stats

MDPI

Article

A Variable Selection Method for Small Area Estimation Modeling of the Proficiency of Adult Competency

Weijia Ren *, Jianzhu Li, Andreea Erciulescu, Tom Krenzke and Leyla Mohadjer

Westat, 1600 Research Boulevard, Rockville, MD 20850, USA; janeli@westat.com (J.L.); andreeaerciulescu@westat.com (A.E.); tomkrenzke@westat.com (T.K.); leylamohadjer@westat.com (L.M.)
* Correspondence: weijiaren@westat.com

Abstract: In statistical modeling, it is crucial to have consistent variables that are the most relevant to the outcome variable(s) of interest in the model. With the increasing richness of data from multiple sources, the size of the pool of potential variables is escalating. Some variables, however, could provide redundant information, add noise to the estimation, or waste the degrees of freedom in the model. Therefore, variable selection is needed as a parsimonious process that aims to identify a minimal set of covariates for maximum predictive power. This study illustrated the variable selection methods considered and used in the small area estimation (SAE) modeling of measures related to the proficiency of adult competency that were constructed using survey data collected in the first cycle of the PIAAC. The developed variable selection process consisted of two phases: phase 1 identified a small set of variables that were consistently highly correlated with the outcomes through methods such as correlation matrix and multivariate LASSO analysis; phase 2 utilized a k-fold cross-validation process to select a final set of variables to be used in the final SAE models.

Keywords: adult competency; cross-validation; multiple data sources; multivariate LASSO; small area estimation

Citation: Ren, W.; Li, J.; Erciulescu, A.; Krenzke, T.; Mohadjer, L. A Variable Selection Method for Small Area Estimation Modeling of the Proficiency of Adult Competency. *Stats* **2022**, *5*, 689–713. https://doi.org/10.3390/stats5030041

Academic Editor: Wei Zhu

Received: 30 June 2022
Accepted: 21 July 2022
Published: 27 July 2022

1. Introduction

Direct estimates based on survey data alone may not be suitable as reliable statistics for small areas (areas defined by geographies or socio-economic groups for which the realized survey sample sizes are too small for deriving reliable estimates, i.e., counties, census tracts) to help with formulating policies or programs specific for each small area. In contrast, indirect estimation methods present benefits for small area estimation (SAE), especially when models and richness of auxiliary information are investigated [1]. Model-based SAE techniques have been widely adopted by federal statistical agencies, including the Census Bureau's Small Area Income and Poverty Estimates Program (SAIPE), the Census Bureau's Small Area Health Insurance Estimates Program (SAHIE), and the Substance Abuse and Mental Health Services Administration's National Survey on Drug Use and Health Program (SAMHSA's NSDUH).

In general, a preferred SAE model would be one that is complex enough to explain relations in the data and provide accurate prediction, but also simple enough to be understood and explainable/interpretable. One of the key processes to achieving such a satisfying model is to carefully select the set of covariates to use in the model. With the increase in auxiliary information available from multiple data sources nowadays, a much larger pool of potential variables exists compared with decades ago when the pioneering work on model-based small area estimation was developed (i.e., [2,3]). Similar variables from various sources could provide redundant information and cause multi-collinearity issues if used in the same model. Too many variables could also result in computational burden, especially when working with a large dataset and complex models. In addition, unnecessary covariates would add noise to the estimation of interest and waste the degrees of

freedom, which could lead to overfitting. On the other hand, including too few variables in the model could lead to ignorance of important relationships, a decrease in the model goodness of fit, and a decrease in the accuracy of model predictions.

Practically, traditional variable selection methods that are commonly applied in linear and generalized linear models include: (1) significance criteria, e.g., likelihood ratio test (or Wald test) and stepwise (forward or backward) variable selection algorithms; (2) information criteria, e.g., the Akaike information criterion (AIC) and Bayesian information criterion (BIC); (3) regularization criteria, e.g., the least absolute selection and shrinkage operator (LASSO) [4]; (4) association criteria, e.g., decision trees and random forests [5]; (5) cross-validation criteria [6]; and (6) expert knowledge criteria. These methods are directly applicable to the SAE models, where variable selection is considered one of the major problems due to unobservable random effects with limited or no information on their distribution [7]. Variable selection methods that are usually applied in the field of SAE include information criteria [7–10] and regularization and regression trees [11]. There is, however, no single universally applicable variable selection method that fits all SAE models, especially due to the wide range of complexity in SAE models. There is also a lack of practical guidance on how to conduct variable selection in the SAE model development process. In this manuscript, we describe the variable selection process adopted for the National Center for Education Statistics' (NCES's) Program for the International Assessment of Adult Competencies (PIAAC) of the United States. The goal was to identify and select the best set of covariates to be used in the SAE models developed for estimating adult competency outcomes. Section 2 provides the background of the PIAAC, and Section 3 lays out the variable selection methods. The results are presented in Section 4 and a discussion is given in Section 5.

2. Background

2.1. PIAAC

The PIAAC is a multicycle survey of adult skills and competencies sponsored by the Organization for Economic Cooperation and Development (OECD). The international survey examines a range of basic skills in the information age and assesses these adult skills consistently across participating countries. In the United States, three rounds of data were collected in the first cycle of the survey. The round 1 data were collected in the year 2012. In round 2, a supplemental sample was drawn to enhance the round 1 sample [12]. The combined PIAAC 2012/2014 sample was nationally representative of the U.S. adult population that was 16–74 years old. The round 3 data were collected in the year 2017 with two core objectives: (1) to produce a nationally representative sample of the U.S. adult population that was 16–74 years old and (2) arrive at a large enough sample size that, when combined with the 2012/2014 sample (called the 2012/2014/2017 sample), could produce small area estimates for the United States' counties. In each year, a four-stage stratified area probability sample was selected. In stage 1, primary sampling units (PSUs) were selected that consisted of counties or groups of contiguous counties. In stage 2, secondary sampling units (SSUs) were selected that consisted of decennial census blocks or block groups. In stage 3, dwelling units (DUs) were selected. In stage 4, a sampling algorithm was implemented to select one or more sample persons among those identified to be eligible.

The PIAAC is the sixth of a series of adult skills surveys sponsored by the NCES that have been implemented in the United States. In 2009, the NCES published SAE model-dependent estimates for states and counties using the National Adult Literacy Survey (NALS) in 1992 and the National Assessment of Adult Literacy (NAAL) survey in 2003. The proficiency assessment instruments and scales used in the NAAL and NALS are different from those used in the PIAAC, and thus, the small area estimates for counties and states from the NAAL and NALS are not comparable with the corresponding estimates from the PIAAC.

For the 2012/2014/2017 sample, there was a total of 185 unique counties with one or more completed cases from the three rounds of surveys, with a total sample of 12,330 respondents. Among them, there were four counties with less than five completed cases and 43 counties with more than 100 completed cases (see Table 1). The variance of the direct estimates can be large, especially for counties with small sample sizes, thus survey regression estimation (SRE) was applied as a model-assisted approach to bring county population estimates in line with external county totals and improve the stability of the survey estimates. In addition, the variances of the estimates were also predicted through modeling approaches inspired by the generalized variance function (GVF) methods of Wolter [13]. Details can be found in Krenzke et al. [14]. The SAE models were developed using the SREs constructed for these 185 counties. Model-based predictions were made for all 3142 U.S. counties.

Table 1. Number of completed cases per county: 2012/2014/2017.

Number of Completed Cases	Number of Counties
Less than 5	4
5 to 10	14
11 to 20	10
21 to 50	58
51 to 100	56
101 or more	43
Total	185

2.2. Proficiency Measures in the PIAAC

The PIAAC assessed three domains of cognitive skills: (1) literacy, (2) numeracy, and (3) problem-solving in a technology-rich environment. The SAE analysis focused on the first two domains (literacy and numeracy). Within each domain, county- and state-level direct survey estimates of adult proficiency were produced for the proportion at or below level 1, the proportion at level 2, the proportion at level 3 and above, and the average, resulting in eight outcome measures for each state and county (see Table 2). Adjustments were applied to the survey direct estimates to improve stability based on a survey regression estimation (SRE) method [15] and a variance smoothing method through generalized variance functions [13]. As a consequence of the SRE, one of the 185 counties was found to have a negative estimate for the literacy proportion at or below level 1, and thus, was excluded from the SAE modeling.

Table 2. Proficiency domains and measures.

