R}^d$ have an inner product. This inner product has the following form. For any $f, g \in L_2$:

$$\langle f, g \rangle \coloneqq \mathbb{E}[f(a, c, z)g(a, c, z)].$$

Because the slope is a linear functional, there exist unique "coefficients" in L_2 that we will denote $\alpha_0(a, c, z)$ such that:

$$\langle \alpha_0, m \rangle = \text{slope}(m), \forall m \in L_2. \tag{11}$$

In $\mathbb{R}^d$, the gradient of a linear function is equal to the coefficients in $\mathbb{R}^d$. Similarly, the gradient $\nabla \text{slope}(a, c, z) = \alpha_0(a, c, z)$. Spaces of general functions that have an inner product are called "Hilbert spaces" and the "coefficients" in these spaces are called the Riesz representer. For a review of Riesz representers for an applied audience in epidemiology see Williams et al. (2025).

C.3 Riesz Representer for the Slope Operator

So far, we have reviewed how to debias a machine learning estimate using the Riesz representer. However, we have discussed the Riesz representer abstractly. In this section, we present the Riesz representer for our particular object of interest: the slope of the age profile. It turns out that the Riesz representer for the slope in our setting has an explicit analytic form. In this section, we present the analytic form but defer the derivation.

Define:

$$\vec{a} \coloneqq \{35, ..., 50\}, \tag{12}$$

$$s \coloneqq \frac{\vec{a} - \bar{a}}{(\vec{a} - \bar{a})^\top (\vec{a} - \bar{a})}, \tag{13}$$

where $\bar{a}$ is the mean of $\vec{a}$. Let $s(a)$ denote the entry of s corresponding to age a.

Then the Riesz representer of the slope is:

$$\alpha_{\text{slope}}(a, c, z) = \frac{s(a)}{P(T = c + a | Z = z)}. \tag{14}$$

There are two contributions to the size of α_{slope}: the numerator $s(a)$ and the denominator $P(T = c + a|Z = z)$, the probability of T given z. Next, we will interpret these in turn.

The entries of the numerator, $s(a)$, sum to zero with the largest magnitude values at the earliest and latest ages. Recall that when $\alpha_{\text{slope}}(a, c, z)$ is large for a given observation, then that observation is especially impactful for estimating the slope. Intuitively, the numerator reflects the fact that the slope estimate is the most sensitive to observations with ages farthest from the mean age. In fact, notice that from the usual simple linear regression math, if we take the inner product of s with any other vector y of the same length, we get the slope of y.

The denominator is especially important, and provides a hint as to how proxies achieve identification. The slope is identified if and only if $P(T = c + a|Z = z)$ is greater than zero; otherwise the weights can be infinite. Using the gradient intuition, if the weights are infinite, it means that an arbitrarily small change at that data point can cause an arbitrarily large change to the slope

estimate — this is precisely what it means to be unidentified.

Consider using a time trend as the proxy Z, i.e. $Z = T$. In this case there is no uncertainty about time conditional on the proxy; $P(T = c + a|Z = z)$ either equals 0 or 1. Therefore the denominator in (14) is not always greater than 0 and the slope is not identified.

Proxy variables can only achieve identification by introducing some uncertainty into the relationship between Z and T. First, notice that this motivates our use of blocked cross-validation. Consider trying to estimate $P(T = c + a|Z = z)$ from the data. Without sample splitting, for continuous valued proxies there is usually a bijective map between Z and T. E.g., consider taking Z to be inflation. Inflation in the US in the year 2005 was 3.39275. There is no other year in which inflation takes on *precisely* this value, so without any sort of sample splitting, we could construct a map that predicts T given inflation with perfect accuracy. $P(T = c + a|Z = z)$ would again always be 0 or 1, and the slope would be unidentified. Instead, if we always split the observations into train and test folds by years, then it is generally not possible to perfectly predict T given Z out-of-sample. As a result $P(T = c + a|Z = z)$ will be greater than 0, and the slope becomes identified.

Second, notice that this means that proxies that are *approximately* but not perfectly trending, such as GDP, yield very poor identification. When Z is GDP, then $P(T = c + a|Z = z)$ is always close to 0 or 1, even when sample splitting by year. All identification comes from the deviations around the trend, and these observations correspond to large values of the Riesz representer. In other words, the estimate of the slope is driven by these deviations from the trend. During debiasing, the Riesz representer is applied as a set of weights, and reweighting by large values translate into large standard errors — correctly reflecting poor identification. For time series like unemployment or inflation, it is generally harder to predict T as a function of Z, resulting in a $P(T = c + a|Z = z)$ that is strictly bounded away from 0. In this case the Riesz representer does not take on extremely large values, and we achieve better identification.

Therefore, the discussion above based on (14) motivates our choices to exclude proxy variables that are very strongly trending (with an $R^2 > 0.75$ in the regression of T on that proxy), and to perform cross-validation by splitting whole years into train and test folds.

Finally, note that instead of minimizing the Riesz loss as described in the previous section, we could instead estimate the conditional density $P(T|Z)$ and then plug in our estimate into (14) to compute the Riesz representer for debiasing. This has two problems. First, conditional density estimation becomes statistically very challenging as the dimensionality of Z grows. Second, taking the inverse in (14), $s(a)/\hat{P}(T = c+a|Z = z)$, has undesirable properties. Even if $\hat{P}(T = c+a|Z = z)$

can be estimated well, we could end up with a poor estimate of $1/\hat{P}(T = c + a|Z = z)$ due to instability for small values of $\hat{P}(T = c+a|Z = z)$. For these reasons, we prefer to directly estimate the Riesz representer using the Riesz loss. However, (14) provides important intuition about the underlying structure of the identification problem.