Proficiency Domain	Proficiency Measure
Literacy	Average score
	Proportion at or below level 1
	Proportion at level 2
	Proportion at or above level 3
Numeracy	Average score
	Proportion at or below level 1
	Proportion at level 2
	Proportion at or above level 3

2.3. PIAAC SAE Models

With a careful literature review and discussion with a group of international experts formed for this project, the progression of the previous research and simulation studies led to the development of an area-level bivariate hierarchical Bayes linear three-fold model for

proportions and an area-level univariate hierarchical Bayes linear three-fold model for averages. Specifically, in the proportion model, two proportions (at or below level 1 and at or above level 3) were modeled jointly, and the third proportion (at level 2) was derived by subtracting the proportions of the other two levels from one; meanwhile, in the average model, only one outcome (average scores) was used. One motivation for modeling two proportions jointly instead of separately is the fact that they are correlated and a joint SAE model would borrow strength from that relationship. The SAE models accounted for random effects at three nested levels: county, state, and census divisions. The benefits of the three-fold modeling are that (1) benchmarking the estimates may not be necessary, as estimates are controlled through random effects (a consensus among the experts); (2) estimates for states without samples will not be fully synthetic (i.e., based only on the fixed-effects part of the model) because all census divisions have PIAAC samples; and (3) the precision of the estimates would be further improved by borrowing strength across counties nested within states, as well as states nested within census divisions.

The PIAAC SAE models employed the traditional SAE structure, including a sampling model and a linking model, using matrix form notation to account for multiple domains, as follows:

$$\begin{aligned} \boldsymbol{Y}_{ijk} &\sim \boldsymbol{N}(\boldsymbol{\theta}_{ijk}, \boldsymbol{\Sigma}_{ijk}) \\ \boldsymbol{\theta}_{ijk} &= \boldsymbol{X}'_{ijk}\boldsymbol{\beta} + \boldsymbol{c}_{ijk} + \boldsymbol{v}_{ij} + \boldsymbol{d}_i \end{aligned} \quad (1)$$

where i is an index for the division, j is an index for the state, and k is an index for the county. In the proportions model, $\boldsymbol{Y}_{ijk}$ is a normally distributed bivariate vector of survey regression estimates for proportions at or below level 1 and at or above level 3, with a mean $\boldsymbol{\theta}_{ijk}$ and an estimated variance–covariance matrix $\boldsymbol{\Sigma}_{ijk}$. In the average model, $\boldsymbol{Y}_{ijk}$ is the average score at the county level, which is normally distributed with mean $\boldsymbol{\theta}_{ijk}$ and an estimated variance $\boldsymbol{\Sigma}_{ijk}$. The covariates are denoted by $\boldsymbol{X}_{ijk}$; the regression coefficients are denoted by $\boldsymbol{\beta}$; and $\boldsymbol{c}_{ijk}$, $\boldsymbol{v}_{ij}$, and $\boldsymbol{d}_i$ are the county-level, state-level, and division-level random effects, respectively. Each of the random effects is assumed to follow a normal distribution with a mean of zero and unknown variance.

In total, four SAE models were fit: a literacy proportion model, a literacy average model, a numeracy proportion model, and a numeracy average model. The same set of variables (same model matrix $\boldsymbol{X}_{ijk}$) was selected for the four models because these outcomes were highly correlated, and having the same set of covariates would ease the explanation. See Krenzke et al. [14] for details on model specification, fit, and validation.

3. Methods

In order to conduct the variable selection process, we had to first identify the covariates that were measured consistently across all counties and that were highly correlated with adult proficiency. The variable selection process would narrow down the variables to a reasonably smaller set so that the final model could be developed based on this reduced set of variables. Section 3.1 provides information on the potential effective variables that were measured consistently across all counties and states. The county- and state-level sources from which the potential variables were selected are given in Section 3.2. Section 3.3 describes the variable selection process, which included two phases (as shown in Figure 1): phase 1 reduced the state and county level variables identified in Section 3.1 to a smaller set, and phase 2, based on the results of phase 1, used a k-fold cross-validation process [16] to arrive at the final set of variables for the SAE models.

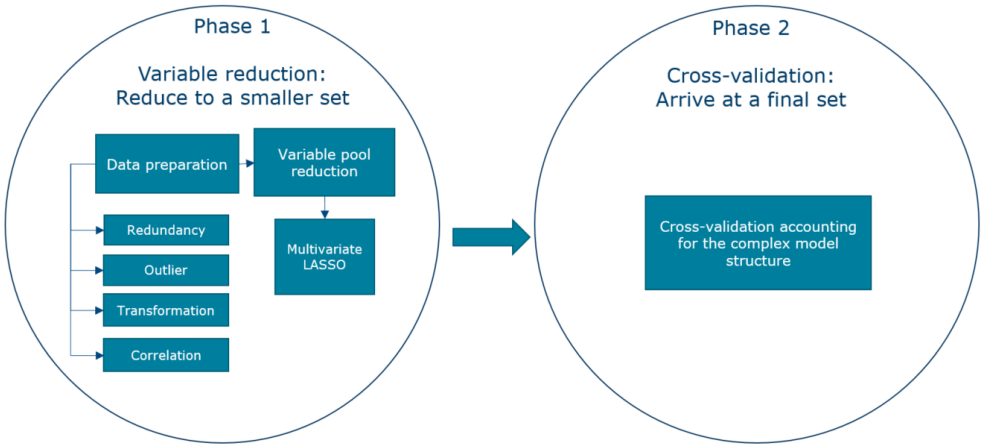

Figure 1. Variable selection process diagram.

The same set of variables was selected for ease of explanation and used in all four SAE models for literacy and numeracy proportions and averages. For example, if a variable was not selected for the literacy proportion model, we still may have wanted to include it in the final set, as it might have been selected for the literacy averages model.

3.1. Identifying County and State Variables

Reliable data sources and variables that are potentially related to adult proficiency levels were initially identified. As a result, more than 70 county-level variables across five major variable types were obtained as potential covariates from eight data sources (see details in Section 3.2). The major county-level variable types included variables related to demographic characteristics (i.e., race/ethnicity, age, gender, marital status), socioeconomic status (i.e., poverty, income, employment status, occupation), education (i.e., education, English-speaking ability), location (i.e., urbanicity, census division), immigration status (i.e., length of stay for foreign-born people, migration), and other (i.e., journey to work, housing unit tenure/phone service, plumbing facilities, health, tax). In addition, the PSU selection probability was also initially included as a potential county-level variable to account for the informative sampling design. However, the models were finally specified for the county-level SREs with associated smoothed variance estimates, and hence, the sample design was accounted for implicitly. The PSU selection probability was not identified as a significant covariate in the variable selection process.

In addition to county-level variables, a set of state-level variables was identified to provide additional information related to adult competency, including 24 potential state-level covariates across different variable types from several major data sources (see the details in Section 3.2). The major state-level variable types included socioeconomic status (i.e., average annual pay, homeownership rate), education (i.e., school enrollment rate, graduation rate, test pass rate, reading/math composite scores), and other area characteristics (i.e., birth rate, fertility rate, infant mortality rate, crime rate, physician availability, federal aid, energy consumption). A listing of all county- and state-level variables considered for modeling is given in the Appendix A. The listing is sorted by major variable type and provides details about the source, year, and level (county level or state level) of each variable.

These variable types were chosen because they were found to be related to the adult proficiency skills in previous studies [17–21] and were available for all the counties in our sample. To ensure that the values of the variables were most relevant to the PIAAC study, we obtained variables collected within the time frame of the PIAAC study. If the variable value was based on a single year of data, we used values from 2015 whenever possible (2015 is the middle time point of the PIAAC study), and if not, the most recent data were used. If the variable was from multiple years, we used the years closest to the PIAAC

study years (i.e., 2013–2017). Sometimes the same variables were selected from multiple data sources. For example, there were two county-level median household income variables selected: one from the American Community Survey (ACS) dataset and the other from the U.S. Census Bureau's SAIPE dataset. Different datasets usually have different sample designs and could incur various sampling and nonsampling errors [22]. Therefore, we gathered variables from multiple readily available sources and attempted to find the most precise and most suitable variables to model adult competency.

3.2. Initial Set of Selected County and State Variable Sources

The selected data sources had reliable data that was publicly available for all counties (or all states) and usually published the updated data regularly (i.e., annually). The following subsections provide brief descriptions of the data sources and the variables chosen from each source. We begin with sources for county-level variables. More details about the variables are given in the Appendix A (see Tables A1 and A2).

3.2.1. Initial Set of Selected Sources for County-Level Variables

Census Bureau's ACS—The Census Bureau's ACS is an ongoing survey that provides up-to-date estimates for a wide range of topics, including socioeconomic, demographic, and housing characteristics of the U.S. population. The 5-year estimates (2013–2017) represent data collected over 5 years for all geographies down to the block-group level (over 578,000 geographic areas). The PIAAC variable pool from ACS includes the number of families in poverty, median household income, population sizes with different education levels, population sizes with English-speaking ability, the population in rural/urban areas, race/ethnicity, the length of stay for foreign-born people, age categories, gender, employment status, occupation, census division, housing unit tenure, phone service, plumbing facilities, marital status, and migration status.

Census Bureau's SAIPE Program—The Census Bureau, with support from other Federal agencies, created the SAIPE program to provide current small area estimates of selected income and poverty statistics. The PIAAC variable pool from SAIPE includes the proportion of families in poverty and median household income.

Bureau of Economic Analysis (BEA)—The BEA prepares estimates of personal income for local areas (counties, metropolitan areas, and the BEA economic areas). The PIAAC variable pool from the BEA includes personal income.

U.S. Department of Agriculture (USDA)—The USDA Economic Research Service provides codes that classify each county according to metro and non-metro classifications. The 2013 Rural-Urban Continuum Codes form a classification scheme that distinguishes metropolitan counties by the population size of their metro area and non-metropolitan counties by the degree of urbanization and adjacency to a metro area. The official Office of Management and Budget (OMB) metro and non-metro categories have been subdivided into three metro and six non-metro categories. Each U.S. county is assigned one of the nine codes. The PIAAC variable pool from USDA includes proportions of metro/non-metro counties.

Bureau of Labor Statistics (BLS)—The Local Area Unemployment Statistics (LAUS) program produces monthly and annual employment, unemployment, and labor force data for census regions and divisions, states, counties, metropolitan areas, and some cities, by place of residence. The PIAAC variable pool from BLS includes the unemployment rate.

Centers for Disease Control and Prevention's Division of Diabetes Translation (DDT)—The CDC collects and provides updated statistics about diabetes in the United States through the U.S. diabetes surveillance system. The PIAAC variable pool from DDT includes the proportions of diagnosed diabetes and obesity.

Centers for Medicare & Medicaid Services (CMS)—The CMS developed a geographic variation public use file with information about the utilization and quality of healthcare services for the Medicare fee-for-service population. The PIAAC variable pool from CMS includes the proportion of the population eligible for Medicaid.

The Statistics of Income Data (SOI)—The SOI bases its county income data on the addresses reported on the individual income tax returns filed with the Internal Revenue Service. The PIAAC variable pool from SOI includes the number of tax returns, returns with unemployment compensation, and returns with taxable Social Security benefits, as well as adjusted gross personal income, personal unemployment compensation amount, and personal taxable Social Security benefit amount.

3.2.2. Initial Set of Selected Sources for State-Level Variables

In addition to county-level variables, a set of state-level variables was also selected to provide additional information not covered by county-level variables. Several variable sources were considered at the state level, as described below.

Bureau of Labor Statistics (BLS)—Besides the BLS LAUS program mentioned above, state-level data were considered from the Current Employment Statistics Program, which surveys more than 160,000 businesses and government agencies each month. The Employment and Wages annual averages were also included in the selection process. The PIAAC variable pool from the BLS includes the average personal annual income.

Adult Education Data (OCTAE)—The Office of Career, Technical, and Adult Education (OCTAE) collects data on adult education program enrollments from each state. Data for 2014–2015 from the National Reporting System (NRS) for Adult Education and Literacy was considered for the small area models. The PIAAC variable pools from the OCTAE are adult basic/secondary education enrollment and English as a second language enrollment.

The Integrated Postsecondary Education Data System (IPEDS)—This NCES program collects data through a system of surveys from primary providers of postsecondary education. The PIAAC variable pool from the IPEDS includes the graduation rate, instructor salary, average financial aid, and annual college cost.

National Assessment of Educational Progress (NAEP)—The NAEP survey is the largest nationally representative and continuing assessment of what our nation's students know and can do in various subject areas. Assessments are conducted periodically in mathematics, reading, science, writing, the arts, civics, economics, geography, U.S. history, and technology and engineering literacy based on representative samples of students in grades 4, 8, and 12 for the main assessments. The PIAAC variable pool from NAEP includes the average 4th- and 8th-grade reading/mathematics composite scale scores, while grade 12 data are not available at the state level.

Other Census Bureau Programs—Besides the ACS, other state demographic data from the Census Bureau were collected from Population Estimates and from data on housing vacancies and home ownership from the Housing Vacancy Survey.

Other Sources—State-level data from other sources were obtained, including the National Highway Safety Traffic Administration's Traffic Safety Facts, National Center for Health Statistics' Vital Statistics of the United States, the American Medical Association's Physician Characteristics and Distribution in the United States, the Federal Bureau of Investigation's Crime in the United States, the Energy Information Administration's State Energy Data Report, the GED Testing Service's Annual Statistical Report on the GED Test, and the Centers for Disease Control and Prevention's National Vital Statistics Reports.

3.3. Variable Selection Process

A key step in model development involves selecting a smaller set of variables from a large pool of potential variables. As mentioned above, for the PIAAC SAE, more than 70 variables in the county-level and more than 20 variables in the state-level variable pool were identified as potential variables. The proposed two-phase process would facilitate researchers to achieve a smaller but reasonable set of variables from a large variable pool.

In the first phase of the selection process, all the county- and state-level variables were considered as fixed effects and the number of variables was reduced. The variable reduction phase implements: (1) data preparation, including a redundancy check, outlier

detection, transformation application, and correlation matrix calculation, and (2) variable pool reduction, specifically, multivariate LASSO was used to select variables for the proportion models and univariate LASSO was used for the average models. This phase resulted in several potential reduced sets of variables. In the second phase of the selection process, the reduced sets of variables from phase 1 were evaluated using a cross-validation process, adding the random effects by using the SAE models (i.e., area-level hierarchical Bayes models) to arrive at the final list of variables. This final list of variables was used when modeling all four SAE models (i.e., literacy/numeracy proportion/average models). Details are provided below.

3.3.1. Phase 1—Variable Reduction

This section describes the variable reduction process in phase 1, which had two major steps: (1) data preparation and (2) variable pool reduction.

Step 1: In the variable selection process, appropriate data preparation is needed before any variable selection algorithm kicks in. In the data preparation step, four data check processes were proposed to ensure the data were well prepared. First, the variables needed to be carefully evaluated for redundancy. In this step, if two variables were found to be redundant, one would be dropped based on level of availability or multicollinearity issues. After examining the redundancy, we identified outliers and influential cases by checking the distributions of the variables, as well as the outcomes. Outliers and influential cases could have a great impact on the variable selection, especially when the sample size is small. Plots (i.e., scatterplot, box plot, or histogram) and interquartile ranges could be used to detect outliers. In addition to outlier detection, the distribution of each continuous variable could be checked in terms of skewness and kurtosis. Skewness assesses the extent to which a variable's distribution is symmetrical, and kurtosis is a measure of whether the distribution is too peaked. Distributions with skewness close to 0 and kurtosis close to 3 are normally distributed. Histogram plots can also be used to visualize the distributions. Lastly, we evaluated whether a transformation was needed. Common transformation methods include standardization, reciprocal, logarithm, square root, squaring or taking the nth power, and categorization/dichotomization. Finally, after removing outliers and doing the transformation, a correlation matrix involving the variables themselves, as well as the outcomes, was created to identify possible multicollinearity between the variables. Variables with high correlations (i.e., 0.7 or 0.8, depending on the data) with another variable were identified as a "highly correlated" pair, and one variable from each pair would be eliminated from the variable pool based on its correlation with the outcome variable.

Step 2: Once the data were well prepared in step 1, a suitable variable pool reduction method was applied to reduce the number of variables further. Before reducing the variables, we had to first identify a target number of variables to include in the final models. This is usually captured by the events-per-variable (EPV) ratio. This is the ratio between the number of observations on the outcome variable and the number of variables included in the model. The EPV ratio quantifies the balance between the amount of information provided by the data and the number of unknown parameters that could be estimated. As a rule of thumb, the EPV ratio could range from 5 to 50, depending on the variables considered and the models being developed [23–25].