C.4 Estimating the Riesz Representer

Finally, we comment on how we estimate $\hat{\alpha}$. We discussed above how α_0 for the slope functional has a closed-form expression. Note that in the expression (14), the only unknown quantity is the conditional probability in the denominator — $P(T|Z)$. One simple way to estimate α_0, would be to first get an estimate of this probability, $\hat{P}(T|Z)$, and then to compute the "plug-in" estimate:

$$\hat{\alpha}(a, c, z) = \frac{s(a)}{\hat{P}(T = c + a|Z = z)}.$$

This strategy is called "inverse probability weighting" and is used in Chernozhukov et al. (2018).

However, the recent literature on debiased machine learning has emphasized that this is usually *not* a reliable way to estimate the Riesz representer—see for example Ben-Michael et al. (2021) and Chernozhukov et al. (2022). Among other issues, inverting the estimated probabilities means that when $P(T|Z)$ is close to zero, very small estimation errors can lead to enormous estimation errors in $\hat{\alpha}$.

Instead in this paper, we adopt the "Riesz loss" from Chernozhukov et al. (2022). The true Riesz representer α_0 is the *unique* minimizer of the optimization problem:

$$\min_{\alpha \in L_2} \{\mathbb{E}[\alpha(a, c, z)^2 - 2 \cdot \text{slope}(\alpha)]\}. \tag{15}$$

Minimizing this loss is equivalent to minimizing the squared-loss between α and α_0:

$$\begin{aligned}
&\min_{\alpha \in L_2} \mathbb{E}[(\alpha(a, c, z) - \alpha_0(a, c, z))^2] \\
&\quad = \min_{\alpha \in L_2} \mathbb{E}[\alpha(a, c, z)^2 - 2 \cdot \alpha(a, c, z)\alpha_0(a, c, z) + \alpha_0(a, c, z)^2] \\
&\quad = \min_{\alpha \in L_2} \mathbb{E}[\alpha(a, c, z)^2] - 2 \cdot \mathbb{E}[\alpha(a, c, z)\alpha_0(a, c, z)] \\
&\quad = \min_{\alpha \in L_2} \mathbb{E}[\alpha(a, c, z)^2] - 2 \cdot \langle \alpha, \alpha_0 \rangle \\
&\quad = \min_{\alpha \in L_2} \mathbb{E}[\alpha(a, c, z)^2 - 2 \cdot \text{slope}(\alpha)],
\end{aligned}$$

where we applied the key property of the Riesz representer (11) for:

$$\text{slope}(\alpha) = \langle \alpha, \alpha_0 \rangle, \forall \alpha \in L_2.$$

Finally, we connect this optimization problem back to the original intuition, where we referred to the Riesz representer as the gradient $\nabla \text{slope}(a, c, z)$. If we take the first order condition of the Riesz loss optimization problem (15), we get that the solution α^* must satisfy:

$$2\alpha^*(a, c, z) - 2\nabla \text{slope}(a, c, z) \implies \alpha^*(a, c, z) = \nabla \text{slope}(a, c, z).$$

D Details for Debiased Machine Learning

D.1 Fitting the Outcome Models

As we discuss in section 5.1, we select 10 proxies that have the strongest relationship with the outcome variable based on a univariate cross-validated kernel ridge regression model. We then train our models on all combinations of these ten chosen candidate features. For each sublist, we also include either age and cohort trends or age and cohort dummies. We therefore train our models on $2 \times (2^{10} - 1) = 2046$ different feature matrices for each country.

For each feature matrix, we remove each column's median and rescale by the interquartile range using sklearn's RobustScaler. For the dummy variable columns we center using the mean and rescale with the standard deviation instead. We train various models over a range of hyperparameters. We use log-spaced grids for hyperparameters. Denote an evenly spaced grid of n elements between endpoints x_0 and x_1 as: $g(x_0, x_1, n)$. Let $\text{log\_grid}(x_0, x_1, n)$ be the grid containing 10 to the power of i for $i \in g(\log x_0, \log x_1, n)$. We train the following models:

- Ridge regression with hyperparameters $\lambda \in \text{log\_grid}(\text{1e-7}, 20, 100)$
- Lasso with hyperparameters $\lambda \in \text{log\_grid}(\text{1e-7}, 20, 100)$
- Kernel ridge regression with $\lambda \in \text{log\_grid}(\text{1e-7}, \text{2e-1}, 20)$, $\gamma \in \text{log\_grid}(\text{1e-6}, \text{1e-1}, 10)$, with both RBF and Laplacian kernels.

We use an evenly spaced grid in logs because the variance can grow very rapidly as the regularization parameters near 0, so we want the grid points to be closer together near 0. This is standard practice in machine learning.

Finally, note that we choose relatively coarse grids with only 100 different combinations of hyperparameters per method. This saves computation in our initial pass that performs this procedure on different combinations of feature matrices. However, after we have found the best combination of feature matrix and hyperparameter for ridge/lasso/KRR, we then go back and refine the hyperparameters on only these feature matrices with finer grids containing 10x more points. Again, this coarse-to-fine strategy is standard practice in machine learning.

D.2 The Choice of Machine Learning Estimators

As we describe in Section 5.1, we fit each machine learning algorithm (ridge, kernel ridge, etc) on every combination of features. With unlimited compute, we would run any conceivable machine learning algorithm, including tree-ensembles and neural networks, on each of the feature matrices. However, in practice, we find that given our very small number of sample sizes (in the 100s) more complicated methods like tree-ensembles and neural networks have two problems: (1) they perform poorly across the board compared to simpler models, and (2) they take much longer to run. Therefore, we exclude these more complicated models from our full run.

To demonstrate these two claims, in this section we present results with tree-ensembles and neural networks on the GRID data for the United States. We begin by selecting the features that yielded in the best results under our main specification. These are: Age, Cohort, Square-root of Age, Log Oil Prices, Lagged Log Oil Prices, and Lagged Log Industrial Production Squared. The best performing model from our main specification is Kernel Ridge Regression, using the Laplacian kernel with bandwidth 0.0328 and regularization penalty 1.86×10^{-5}. This model achieves a cross-validated root mean squared error of 0.013. We compare this result to several tree and neural network specifications in Table D.1.

Table D.1: Comparison of Tree and Neural Network Models to Kernel Ridge Regression

Machine Learning Method	Cross-validated RMSE
Kernel Ridge Regression	**0.013**
Random Forest	0.022
Gradient-Boosted Trees	0.021
Multilayer Perceptron	0.111
RealMLP	0.021

Note: Inputs to the models are: Age, Cohort, Square-root of Age, Log Oil Prices, Lagged Log Oil Prices, and Lagged Log Industrial Production Squared. For cross-validation, we use year-blocked splits of 8 contiguous years.

The much simpler kernel ridge regression method achieves *nearly half* the RMSE of the more

complicated models. Note that the typical neural network used for regression in small datasets, the Multilayer Perceptron (MLP) performs particularly poorly. RealMLP (Holzmüller et al., 2024) is a recent improved architecture for small datasets that incorporates an extensive series of tricks and transformations — even this state-of-the-art neural architecture performs poorly in this setting compared to kernel ridge.

Finally, kernel ridge regression tends to suffer in high dimensions. This is one reason why we have to iterate through all possible subsets of the original features to find the best model. Tree ensembles and neural networks on the other hand typically perform well with high-dimensional inputs. Therefore, we also test these models using all available features including all transformations of age, cohort, and all proxy variables. For the GRID US data, this is 53 features. The results are presented in Table D.2

Table D.2: Tree and Neural Network Models with High-Dimensional Input

Machine Learning Method	Cross-validated RMSE
Random Forest	0.023
Gradient-Boosted Trees	0.024
Multilayer Perceptron	0.722
RealMLP	0.025

Note: Models are trained using all 53 input features (excluding only those proxies with an $R^2 > 0.75$ in the regression of calendar year on that proxy). For cross-validation, we use year-blocked splits of 8 contiguous years.

Reflecting the difficulties of the very small sample size, the results are always worse than what was achieved in the more carefully selected subset of features in Table D.1.

D.3 Fitting the Weighting Models

In this section, we provide details on how we estimate the weights $\hat{w}$. In general, these weights are found by minimizing the loss from Chernozhukov et al. (2022):

$$\min_{w \in \mathcal{F}} \left\{ E[w(A, C, Z)^2 - 2 \cdot \text{slope}(w)] \right\}. \tag{16}$$

In the most general case, $\mathcal{F}$ can be an arbitrary function class. Different choices of $\mathcal{F}$ will results in different estimates with different properties. For the applications used in Chernozhukov et al. (2022), $\mathcal{F}$ is the set of linear functions whose coefficients are within an ℓ_1-norm ball — just as in the Lasso. See Chernozhukov et al. (2022) for details on the "Lasso" type weights.

In this work, we consider two function classes that can be solved analytically: linear functions with coefficients in an ℓ_2-norm ball (ridge weights), and functions within an RKHS ball

(kernel ridge weights). Access to a closed-form solution offers substantial computational advantages. These ridge-type estimators of the Riesz representer have been studied theoretically in Singh (2021).

In the subsections below, we will walk through the derivation of the closed-form solutions for ridge weights and kernel ridge weights.

But other than the loss function, we estimate the weights with exactly the same machine learning setup as for our basic prediction problem. This includes both the choice of features, and using blocked cross-validation for model selection. We train models for the weights with the following hyperparameters:

1. Ridge weights with $\delta \in \text{log\_grid}(\text{1e-7}, \text{5e-1}, 25)$.
2. Kernel Ridge weights with with $\delta \in \text{log\_grid}(\text{1e-7}, \text{5e-1}, 20)$, $\gamma \in \text{log\_grid}(\text{1e-6}, \text{1e-1}, 10)$, with the RBF kernel.

As in Appendix D.1, we refine the hyperparameters on a finer grid for the best performing feature matrix / model combinations.

D.3.1 Linear Weights

First, we consider linear function classes with ℓ_2-norm regularization, aka Ridge for weighting. Write $X(A, C, Z) \in \mathbb{R}^{n \times d}$ for a design matrix with a row for each observation, and columns for each feature. For example, we might have $d = 3$ with simple linear features A, C, and Z. Or we might have $d = 5$ with features A, A^2, C, C^2, Z. Note that we have written X as a function of A, C, Z. This will be important later when computing the optimal weights. When this dependence on A, C, Z is not important for the context, we will simply write X for brevity. Without loss of generality, we assume that X has mean zero, otherwise we can just recenter first.

The loss in the linear case is:

$$\min_{\beta \in \mathbb{R}^d} \{ \frac{1}{n} \beta^\top X^\top X \beta - 2\text{slope}(X\beta) + \delta \beta^\top \beta \}, \tag{17}$$

where δ is a hyperparameter. To get the optimal weights, we will exploit the fact that because slope is a linear functional, there must be a way to write the slope as a linear function of β. We derive this form next.

Recall that our estimand is the slope of the age profile. Then by straightforward OLS math, the slope of the age profile is a linear combination of the elements of the age profile. In particular,

recall that the value of the age profile at age a is $\sigma_a^2 = \mathbb{E}[m(a, C, Z)]$. Let $\sigma_{\vec{a}}^2$ be the full age profile as a vector whose elements are $\hat{\sigma}_a^2$. Furthermore, define $\vec{a} \coloneqq \{a_{\text{min}}, ..., a_{\text{max}}\}$ and let $\bar{a}$ denote the mean of $\vec{a}$. Then, we have:

$$\text{slope} = \frac{(\vec{a} - \bar{a})^\top \sigma_{\vec{a}}^2}{(\vec{a} - \bar{a})^\top (\vec{a} - \bar{a})}.$$

Denote:

$$s \coloneqq \frac{(\vec{a} - \bar{a})}{(\vec{a} - \bar{a})^\top (\vec{a} - \bar{a})}.$$

Notice that for m of the form $X\beta$ we have $\sigma_a^2 = \mathbb{E}[\beta^\top X(a, C, Z)]$. Putting this all together:

$$\begin{aligned} \text{slope}(X\beta) &= s^\top \mathbb{E}[\beta^\top X(\vec{a}, C, Z)] \\ &= \beta^\top \mathbb{E}[s^\top X(\vec{a}, C, Z)], \end{aligned}$$

where $\mathbb{E}[s^\top X(\vec{a}, C, Z)] =: \bar{X}_{\text{cf}} \in \mathbb{R}^d$ are the average "counterfactual features," to borrow the language from Chernozhukov et al. (2020). Not that when we are solving for the weights in practice, the expectation in the definition of $\bar{X}_{\text{cf}}$ will be the sample average.