After the target number of variables was determined, we investigated several variable reduction methods and decided to use the LASSO method for the reduction. The LASSO method was selected because of its applicability to multivariate model structure, which was the structure of the SAE proportion models. LASSO [4] is a method that applies shrinkage factors to regression coefficients, and thus, can more efficiently perform stable variable selection. The LASSO model included the fixed effects but not the random effects. The goal of the method is to minimize the sum of squares plus a penalty term, which is a multiple of the sum of the absolute values of the regression coefficients, i.e., $\lambda \sum_{j=1}^{p} |\beta_j|$. The tuning parameter λ controls the strengths of the penalty. The procedure can select a few variables that are related to the dependent variable from a large number of possible variables. LASSO-based

methods use "penalized regression" models that impose constraints on the estimated coefficients that tend to shrink the magnitude of the regression coefficients, often eliminating the variables by shrinking their coefficients to zero. Therefore, nonzero coefficients are estimated for true variables, whereas the coefficients for irrelevant variables are zeroed out.

The LASSO estimation was carried out in R using the *glmnet* package [26] in our analysis. LASSO estimation is highly dependent on the scale of the covariates; therefore, LASSO performs an internal standardization to unit variance first, before the coefficient shrinkage takes place. The final variable reduction process was based on applying the LASSO model with standardized covariates and a LASSO penalty.

3.3.2. Phase 2—Cross-Validation

It is possible that the relationship between a variable and the outcome from a simple additive model might change in the complex model. Therefore, it would be risky to directly use the selected variables from a selection algorithm in the final models.

These sets of candidate variables thus needed to be evaluated in phase 2, where a cross-validation process took place. In this phase, complex models with all the features of the final model (including both fixed and random effects) were applied. The final selected set of variables was the one with decent predictive power and presumably interpretable.

In our study, the SAE models were used to make predictions for the non-sampled counties (the counties that had no PIAAC sample or had too few sampled cases to be usable). The cross-validation analysis evaluated the prediction power of the model as compared with other models using alternative sets of variables selected from the LASSO models (from phase 1) through k-fold cross-validation.

The k-fold cross-validation was implemented in the following steps to select the best set of variables for the bivariate model of literacy proportions:

- We sorted the 184 sampled counties from the largest to the smallest by sample size and divided them into groups of 10 counties, with the last group having only 4 counties. There were 19 groups in total.
- For each group of 10 counties, the counties were randomly assigned to 10 subsets, with each subset containing 1 county from the group. For the group with four counties, the counties were randomly assigned to four subsets. At the end of this step, each subset contained 18 or 19 counties with varying sample sizes.
- Excluding the counties in the first subset, the counties in the remaining nine subsets were used to fit the bivariate small area estimation model for each given set of variables and made predictions for the group of counties that were deleted.
- The previous step was repeated by excluding subsets 2 through 10, one at a time. At the end of this process the predicted proportions at or below level 1, at level 2, and at or above level 3 were calculated for all the counties.

We compared the predicted proportions against the direct estimates for all 184 counties, as well as only for the counties with large sample sizes (sample size greater than 100). The sums of the squared differences were calculated. The smaller the sum of squared differences was, the better the set of variables predict the proportions for the counties that were excluded from modeling.

4. Results

Specific results from the variable selection method described above for the PIAAC 2012/2014/2017 data are presented below. For each phase, we describe the selection process in more detail and motivate the selection of the variables. The estimation of the literacy proportion at or below level 1 was used as an example, but a similar process and results were obtained for the numeracy models and the average models.

4.1. Phase 1—Variable Reduction

Step 1: In the data preparation step, a redundancy check revealed that since our variable pool contained both county- and state-level variables, we had the same variables (i.e., race/ethnicity, poverty rate) available at both levels from the same data source (i.e., ACS). Therefore, we dropped the state-level variables with the assumption that the county-level variables contained more information. In addition, similar variables (i.e., median household income) were found across different data sources (i.e., ACS vs. SAIPE); therefore, we kept both in this step. Their correlation was then explored, and if deemed high, one of the two variables was dropped in this step.

In the outlier and influential case detection step, the three income variables (i.e., median household income) in our variable pool were log-transformed to support an assumption of a linear relationship with the outcomes.

In the correlation check process, a correlation matrix involving the variables themselves, as well as the eight outcomes, was created to identify the multicollinearity between variables. Specifically, the Pearson correlation matrix was computed between each pair of the potential county-level variables (observed for the 3142 counties), and for each pair of the potential state-level variables (observed for the 50 states and the District of Columbia). In addition, the Pearson correlations between the variables and each of the eight outcomes (proportion at or below level 1, proportion at level 2, proportion at or above level 3, and average proficiency score for both literacy and numeracy) were constructed for all 184 counties with valid SREs.

It should be noted that all the variables in our analysis were continuous; therefore, the Pearson correlation was applicable. For studies where categorical variables were involved, other association tests (i.e., Cramer's V) could be conducted to test for the relationships between variables. Variables with high correlations with the outcomes turned out to be the education-related variables (i.e., $|\rho| = 0.7$ for the proportion of the population with lower than high school education vs. the proportion at or below level 1 literacy), poverty-related variables (i.e., $|\rho| = 0.6$ for the proportion of the population lower than the poverty threshold versus the proportion at or below level 1 literacy), employment-related variables (i.e., $|\rho| = 0.6$ for the proportion of the population not in the labor force vs. the proportion at or below level 1 literacy), and health-related variables (i.e., $|\rho| = 0.5$ for the proportion of the population with no health insurance vs. the proportion at or below level 1 literacy). Variables with high pair-wise correlations (i.e., $|\rho| > 0.7$) with other variables were treated as with "high multicollinearity", and one variable from each pair was dropped from the variable pool. Specifically, the variable with a lower correlation with the outcomes was dropped in each pair. In the cases where two highly correlated variables were correlated by definition and found to have a key impact on the outcomes (i.e., proportion of the population less than high school, proportion of the population more than high school), both were kept for the following variable reduction process.

Step 2: In the variable pool reduction step, we chose the EPV ratio of 30 as a target EPV ratio. With 184 sampled cases, we aimed to select six variables for the final model.

In our analysis, four LASSO models were created. For proportions models, multivariate LASSO was used with the option "family = "mgaussian"", whereby a multi-task learning method was applied when there were several correlated responses. Using a "group LASSO", the multivariate LASSO selected the same set of variables for all the outcomes. For the average models, we used univariate LASSO to conduct variable selection because the outcome was univariate. For each of the four LASSO models, the random effects were dropped from the model specification and the LASSO penalty parameter λ (the parameter that controls the overall strength of the penalty) was adjusted using various values close in magnitude to the λ that minimized the mean cross-validated error (0.02 and 0.03 for the proportion model, and 2 and 3 for the average model). As a result, we constructed two sets of variables (with non-zero coefficients) for each of the four models. Each set contained 10 or fewer variables with some variation among the sets. The lambda values and the estimated coefficients of the covariates were obtained using the cv.glmnet

function. Two lambda values were used to obtain two sets of selected variables, with one being more parsimonious than the other.

In Table 3, we report the list of the selected phase 1 variables with the source, year, description, and label. In Table 4, we report the selected variables with the marker "✓" identifying the selected variables for each of the LASSO models (with two λ options). It should be noted that for the numeracy proportion model, both lambda options (0.02 and 0.03) resulted in the same set of selected variables.

Table 3. List of phase 1 selected variables, including their source, year, description, and label.

Source	Year(s)	Description	Label
American Community Survey	2013–2017	Percentage of population aged 25 and over with less than high school education (no high school diploma)	Education—LH
		Percentage of population aged 25 and over with more than high school education (including some college, no degree)	Education—MH
		Percentage of population below 100 percent of the poverty line	Poverty
		Percentage of Black or African American population	Black
		Percentage of Hispanic population	Hispanic
		Percentage of civilian non-institutionalized population who has no health insurance coverage	No health insurance
		Percentage of population aged 16 and over with service occupations	Service occupations
		Percentage of foreign-born people who entered the United States after the year 2010 among the population born outside the United States	Enter U.S. 2010
		Percentage of the population born outside of the United States	Foreign born
		Percentage of population 16 and over who did not work at home who spent more than 60 min traveling to work	Journey to work
Bureau of Labor Statistics	2015	Unemployment rate	Unemployment rate
Division of Diabetes Translation	2013	Percentage of diabetes diagnosed	Diabetes rate
National Vital Statistics Reports	2015	Birth rate per 1000 women	Birth rate
The Integrated Postsecondary Education Data System	2014–2015	Average amount of grant and scholarship aid received	Grant/ scholarship received

Table 4. Covariates selected in phase 1 by outcome and LASSO lambda option.