Now we are ready to get a closed-form solution to Equation (17). Using the first order conditions, we have the coefficients for the optimal weights:

$$\beta^* = n(X^\top X + \delta I)^{-1} \bar{X}_{\text{cf}}.$$

These closed-form coefficients can be computed quickly in just a few lines of code. The first term, $(X^\top X + \delta I)^{-1}$ is just the standard inverse regularized covariance matrix from ridge regression. The only new term we must compute is $\bar{X}_{\text{cf}}$. For each observation, we have to compute the counterfactual features $X(a, C_i, Z_i)$ and take their inner product with s. Then we just take the sample average of the result.

The final estimated weighting function is:

$$\hat{w}(x) = x^\top \beta^*.$$

D.3.2 Kernel Weights

We adopt the same setup with a feature matrix X as for linear weights, but now we will solve the weighting optimization problem over an Reproducing Kernel Hilbert Space (RKHS) — see

Scholkopf and Smola (2018) for a review. We will broadly adopt the setup from Hirshberg et al. (2019); Singh (2021); Bruns-Smith et al. (2023). Let $\mathcal{H}$ be a possibly-infinite-dimensional RKHS with kernel $\mathcal{K}$. Let $\|\cdot\|_{\mathcal{H}}$ denote the norm of $\mathcal{H}$. Let K be the matrix with entries $\mathcal{K}(x_i, x_j)$, where x_i, x_j are the ith and jth entries of X. Then the weighting problem is:

$$\min_{a \in \mathbb{R}^d} \{\frac{1}{n}\alpha^\top K^\top K\alpha - 2\text{slope}(K\alpha) + \delta\alpha^\top K\alpha\}$$

where

$$\text{slope}(K\alpha) = \alpha^\top \bar{K}_{\text{cf}},$$

and the counterfactual kernel matrix $\bar{K}_{\text{cf}}$ has a closed-form. Define K_a to be the matrix with entries $\mathcal{K}(x_i, x'_j)$ where x_i is the ith row of $X(A, C, Z)$ and x'_j is the (counterfactual) jth row of $X(a, C, Z)$. Define $\bar{K}_a \in \mathbb{R}^n$ to be the average for each row of K_a. Define $\bar{K}_{\vec{a}} \in \mathbb{R}^{n \times n_{\text{age}}}$ to be the matrix whose columns are the $\bar{K}_a$ for each a. Then:

$$\bar{K}_{\text{cf}} \coloneqq \bar{K}_{\vec{a}} s,$$

for s defined as in Appendix D.3.1. This follows directly from algebra using the definition of the slope applied to the kernel ridge model with dual coefficients α.

There is a closed-form solution to the optimization problem:

$$\alpha^* = n(K^T K + \delta K)^{-1} \bar{K}_{\text{cf}}.$$

For a single point x, define K_x as the vector with entry $\mathcal{K}(x_i, x)$ for each row x_i in X. Then the resulting estimated weighting function is:

$$\hat{w}(x) = K_x^\top \alpha^*.$$

D.4 Standard Errors for the Slope

We now describe how to compute asymptotic normal standard errors for the debiased machine learning estimate. We do this in the usual way taken directly from Chernozhukov et al. (2022), but give details here for concreteness.

Let $\text{slope}(\hat{m}, c, z)$ denote the slope of model $\hat{m}$ evaluated at a fixed c and z. In other words, we take the age profile at a fixed c and z, $\hat{m}(35, c, z), \hat{m}(36, c, z), ..., \hat{m}(50, c, z)$, and then compute the slope of the best fit line through this profile. Then by linearity, we have $\text{slope}(\hat{m}) =$

$E[\text{slope}(\hat{m}, c, z)]$.

Then our debiased point estimate using predictive model $\hat{m}$ and weighting model $\hat{w}$ can be written:

$$\hat{\theta}_{\text{dr}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big\{\text{slope}(\hat{m}, c_i, z_i) + \hat{w}(a_i, c_i, z_i)(\sigma_i^2 - \hat{m}(a_i, c_i, z_i))\Big\} \tag{18}$$

This is a sample average, and so the asymptotic variance is estimated in the usual way:

$$\hat{V}_{\text{dr}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big\{\big(\text{slope}(\hat{m}, c_i, z_i) + \hat{w}(a_i, c_i, z_i)(\sigma_i^2 - \hat{m}(a_i, c_i, z_i)) - \hat{\theta}_{\text{dr}}\big)^2\Big\}$$

and our asymptotic normal standard error is $\sqrt{\hat{V}_{\text{dr}}/n}$.

This contrasts with the biased machine learning plug-in estimator:

$$\hat{\theta}_{\text{ml}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big\{\text{slope}(\hat{m}, c_i, z_i)\Big\}$$

and the naive standard errors that would result from erroneously assuming that $\hat{\theta}_{\text{ml}}$ is unbiased:

$$\hat{V}_{\text{ml}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big\{\big(\text{slope}(\hat{m}, c_i, z_i) - \hat{\theta}_{\text{ml}}\big)^2\Big\}$$

The standard errors using $\hat{V}_{\text{ml}}$ will not generally yield a confidence interval with the desired coverage, as we illustrate in simulation in Appendix G.