Variable	Literacy				Numeracy			
	Proportion Model		Average Model		Proportion Model		Average Model	
	$\lambda = 0.02$	$\lambda = 0.03$	$\lambda = 2$	$\lambda = 3$	$\lambda = 0.02$	$\lambda = 0.03$	$\lambda = 2$	$\lambda = 3$
Education—LH	✓	✓	✓	✓	✓	✓	✓	✓
Education—MH	✓	✓	✓	✓	✓	✓	✓	✓
Poverty	✓	✓	✓	✓	✓	✓	✓	✓
Black	✓		✓		✓	✓	✓	✓
Hispanic							✓	

No health insurance	✓	✓	✓	✓	✓	✓	✓
Service occupations		✓	✓			✓	✓
Enter U.S. 2010	✓	✓					
Foreign born	✓						
Journey to work		✓					
Unemployment rate		✓				✓	✓
Diabetes rate				✓	✓		
Birth rate	✓						
Grant/scholarship received	✓	✓				✓	

Appendix A Tables A3 and A4 provide the listings of the county- and state-level selected variables with the correlation estimates and LASSO standardized regression estimates, which were sorted in descending order by the correlations for each model.

4.2. Phase 2—Cross-Validation

For the literacy proportions model, five sets of variables (all county level) were used to fit the models and to compare the predicted proportions against the direct estimates. The results are summarized in Table 5.

Table 5. Variables used in the cross-validation for literacy proportions and results of the summed squared differences between predicted proportions and direct estimates: 2012/2014/2017.

Variable	**Scenarios**				
	1	**2**	**3**	**4**	**5**
Education—LH	✓	✓	✓	✓	✓
Education—MH	✓	✓	✓	✓	✓
Poverty	✓	✓	✓	✓	✓
Black		✓	✓	✓	✓
Enter U.S. 2010		✓	✓		
No health insurance		✓		✓	✓
Birth rate		✓			
Grant/scholarship received		✓			
Foreign born		✓			
Hispanic			✓	✓	✓
Service occupations					✓
Sums of squared differences between the predicted proportions and direct estimates over 44 counties with a sample size of at least 100					
P1	0.109	0.078	0.081	0.076	0.076
P2	0.136	0.137	0.144	0.141	0.143
P3	0.212	0.155	0.186	0.170	0.183

For the cross-validation analysis, scenarios 1 and 2 were chosen from the LASSO models with $\lambda = 0.03$ and $\lambda = 0.02$, respectively. Scenario 3 used the five covariates adopted by the hierarchical Bayes model in the NAAL study to predict the proportion of adults lacking basic prose literacy skills and added the percent of Hispanics as a covariate, which was highly correlated with the proportion at or below level 1. Compared with scenario 3, scenario 4 added another covariate, namely, the proportion of people with no health insurance coverage, which was shown to be significant in the LASSO models for predicting proportions and averages for both literacy and numeracy. Scenario 5 added an extra covariate, namely, the proportion in the service occupation, to the set of variables used in scenario 4 because this variable was shown to be a significant covariate in the LASSO models for predicting averages for both literacy and numeracy.

The results in Table 5 show that scenarios 2, 4, and 5 had similar performances and their sums of the squared differences between the model predictions and direct estimates were smaller for all three proportions than those from scenarios 1 and 3. Combining these results with the other cross-validation results for the literacy average and numeracy proportions and average, a decision was made to use the seven county-level variables from the 2013–2017 ACS data in all four models fitted for proportions and averages for literacy and numeracy, as shown in Table 6. It should be recognized that the ACS data was subject to sampling error, and the ACS estimates for counties with small sample sizes might have been associated with greater uncertainty. Table 7 shows the correlation coefficients between these variables. The seven variables were highly correlated with the proportions and averages. For example, the adjusted R-square was 0.58 for the linear regression (without random effects) of literacy proportions at or below level 1 for the seven variables.

Table 6. List of variables for the final small area models.

Variables	Label
Percentage of population aged 25 and over with less than high school education	Education—LH
Percentage of population aged 25 and over with more than high school education	Education—MH
Percentage of population below 100 percent of the poverty line	Poverty
Percentage of Black or African American population	Black
Percentage of Hispanic population	Hispanic
Percentage of civilian non-institutionalized population who had no health insurance coverage	No health insurance
Percentage of population aged 16 and over with service occupations	Service occupations

Table 7. Correlation coefficients among variables for the final small area model: 2012/2014/2017.

Variable	Education—MH	Poverty	Black	Hispanic	No Health Insurance	Service Occupation
Education—LH	−0.76	0.64	0.34	0.42	0.58	0.21
Education—MH		−0.53	−0.20	−0.04	−0.38	−0.13
Poverty			0.47	0.08	0.47	0.37
Black				−0.11	0.19	0.15
Hispanic					0.40	0.15
No health insurance						0.19

The final SAE models were evaluated internally and externally to ensure the goodness of fit. We briefly present selected diagnostics plots below. First, shrinkage plots were created to show how the indirect estimates changed from the SREs by sample size. The plots in Figure 2 show that the model predictions were pulled toward the average. The direction of the arrow corresponds to the direction of the shrinkage and the length of the arrow corresponds to the shrinkage amount. Longer arrows correspond to counties with smaller sample sizes, which was expected since the model predictions could deviate from the survey estimates in small areas more than in large areas.

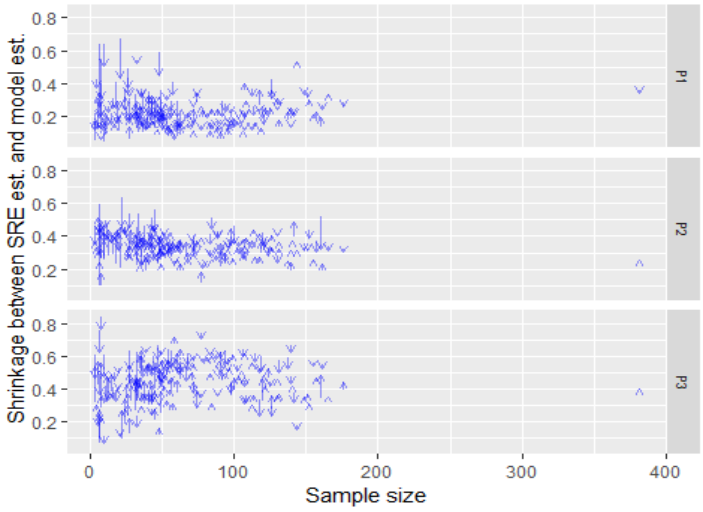

Figure 2. Literacy proportion—shrinkage plots of point estimates by sample size: 2012/2014/2017. P1: proportion at or below level 1; P2: proportion at level 2; P3: proportion at or above level 3.

Furthermore, scatterplots of both the estimates and variance estimates were constructed to show the relationship between the survey regression estimates and the model predictions. As seen in Figure 3, for the estimates of the proportion of literacy at or below level 1, the majority of the points were around the 45-degree line, indicating that the model predictions were close to the SREs, and counties with larger sample sizes (i.e., larger bubbles) had closer estimates than those with small sample sizes.

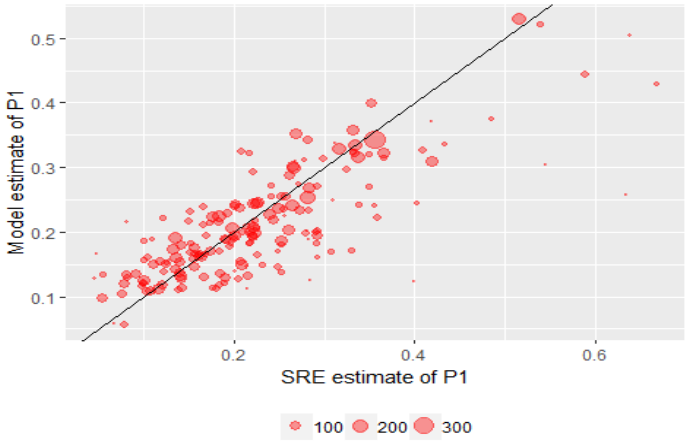

Figure 3. Literacy proportion—comparison between survey regression estimates and indirect estimates: 2012/2014/2017. P1: proportion at or below level 1.

Lastly, the plot in Figure 4 shows the smoothed standard errors of the SREs and the posterior standard deviations from the small area model. The plot shows that the model produced smaller posterior standard deviations than the smoothed standard errors, especially for areas of very small sample sizes.