Note that the debiased variance is inflated by the inclusion of the weights and residuals. Ignoring cross-terms, we can use the following rough bound for intuition:

$$\begin{aligned}\hat{V}_{\text{dr}} &\approx \hat{V}_{\text{ml}} + \frac{1}{n}\sum_{i=1}^{n}\hat{w}(a_i, c_i, z_i)^2(\sigma_i^2 - \hat{m}(a_i, c_i, z_i))^2 \\ &\leq \hat{V}_{\text{ml}} + \underbrace{\left(\sum_{i=1}^{n}\hat{w}(a_i, c_i, z_i)^2\right)}_{\text{size of } \hat{w}}\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}(\sigma_i^2 - \hat{m}(a_i, c_i, z_i))^2\right)}_{\text{MSE of } \hat{m}}.\end{aligned}$$

In other words, the difference between the debiased confidence interval and the naive confidence interval depends roughly (1) the squared sum of the weights, and (2) the mean squared error of the predictor $\hat{m}$. If the predictor $\hat{m}$ were perfect, we wouldn't need to perform any bias correction. If the predictor is imperfect, then intuitively the weights $\hat{w}$ control how the uncertainty from the

predictions passes through to our uncertainty about $\hat{\theta}_{\text{dr}}$.

D.5 Cross-Fitting

Notice that the added bias correction term in Equation (18) takes the form:

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\hat{w}(a_i, c_i, z_i)(\sigma_i^2 - \hat{m}(a_i, c_i, z_i))\Big\}. \tag{19}$$

When the prediction residuals $\sigma_i^2 - \hat{m}(a_i, c_i, z_i)$ are small, then the impact of bias correction is also small. This can be an issue if we use the data to fit the model and then subsequently use the same data to calculate the residuals. In the worst-case, we could severely overfit, driving the residuals to zero. In practice, we select $\hat{m}$ using blocked cross-validation, so we will never drive the residuals to zero. However, there could be a concern that the residuals would still be too small, since we use the same data to select $\hat{m}$ and to compute the residuals. This has been an area of active debate in the debiased machine learning literature. Theoretically, when the machine learning algorithm used to select $\hat{m}$ satisfies a "Donsker condition" (van der Vaart and Wellner, 2023), then reusing the samples is not a problem. However, the Donsker condition essentially requires that the algorithm cannot be too flexible, which may be violated by many modern machine learning tools. Therefore many debiased machine learning papers (e.g. see Chernozhukov et al., 2018) suggest using "cross-fitting", where the models are trained in folds and then applied pseudo-out-of-sample as with cross-validation. A recent simulation study in Bach et al. (2024) finds that using the whole sample or doing cross-fitting perform roughly the same, but to guarantee that our confidence intervals are valid, we choose to use cross-fitting for our final estimates. Unfortunately, this does make notation a bit messier, and so we suppress the details from the main text and present them below.

We define folds according to our blocked cross-validation scheme. The test folds are observations from within blocks of 8 contiguous years. Recall from Section 5.2 that we do not use a single fixed partition into folds. Instead we consider each possible combination of 8 contiguous years as a test fold — a sliding window.

Denote each of these test folds ℓ. Let L be the total number of folds. If B is the block size, then $L = n_{\text{year}} - B + 1$ using our sliding window folds. Let I_ℓ denote the indices of the observations within fold ℓ and let $I_{\neg\ell}$ denote the indices of observations outside of fold ℓ. Let $\hat{m}_\ell$ and $\hat{w}_\ell$ be the predictors and weights fit on the data in $I_{\neg\ell}$. Let $X_i \coloneqq \{A_i, C_i, Z_i\}$ denote the data for observation i.

Then the cross-fit point estimate is:

$$\hat{\theta}_{\text{dr-cf}} \coloneqq \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{|I_\ell|}\sum_{i\in I_\ell}\Big\{\text{slope}(\hat{m}_\ell, c_i, z_i) + \hat{w}_\ell(a_i, c_i, z_i)(\sigma_i^2 - \hat{m}_\ell(a_i, c_i, z_i))\Big\}, \tag{20}$$

and the cross-fit variance is:

$$\hat{V}_{\text{dr-cf}} \coloneqq \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{|I_\ell|}\sum_{i\in I_\ell}\Big\{\big(\text{slope}(\hat{m}, c_i, z_i) + \hat{w}(a_i, c_i, z_i)(\sigma_i^2 - \hat{m}(a_i, c_i, z_i)) - \hat{\theta}_{\text{dr-cf}}\big)^2\Big\}. \tag{21}$$

Finally, we compute the standard errors:

$$\widehat{\text{std err}}_{\text{cf}} = \sqrt{\hat{V}_{\text{df-cf}}/n}\,. \tag{22}$$

D.6 Clustered Standard Errors

Our blocked cross-validation scheme splits the data by whole years. This guarantees that observations from the same year (which are strongly dependent) do not simultaneously appear in both a training set and test set. For the same reason, we also extend (21) to handling clustering at the year level.

Let $I_{\ell,t}$ be the index set of observations in year t and fold ℓ. We compute the squared cluster sums:

$$g_{\ell,t}^2 = \left(\sum_{i\in I_{\ell,t}}\big(\text{slope}(\hat{m}, c_i, z_i) + \hat{w}(a_i, c_i, z_i)(\sigma_i^2 - \hat{m}(a_i, c_i, z_i)) - \hat{\theta}_{\text{dr-cf}}\big)\right)^2$$

Let g_t^2 denote the average of $g_{\ell,t}^2$, averaging over all folds ℓ that contain t.

Then our estimate of the asymptotic variance (including finite sample correction for our number of clusters, n_{year}) is:

$$\hat{V}_{\text{dr-cluster}} \coloneqq \left(\frac{n_{\text{year}}}{n_{\text{year}} - 1}\right)\frac{1}{n}\sum_t g_t^2.$$

E Debiased Estimate of the Age Profile

E.1 Estimating the Full Age Profile

In the main text and in Appendix D, we have discussed estimation of the slope of the age profile. Beyond estimating the slope, we can also debias each point σ_a^2 on the age profile individually to

obtain a full debiased age profile. For fixed values of the hyperparameter and proxies used for the weights, the two are *equivalent*. That is, if we were to debias each point individually, and then compute the slope of the resulting age profile, this is numerically equivalent to the debiased estimate of the slope.

We use this fact to plot a corresponding debiased age profile for each of our slope point estimates — these are what we plot in Figure 5 and Figure 7, and what we use to get the "Debiased ML" columns of Table 2 and Table 3. We take the cross-validated kernel weighting hyperparameters and proxy variables that are selected for the slope, and use them to debias each point on the age profile individually like so:

Consider a fixed point on the age profile. For concreteness, we'll use σ^2_{35}, the value of the age profile at age 35. The weighting problem for debiasing σ^2_{35} is:

$$w^*_{35} = \underset{w}{\text{argmin}}\big\{E[w(a,c,z)^2 - 2 \cdot w(35,c,z)]\big\}. \tag{23}$$

We can then run exactly the same debiased machine learning procedure that we outlined for the slope, but for σ^2_{35}. The prediction problem for σ^2_{at} is the same, so we will use the same predictive model, $\hat{m}$. If we have a weighting model $\hat{w}_{35}$, the resulting debiased estimate of the age profile at age 35 is:

$$\hat{\sigma}^2_{35} := \frac{1}{n}\sum_{i=1}^{n}\Big\{\hat{m}(35,c_i,z_i) + \hat{w}_{35}(a_i,c_i,z_i)(\sigma^2_i - \hat{m}(a_i,c_i,z_i))\Big\} \tag{24}$$

The slope of the resulting profiles are numerically identical to the reported slope estimates in Table 1 and Table 4.

E.2 Proof of the Equivalence

For those interested in the details, we now provide a derivation of the equivalence, described in the previous section, between the debiased estimate of the slope and the slope of a debiased estimate of the age profile.

We will write $\vec{a}$ to indicate we mean the vector of values we get from setting age equal to $35, 36, ..., 50$. So $\hat{\sigma}^2_{\vec{a}}$ is the whole debiased estimated age profile following the form of $\hat{\sigma}^2_{35}$ in Equa-

tion (24). Using this notation, the slope through this estimated age profile would be:

$$\begin{aligned}\text{slope}(\hat{\sigma}^2_{\vec{a}}) &= \text{slope}\left(\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}(\vec{a}, c_i, z_i) + \hat{w}_{\vec{a}}(a_i, c_i, z_i)(\sigma_i^2 - \hat{m}(a_i, c_i, z_i))\right\}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\text{slope}(\hat{m}, c_i, z_i) + \text{slope}(\hat{w}_{\vec{a}}(a_i, c_i, z_i))(\sigma_i^2 - \hat{m}(a_i, c_i, z_i))\right\}. \end{aligned} \tag{25}$$

We are relying on the fact that the slope is calculated with a linear combination, and therefore can move inside the sum. Compare Equation (25) to the debiased point estimate for the slope from Equation (18). The only difference is that while Equation (18) has a single weighting model for the slope $\hat{w}$, in Equation (25), we have the combination of individual weighting models, $\text{slope}(\hat{w}_{\vec{a}}(a_i, c_i, z_i))$.

These turn out to be equivalent in finite sample for the linear and kernel weighting models we consider in Appendix D.3. We provide the derivation for the linear case below (the argument for the kernel setting is virtually identical).

We first drive the solution to the linear weighting loss for a fixed point on the age profile (again, we use age 35 as an example). Let X_{35} denote the covariates where all ages are set to equal 35, and write $\bar{X}_{35} \in \mathbb{R}^d$ for the sample average of the columns of X_{35}. Then the weighting problem is:

$$\min_{\beta \in \mathbb{R}^d}\left\{\frac{1}{n}\beta^\top X^\top X\beta - 2\bar{X}_{35}\beta + \delta\beta^\top\beta\right\}. \tag{26}$$

Taking the derivative with respect to β, the solution is:

$$\beta^*_{35} = n(X^\top X + \delta I)^{-1}\bar{X}_{35},$$

with corresponding weights $\hat{w}_{35} = X\beta^*_{35}$. Then following the same argument from Appendix D.3.1, for

$$s := \frac{(\vec{a} - \bar{a})}{(\vec{a} - \bar{a})^\top(\vec{a} - \bar{a})},$$

the slope through the weights are:

$$\begin{aligned}\text{slope}(X\beta^*_{\vec{a}}) &= s^\top(X\beta^*_{\vec{a}}) \\ &= ns^\top(X(X^\top X + \delta I)^{-1}\bar{X}_{\vec{a}}) \\ &= nX(X^\top X + \delta I)^{-1}(s^\top\bar{X}_{\vec{a}}) \\ &= n(X^\top X + \delta I)^{-1}\bar{X}_{\text{cf}},\end{aligned}$$

which is exactly the solution to the weights for the slope derived in Appendix D.3.1.

F GMM Details

In this section, we describe in detail how we estimate the persistence parameter ρ using GMM.

We use our estimated age profiles as inputs to GMM estimation for the parameters of the income process:

$$y_{i,a} = \alpha_i + z_{i,a} + \epsilon_{i,a} \tag{27}$$

$$z_{i,a} = \rho z_{i,a-1} + \eta_{i,a}. \tag{28}$$

Recall that we assume for simplicity that the processes are centered around zero — $E[y_{i,a}] = E[z_{i,0}] = E[\alpha_i] = E[\epsilon_{i,a}] = E[\eta_{i,a}] = 0$. We denote the variances of the random variables as $\sigma^2_\alpha, \sigma^2_\epsilon$, and σ^2_η. We write the initial permanent variance as $\sigma^2_{z0} \coloneqq \text{Var}[z_{i,0}]$.