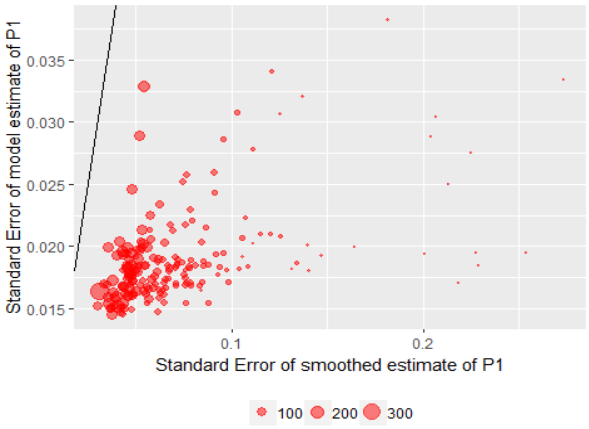

Figure 4. Literacy proportion—comparison between model standard errors and smoothed standard errors: 2012/2014/2017. P1: proportion at or below level 1.

The small area estimates for all the states and counties in the United States are available to the public on the Skills Map website at https://nces.ed.gov/surveys/piaac/skillsmap/ (accessed on 29 June 2022).

5. Discussion

Variable selection has become an issue that almost all modeling processes would encounter, especially with the existence of abundant auxiliary information. Exclusion of the variables that should be included in a model or inclusion of variables that should be excluded could directly affect the reliability and stability of the model. This study provided a practical example for researchers to apply variable selection methods in complex models, such as SAE models. It is not recommended to include all the variables in a variable selection algorithm and solely rely on the model to decide the selected variables without any exploration of these variables first. The choice of variable reduction method should be based on the nature of the final model.

In our approach, two phases were conducted. In the first phase, all the state- and county-level variables were considered as fixed effects and the number of variables was reduced as follows: (1) a correlation matrix was created for all the variables to identify highly correlated variables; then, (2) one variable in each of the highly correlated pairs was dropped to avoid multicollinearity. Subsequently, the LASSO method was used to select several sets of variables for each of the four outcome models for literacy and for numeracy. The multivariate nature of our final models resulted in the choice of a multivariate LASSO. In general, our recommendation is to select several sets of candidate variables by the end of phase 1. To assess the stability of the selected set of variables in phase 1, we conducted a sensitivity analysis and evaluated the phase 1 variable selection process by reapplying the LASSO selection approach to leave out each of the subsets described in the cross-validation stage (phase 2) one at a time and found out that the selected set of variables was robust to the selection approach. Nevertheless, we would recommend that future users consider conducting the variable selection approach described in this paper under the cross-validation scheme to increase the stability and validity of the selection results [27]. In the second phase, these various selected sets of variables selected in phase 1 were evaluated and a final list of variables was determined using a cross-validation process that took into account the random effect estimations.

For the PIAAC SAE application of the variable selection process, we identified variables related to education, poverty, race/ethnicity, health insurance coverage, and service occupation as associated with adult proficiency. Education, poverty, and race/ethnicity were shown to have an association with literacy/numeracy proficiency in previous studies. All

the variables were county-level variables from the ACS 5-year dataset, indicating that the county-level variables might have stronger predictive power than the state-level variables.

There were also several challenges encountered during the application. First, the creation of the auxiliary variable pool was an intensive process. Most of the data were extracted from publicly available datasets, and appropriate variables were derived from the datasets. For our application, all the potential variables should be available for all the U.S. counties (or states), and when there were multiple years of data available (i.e., from the ACS), decisions should be made regarding which data to use. Second, the direct estimates and covariates were subject to sampling error; therefore, the correlation coefficients constructed in the first phase of the selection process were biased and attenuated. As pointed out in Lahiri and Suntornchost [28], the true population correlations were higher, and the correlation estimates could be improved if the sampling error was taken into account. Third, we had four complex final SAE models to fit. Because it was decided to use the same set of variables for all four final models due to the high correlations between the eight outcomes (i.e., literacy/numeracy proficiency levels/scores) when forming the candidate sets of variables, we considered variables that were found to be important for both the proportion and average models. Moreover, in the models, we included counties with sample sizes as small as 4. As a result, the direct estimates from some counties were not stable, which led us to calculate the sums of square differences in phase 2 based on the 44 counties with sample sizes of 100 or more.

It should be noted that the variable selection process varies study by study in practice, depending on the datasets and final models to be fit. We recommend carefully exploring the variables and deciding upon the variable selection method to be used. The final selected variables should consider both the data-driven results from variable selection algorithms and variables shown to be important from theories and previous studies.

Author Contributions: Conceptualization, W.R., J.L., A.E., T.K. and L.M.; Data curation, W.R. and J.L.; Formal analysis, W.R. and J.L.; Funding acquisition, T.K. and L.M.; Investigation, W.R., J.L., A.E., T.K. and L.M.; Methodology, W.R., J.L., A.E., T.K. and L.M.; Project administration, T.K. and L.M.; Resources, T.K. and L.M.; Software, W.R., J.L. and A.E.; Supervision, T.K. and L.M.; Validation, W.R., J.L. and A.E.; Visualization, W.R.; Writing—original draft, W.R.; Writing—review & editing, J.L., A.E., T.K. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education: EDIES12C0072.

Data Availability Statement: The survey data are confidential to the National Center for Education Statistics and cannot be shared. Auxiliary data and model estimates are available in the Skills Map website, at https://nces.ed.gov/surveys/piaac/skillsmap/ (accessed on 29 June 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. List of county-level variables by source and year.

County Characteristics	Source	Year
Poverty		
Percentage of population below 150 percent of poverty line	ACS	2013–2017
Percentage of population receiving SNAP/food stamps	ACS	2013–2017
Percentage of population below 100 percent of poverty line	ACS	2013–2017
Percentage of population in poverty (all ages)	SAIPE	2015
Income		
Median household income—ACS	ACS	2013–2017
Median household income—SAIPE	SAIPE	2015
Per capita personal income	BEA	2015
Education		

Percentage of population aged 25+: with education less than high school (no high school diploma)	ACS	2013–2017
Percentage of population aged 25+: with high school diploma, no college	ACS	2013–2017
Percentage of population aged 25+: with education more than high school (including some college, no degree)	ACS	2013–2017
English-speaking ability for people who speak another language		
Percentage of population aged 5+: speaking other languages and speaking English not at all or not well	ACS	2013–2017
Percentage of population aged 5+: speaking other languages	ACS	2013–2017
Urban/rural		
Metro or non-metro counties	ACS	2013–2017
Counties in metro areas of 1 million population or more	USDA	2013
Counties in metro areas of less than 1 million population	USDA	2013
Non-metro counties	USDA	2013
Race/ethnicity		
Percentage of Hispanics	ACS	2013–2017
Percentage of Whites	ACS	2013–2017
Percentage of Blacks	ACS	2013–2017
Percentage of Asians	ACS	2013–2017
Percentage of American Indians and Alaska Natives	ACS	2013–2017
Percentage of Native Hawaiians and Pacific Islanders	ACS	2013–2017
Percentage of other races	ACS	2013–2017
Foreign-born status		
Percentage of foreign-born people who entered United States after year 2010	ACS	2013–2017
Percentage of foreign-born people who entered United States between years 1990 and 2009	ACS	2013–2017
Percentage of foreign-born people who entered United States after year 1990	ACS	2013–2017
Percentage of foreign-born people who entered United States before year 1990	ACS	2013–2017
Percentage of population born outside of United States	ACS	2013–2017
Age		
Percentage of population 16–54 years old	ACS	2013–2017
Percentage of population 55–64 years old	ACS	2013–2017
Percentage of population 65+ years old	ACS	2013–2017
Gender		
Percentage of male population	ACS	2013–2017
Employment status		
Unemployment rate	BLS	2015
Percentage of population aged 20–64: in armed forces	ACS	2013–2017
Percentage of population aged 20–64: in labor force and employed	ACS	2013–2017
Percentage of population aged 20–64: in labor force and unemployed	ACS	2013–2017
Percentage of population aged 20–64: not in labor force	ACS	2013–2017
Occupation		
Percentage of population aged 16+: management/professional occupations	ACS	2013–2017
Percentage of population aged 16+: service occupation	ACS	2013–2017
Percentage of population aged 16+: sales/office occupation	ACS	2013–2017
Percentage of population aged 16+: natural resources/construction/maintenance occupation	ACS	2013–2017
Percentage of population aged 16+: military	ACS	2013–2017
Percentage of population aged 16+: production/transportation/moving occupation	ACS	2013–2017
Census division		
New England	ACS	2013–2017