The relationship between the age profile and the model parameters is as follows:

$$\begin{aligned}
\sigma^2_a &= E[y^2_{i,a}] \\
&= \sigma^2_\alpha + E[z^2_{i,a}] + \sigma^2_\epsilon \\
&= \sigma^2_\alpha + (\rho^2)^a \sigma^2_{z0} + \sum_{k=1}^{a} (\rho^2)^{a-k} \sigma^2_\eta + \sigma^2_\epsilon.
\end{aligned} \tag{29}$$

Note that we begin indexing at $a = 0$, which corresponds to age 35 in our baseline working age specification.

The full set of parameters $\rho, \sigma^2_\alpha, \sigma^2_\epsilon, \sigma^2_\eta$, and σ^2_{z0} are not identified from the age profile alone — we would need to use the covariance moments. However, ρ is pinned down by the curvature of the age profile.

We estimate ρ jointly with $\sigma^2_\eta, \sigma^2_{z0}$, and $\sigma^2_\alpha + \sigma^2_\epsilon$ using the scipy.optimize library in python. We minimize the (equally weighted) sum of squares of the errors to the estimated data moments, i.e. the age profile, $\hat{\sigma}^2_a$. For computational convenience, we impose the bounds: $\sigma^2_\eta, \sigma^2_{z0}, \sigma^2_\epsilon + \sigma^2_\alpha \in [0, 2]$ and $\rho \in [0.01, 2]$. These constraints never bind. We repeat the solve 200 times using uniform random initializations of the parameters within the bounds. The estimates for ρ are highly stable for every dataset.

In the case where we estimate the age profile $\hat{\sigma}^2_a$ using our debiased machine learning method (with predictive model $\hat{m}$ and weighting model $\hat{w}$), then we can also get a valid confidence interval

for the GMM estimate of ρ. We do this using the standard computation (see for example, Cameron and Trivedi (2005)):

$$\hat{V}_{\text{gmm}} = (D^\top D)^{-1} D^\top \Sigma D (D^\top D)^{-1},$$

where $D \in \mathbb{R}^{n_{\text{age}} \times 4}$ is the Jacobian of the GMM errors with respect to the parameters evaluated at the GMM solution, and $\Sigma \in \mathbb{R}^{n_{\text{age}} \times n_{\text{age}}}$ is the covariance of the errors. We numerically approximate D with a centered second-order finite difference calculation.

For estimating Σ, we make sure to propagate the uncertainty involved in estimating the age profile using the debiased machine learning estimating equation. Each individual element of the age profile is estimated using Equation (24). As described in Appendix E, we estimate the individual weights using the same model we selected via blocked cross-validation for the slope weights. Let $\sigma^2_{\text{gmm}} \in \mathbb{R}^{n_{\text{age}}}$ denote the age profile implied by plugging in the GMM-estimated parameters into Equation (29). Then the GMM error vector calculated at a single observation i are:

$$\text{err}_i := \hat{m}(\vec{a}, c_i, z_i) + \hat{w}_{\vec{a}}(a_i, c_i, z_i)(\sigma_i^2 - \hat{m}(a_i, c_i, z_i)) - \sigma^2_{\text{gmm}},$$

with $\text{err}_i \in \mathbb{R}^{n_{\text{age}}}$, and the error covariance matrix is:

$$\frac{1}{n} \sum_{i=1}^{n} \text{err}_i \text{err}_i^\top.$$

Let the diagonal entry of $\hat{V}_{\text{gmm}}$ that corresponds to the persistence parameter be denoted $\hat{V}_\rho$. Then the standard errors for our estimated persistence are $\sqrt{\hat{V}_\rho / n}$. In practice, we do the variance and standard error calculations using the cross-fitting procedure in Appendix D.5 to avoid underestimating the uncertainty.

Finally, note that we can obtain clustered standard errors by correcting for clustering in the construction of Σ. Let $I_{\ell,t}$ be the index set of observations in year t and fold ℓ. We compute the cluster sums:

$$g_t = \sum_{\ell} \sum_{i \in I_{\ell,t}} \text{err}_i.$$

Let G be the matrix in $\mathbb{R}^{n_{\text{year}} \times n_{\text{age}}}$ whose rows are the g_t for each year. Let A be a diagonal matrix with a row for each year, where the diagonal element is the number of folds in which that year appears.

Then the clustered error covariance is:

$$\Sigma = \left(\frac{n_{\text{year}}}{n_{\text{year}} - 1}\right) \frac{1}{n} G^\top A^{-1} G.$$

The matrix A corrects for double counting of observations across folds (exactly analogous to how we averaged $g_{t,\ell}^2$ across folds in Appendix D.6). We then compute $\hat{V}_{\text{gmm}}$ as above.

G Simulation Study

We perform a simulation study to validate our debiased proxy machine learning methodology. In this study we illustrate (1) the role of the proxy variables in identification, and (2) the importance of the debiasing step for obtaining a valid confidence interval. A key concern is that regularized machine learning might systematically yield a small value for the slope, and perhaps that drives our main results. Therefore, we construct a simulation that serves as a kind of placebo test. As in the US data, the age-cohort fixed effect estimate is larger and the age-year fixed effect estimate is smaller. However, in our simulation, the larger value is actually the truth, and we want our debiased proxy machine learning approach to reliably estimate it. This way we can demonstrate that our estimator doesn't just mechanically pick the smaller value.

We use a data-generating process (DGP) that is calibrated to the US GRID data. we use the ages and years from the real data. We choose a fixed value for the age slope of 0.009, chosen to be higher value for the slope. We then consider the residual after accounting for this age slope: $r_{at} := \sigma_{at}^2 - 0.009 \cdot a$. We use kernel ridge regression to regress r_{at} on cohort and the selected proxies for grid, log oil price, lagged log oil price, and lagged log industrial production squared. Call the resulting predictive model $f(c, z_t)$.

For each simulation draw, we redraw the proxies z_t and the outcomes σ_{at}^2, resulting in a new dataset with $n = 682$ observations. We redraw the proxies by adding random iid noise to the original proxies. The standard deviation of the noise is set to be 1/4th the standard deviation of the corresponding original proxy variable. For non-negative proxies we use an exponential distribution for the noise, for all others we use a normal distribution. We redraw the outcomes as:

$$\sigma_{at}^2 = 0.009 \cdot a + f(c, z_t) + \nu_{at},$$

where ν_{at} is mean zero iid normal noise with standard deviation 0.006 — calibrated to match the in-sample R^2 in the real data.

We run our proxy machine learning methodology with kernel ridge for both the predictive and weighting models, selecting the best performing model with blocked cross-validation, as we do with the real data. We compute the debiased point estimate and confidence interval as described in the main text. For illustrative purposes, we also compute a naive confidence interval for the machine learning estimate without debiasing.

We repeat this process for 2000 draws. We summarize the results in Table G.1. Notice that the biased ML estimate catastrophically undercovers: in 0 out of 2000 draws does the naive confidence interval cover the true slope. By contrast, our debiased procedure achieves coverage of 96.6% — in this simulation we actually over-cover slightly. We achieve this improved coverage in part due to a full 10x decrease in bias.

Table G.1: Simulation Study Results

Estimator	Bias	Coverage
Debiased Machine Learning	0.0003	96.6%
Biased Machine Learning	0.0034	0.0%

H Additional Results

H.1 Age Profile using the PSID

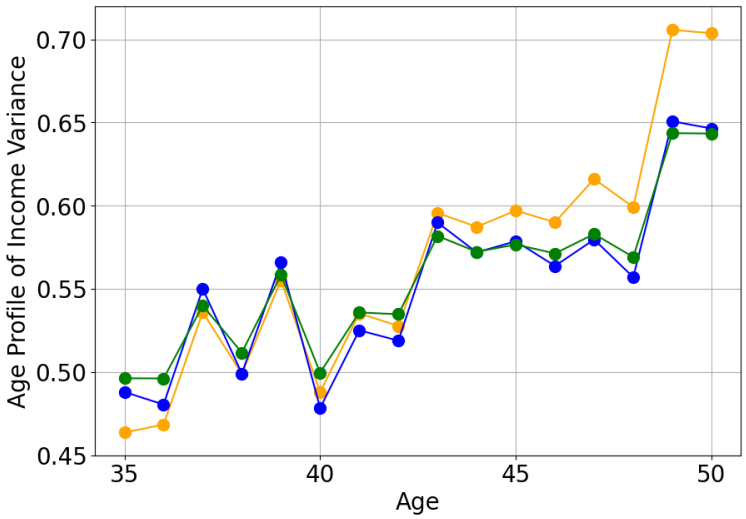


Figure H.1: Age Profile of Cross-Sectional Income Variance in the US

Note: The figure plots the estimated age profile of the cross-sectional variance of income for three models using PSID data.

H.2 Persistence of Income Shocks Across Countries

Table H.1: International Persistence, ρ.

Country	Years	Age-Cohort	Age-Year	Biased ML	Debiased ML
Argentina	1996-2015	$0.990^{\dagger}$	0.955	$0.999^{\dagger}$	0.912 ± 0.107
Brazil	1993-2018	$0.994^{\dagger}$	0.980	$0.999^{\dagger}$	0.850 ± 0.240
Canada	1996-2016	0.952	0.909	0.913	0.910 ± 0.019
Denmark	1996-2016	0.991	0.979	0.984	0.989 ± 0.009
France	1996-2016	0.955	0.938	0.953	0.949 ± 0.024
Germany	2001-2016	0.999	1.012	0.995	0.996 ± 0.001
Italy	1996-2016	0.980	$0.998^{\dagger}$	0.815	0.841 ± 0.225
Mexico	2005-2019	0.740	0.937	0.778	0.853 ± 0.031
Norway	1996-2017	0.994	1.003	0.996	0.995 ± 0.002
Spain	2005-2018	1.001	0.973	1.000	0.76 ± 11.85
Sweden	1996-2016	$0.997^{\dagger}$	0.974	0.961	0.01 ± 1238
USA	1998-2019	0.934	0.878	0.882	0.854 ± 0.027

Note: The table presents results on the persistence of income shocks—ρ in Section 3—for the twelve countries we have data for in GRID. Items with † indicate that the slope is negative. See the note for Table 1 for more detail.

H.3 Comparing In-Sample Fit for Fixed Effects

Data	Age FE	Age Year FE	Age Cohort FE	Year Cohort FE	All FE
PSID 1978-2019	0.168	0.357	0.496	0.535	0.551
PSID 1998-2019	0.269	0.361	0.563	0.561	0.598

Table H.2: PSID In-Sample FE Fit

Data	Age FE	Age Year FE	Age Cohort FE	Year Cohort FE	All FE
Argentina GRID	0.149	0.941	0.828	0.956	0.983
Brazil GRID	0.350	0.971	0.962	0.973	0.994
Canada GRID	0.749	0.905	0.856	0.950	0.980
Denmark GRID	0.783	0.969	0.954	0.663	0.986
France GRID	0.643	0.825	0.823	0.800	0.942
Germany GRID	0.632	0.906	0.920	0.507	0.984
Italy GRID	0.112	0.927	0.861	0.962	0.972
Mexico GRID	0.921	0.992	0.966	0.883	0.998
Norway GRID	0.745	0.939	0.942	0.484	0.972
Spain GRID	0.228	0.963	0.455	0.953	0.980
Sweden GRID	0.760	0.965	0.974	0.519	0.993
USA GRID	0.875	0.980	0.941	0.912	0.995

Table H.3: Grid In-Sample FE Fit