Middle Atlantic	ACS	2013–2017
East North Central	ACS	2013–2017
West North Central	ACS	2013–2017
South Atlantic	ACS	2013–2017
East South Central	ACS	2013–2017
West South Central	ACS	2013–2017
Mountain	ACS	2013–2017
Pacific	ACS	2013–2017
Journey to work		
Percentage of population aged 16+ and did not work at home: less than 30 min to work	ACS	2013–2017
Percentage of population aged 16+ and did not work at home: 30–44 min to work	ACS	2013–2017
Percentage of population aged 16+ and did not work at home: 45–59 min to work	ACS	2013–2017
Percentage of population aged 16+ and did not work at home: 60+ minutes to work	ACS	2013–2017
Housing unit tenure and phone service		
Percentage of owner-occupied housing units	ACS	2013–2017
Percentage of renter-occupied housing units	ACS	2013–2017
Percentage of owner-occupied housing units with phone service available	ACS	2013–2017
Percentage of renter-occupied housing units with phone service available	ACS	2013–2017
Percentage of occupied housing units	ACS	2013–2017
Plumbing facilities		
Percentage of housing units with plumbing facilities	ACS	2013–2017
Marital status		
Percentage of population 15+: never married	ACS	2013–2017
Percentage of population 15+: married	ACS	2013–2017
Percentage of population 15+: widowed	ACS	2013–2017
Percentage of population 15+: divorced	ACS	2013–2017
Migration		
Percentage of population 1+: in different house in the past year	ACS	2013–2017
Percentage of population 1+: in different county in the past year	ACS	2013–2017
Percentage of population 1+: in different state in the past year	ACS	2013–2017
Percentage of population 1+: moved from abroad in the past year	ACS	2013–2017
Health		
Percentage of civilian non-institutionalized population with one type of health insurance coverage	ACS	2013–2017
Percentage of civilian non-institutionalized population with two or more types of health insurance coverage	ACS	2013–2017
Percentage of civilian non-institutionalized population with no health insurance coverage	ACS	2013–2017
Percentage of diagnosed diabetes	DDT	2013
Percentage of obesity	DDT	2013
Percentage of population eligible for Medicaid	CMS	2015
Tax		
Average number of tax returns per person	SOI	2014
Average number of returns with unemployment compensation per person	SOI	2014
Average number of returns with taxable Social Security benefits per person	SOI	2014
Proportion of the amount of unemployment compensation among all tax return amounts	SOI	2014
Proportion of the amount of taxable Social Security benefits among all tax return amounts	SOI	2014

ACS: American Community Survey; SNAP: Supplemental Nutrition Assistance Program; SAIPE: Small Area Income and Poverty Estimates Program; BEA: Bureau of Economic Analysis; USDA: U.S. Department of Agriculture; BLS: Bureau of Labor Statistics; DDT: Centers for Disease Control and Prevention's Division of Diabetes Translation; CMS: Centers for Medicare & Medicaid Services; SOI: The Statistics of Income Data.

Table A2. List of state-level variables by source and year.

State Characteristics	Source	Year
Socioeconomic status		
Average annual pay	BLS	2015
Homeownership rate	Housing Vacancies and Home Ownership (CPS/HVS)	2015
Education		
Adult basic education enrollment rate	OCTAE	2015
Adult secondary education enrollment rate	OCTAE	2015
English as a second language enrollment rate	OCTAE	2015
Graduation rate of postsecondary institutes	IPEDS	2014–2015
Average weighted monthly salary for full-time instructional staff	IPEDS	2014–2015
Average amount of grant and scholarship aid received	IPEDS	2014–2015
Annual college cost (tuition and fees)	IPEDS	2014–2015
GED test completion rate	GED Testing Service (GEDTS)	2013
Average 4th-grade reading composite scale scores	NAEP	2015
Average 4th-grade math composite scale scores	NAEP	2015
Average 8th-grade reading composite scale scores	NAEP	2015
Average 8th-grade math composite scale scores	NAEP	2015
Other area characteristics		
Infant mortality rate per 1000 live births	NCHS, Vital Statistics of the United States, annual, and unpublished data	2013
Women 15–50 years old who gave birth in the past 12 months (per 1000 15–50-year-old women)	ACS	2011–2015
Physicians per 100,000 population	AMA, Chicago, IL, Physician Characteristics and Distribution in the United States, 2014	2015
Violent crime rate per 100,000 population	FBI, Crime in the United States, annual	2015
Federal aid to state and local governments per capita	Census Bureau, Federal Aid to States for Fiscal Year 2010	2010
State government general revenue per capita	Census Bureau; State and Local Government Finance Estimates by State, annual, and unpublished data	2014
Energy consumption per person	EIA, State Energy Data Report, 2014	2014
Traffic fatalities per 100 million vehicle miles	NHTSA, Traffic Safety Facts, annual	2015
Birth rate	National Vital Statistics Reports, 2015	2017
Birth rate for teenagers aged 15–19	National Vital Statistics Reports, 2015	2017

BLS: Bureau of Labor Statistics; CPS/ HVS: Housing Vacancies and Homeownership; OCTAE: Office of Career, Technical, and Adult Education; IPEDS: Integrated Postsecondary Education Data System; GED: General Educational Development; NAEP: National Assessment of Educational Progress; NCHS: National Center for Health Statistics; ACS: American Community Survey; AMA: American Medical Association; FBI: Federal Bureau of Investigation; EIA: Energy Information Administration; NHTSA: National Highway Traffic Safety Administration.

Table A3. PIAAC county- and state-level variable correlations with literacy/numeracy proficiency outcomes: 2012/2014/2017.

Variable	Literacy P1	Literacy P2	Literacy P3	Literacy average	Numeracy P1	Numeracy P2	Numeracy P3	Numeracy Average
County level								
Percentage of population aged 25+: with education less than high school	0.72	0.22	−0.70	−0.73	0.74	−0.11	−0.63	−0.73
Percentage of population aged 25+: with high school diploma, no college	0.28	0.59	−0.59	−0.44	0.36	0.41	−0.59	−0.44
Percentage of population aged 25+: with education more than high school	−0.56	−0.52	0.77	0.68	−0.63	−0.22	0.73	0.68
Percentage of population below 100 percent of poverty line	0.65	0.24	−0.65	−0.67	0.74	−0.10	−0.64	−0.71
Percentage of population receiving SNAP/food stamps	0.59	0.31	−0.66	−0.64	0.69	0.01	−0.66	−0.68
Percentage of population below 150 percent of poverty line	0.67	0.28	−0.70	−0.70	0.75	−0.05	−0.68	−0.73
Percentage of population in poverty (all ages)	0.64	0.23	−0.64	−0.64	0.71	−0.09	−0.62	−0.68
ACS median household income—log-transformed	−0.49	−0.42	0.65	0.56	−0.59	−0.13	0.64	0.59
SAIPE median household income	−0.49	−0.42	0.65	0.56	−0.59	−0.13	0.64	0.59
Per capita personal income—log-transformed	−0.17	−0.34	0.35	0.23	−0.20	−0.15	0.28	0.21
Percentage of population aged 5+: speak another language and speak English not at all or not well	0.15	−0.15	−0.02	−0.10	0.11	−0.18	0.01	−0.12
Percentage of population aged 5+: speaking other languages	0.24	−0.37	0.05	−0.15	0.14	−0.33	0.07	−0.13
Percentage of Hispanics	0.33	−0.25	−0.10	−0.27	0.26	−0.26	−0.09	−0.26
Percentage of Blacks	0.37	−0.03	−0.27	−0.32	0.46	−0.24	−0.28	−0.39
Percentage of Asians	−0.04	−0.40	0.28	0.13	−0.15	−0.27	0.31	0.16
Percentage of American Indians and Alaska Natives	0.01	−0.04	0.02	−0.03	0.01	<0.001	−0.01	−0.05
Percentage of Whites	−0.33	0.27	0.08	0.23	−0.34	0.37	0.09	0.28
Percentage of Native Hawaiians and Pacific Islanders	−0.04	−0.13	0.12	0.07	−0.08	−0.05	0.11	0.08
Percentage of other races	0.20	−0.29	0.03	−0.12	0.14	−0.29	0.05	−0.11
Percentage of foreign-born people who entered United States after year 2010	−0.22	−0.27	0.34	0.31	−0.21	−0.18	0.31	0.26
Percentage of foreign-born people who entered United States between years 1990 and 2009	0.16	−0.19	−0.01	−0.10	0.13	−0.23	0.02	−0.13
Percentage of foreign-born people who entered United States after year 1990	<0.001	−0.31	0.19	0.09	−0.01	−0.29	0.20	0.05
Percentage of foreign-born people who entered United States before year 1990	−0.02	−0.02	0.03	0.02	−0.07	0.07	0.02	0.06
Percentage of population born outside of United States	0.12	−0.40	0.16	−0.03	0.02	−0.33	0.18	−0.01
Percentage of population 16–54 years old	0.11	−0.20	0.04	−0.05	0.07	−0.17	0.04	−0.05
Percentage of population 55–64 years old	−0.16	0.32	−0.08	0.07	−0.14	0.31	−0.06	0.09
Percentage of population 65+ years old	−0.07	0.36	−0.17	−0.03	−0.03	0.33	−0.17	−0.02

Percentage of male population	0.19	<0.001	−0.15	−0.18	0.14	−0.12	−0.06	−0.12
Percentage of population aged 20–64: in armed forces	−0.10	−0.04	0.10	0.11	−0.06	0.01	0.05	0.09
Percentage of population aged 20–64: in labor force and employed	−0.52	−0.41	0.66	0.58	−0.60	−0.12	0.64	0.60
Percentage of population aged 20–64: in labor force and unemployed	0.33	0.05	−0.29	−0.33	0.39	−0.07	−0.33	−0.39
Percentage of population aged 20–64: not in labor force	0.62	0.24	−0.63	−0.63	0.67	−0.07	−0.59	−0.64
Percentage of population aged 16+: management/ professional occupations	−0.38	−0.50	0.61	0.51	−0.44	−0.33	0.62	0.52
Percentage of population aged 16+: service occupation	0.34	0.07	−0.31	−0.37	0.39	−0.07	−0.33	−0.39
Percentage of population aged 16+: sales/office occupation	−0.05	0.17	−0.07	−0.04	0.02	0.24	−0.16	−0.09
Percentage of population aged 16+: natural resources/construction/maintenance occupation	0.22	0.36	−0.40	−0.30	0.22	0.23	−0.35	−0.28
Percentage of population aged 16+: military	−0.09	−0.02	0.08	0.10	−0.06	0.03	0.04	0.09
Percentage of population aged 16+: production/transportation/moving occupation	0.29	0.43	−0.50	−0.39	0.32	0.29	−0.48	−0.38
Percentage of population aged 16+ and did not work at home: less than 30 min to work	<0.001	−0.02	0.01	0.01	0.02	0.01	−0.03	−0.02
Percentage of population aged 16+ and did not work at home: 30–44 min to work	−0.02	−0.09	0.08	0.05	−0.01	−0.16	0.11	0.05
Percentage of population aged 16+ and did not work at home: 45–59 min to work	−0.07	0.04	0.03	0.06	−0.09	<0.001	0.08	0.09
Percentage of population aged 16+ and did not work at home: 60+ min to work	0.07	0.11	−0.12	−0.11	0.02	0.14	−0.10	−0.07
Percentage of owner-occupied housing units	−0.20	0.32	−0.05	0.08	−0.20	0.38	−0.04	0.12
Percentage of renter-occupied housing units	0.20	−0.32	0.05	−0.08	0.20	−0.38	0.04	−0.12
Percentage of owner-occupied housing units with phone service available	−0.30	−0.11	0.30	0.29	−0.32	0.04	0.28	0.31
Percentage of renter-occupied housing units with phone service available	−0.21	−0.05	0.20	0.19	−0.21	0.02	0.18	0.19
Percentage of occupied housing unit	−0.10	−0.26	0.24	0.15	−0.15	−0.14	0.23	0.16
Percentage of housing units with plumbing facilities	−0.15	−0.07	0.16	0.15	−0.14	0.02	0.12	0.14
Percentage of population aged 15+: never married	0.24	−0.37	0.05	−0.11	0.23	−0.41	0.03	−0.16
Percentage of population aged 15+: married	−0.35	0.19	0.15	0.26	−0.40	0.31	0.19	0.33
Percentage of population aged 15+: widowed	0.35	0.47	−0.57	−0.45	0.41	0.28	−0.57	−0.46
Percentage of population aged 15+: divorced	0.07	0.34	−0.27	−0.15	0.18	0.23	−0.31	−0.19
Percentage of population aged 1+: in different house in the past year	−0.10	−0.20	0.20	0.16	−0.05	−0.23	0.19	0.13
Percentage of population aged 1+: in different county in the past year	0.10	−0.01	−0.07	−0.10	0.11	−0.11	−0.04	−0.08
Percentage of population aged 1+: in different state in the past year	−0.20	−0.17	0.26	0.27	−0.16	−0.20	0.28	0.25
Percentage of population aged 1+: moved from abroad in the past year	−0.15	−0.52	0.45	0.32	−0.23	−0.44	0.49	0.33

Percentage of civilian non-institutionalized population with one type of health insurance coverage	−0.43	−0.20	0.46	0.43	−0.48	<.001	0.46	0.46
Percentage of civilian non-institutionalized population with two or more types of health insurance coverage	−0.04	0.29	−0.15	−0.02	0.01	0.23	−0.15	−0.01
Percentage of civilian non-institutionalized population with no health insurance coverage	0.52	<0.001	−0.41	−0.48	0.53	−0.17	−0.40	−0.51
Percentage of diagnosed diabetes	0.39	0.45	−0.58	−0.49	0.50	0.19	−0.59	−0.52
Percentage of obesity	0.40	0.38	−0.55	−0.47	0.48	0.17	−0.56	−0.49
Percentage of population eligible for Medicaid	0.54	0.09	−0.48	−0.52	0.55	−0.06	−0.49	−0.54
Average number of tax returns per person	−0.12	−0.40	0.35	0.18	−0.20	−0.22	0.33	0.19
Average number of returns with unemployment compensation per person	−0.04	<0.001	0.03	0.03	−0.06	0.07	0.01	0.04
Average number of returns with taxable Social Security benefits per person	−0.38	0.28	0.12	0.28	−0.36	0.38	0.10	0.30
Proportion of the amount of unemployment compensation among all tax return amounts	0.09	0.10	−0.14	−0.12	0.09	0.09	−0.14	−0.12
Proportion of the amount of taxable Social Security benefits among all tax return amounts	−0.09	0.42	−0.19	−0.02	−0.03	0.39	−0.21	−0.03
Unemployment rate	0.48	0.15	−0.46	−0.48	0.54	−0.09	−0.45	−0.51
Counties in metro areas of 1 million population or more	−0.12	−0.23	0.24	0.17	−0.15	−0.12	0.22	0.16
Counties in metro areas of less than 1 million population	−0.07	−0.01	0.06	0.06	−0.07	0.05	0.04	0.05
Non-metro counties	0.22	0.28	−0.35	−0.25	0.26	0.08	−0.29	−0.24
New England	−0.16	−0.02	0.14	0.15	−0.16	0.02	0.14	0.16
Middle Atlantic	−0.07	−0.01	0.06	0.06	−0.08	0.05	0.04	0.06
East North Central	−0.17	0.08	0.08	0.11	−0.13	0.11	0.06	0.09
West North Central	−0.11	0.03	0.07	0.10	−0.18	0.10	0.11	0.14
South Atlantic	0.07	<0.001	−0.06	−0.05	0.11	−0.03	−0.09	−0.10
East South Central	0.16	0.23	−0.27	−0.19	0.23	0.07	−0.26	−0.18
West South Central	0.27	−0.09	−0.15	−0.23	0.26	−0.16	−0.15	−0.25
Mountain	−0.14	−0.01	0.12	0.15	−0.14	−0.03	0.16	0.17
Pacific	0.06	−0.26	0.12	<0.001	−0.02	−0.16	0.12	0.01
State level								
Adult basic education enrollment rate	0.20	0.31	−0.35	−0.25	0.28	0.14	−0.36	−0.28
Physicians per 100,000 population	−0.18	−0.15	0.23	0.23	−0.20	−0.07	0.23	0.23
Birth rate for teenagers aged 15–19	0.34	0.21	−0.39	−0.36	0.40	0.02	−0.39	−0.38
Average annual pay	−0.11	−0.27	0.25	0.16	−0.15	−0.14	0.23	0.16

Adult secondary education enrollment rate	−0.12	0.13	0.01	0.09	−0.10	0.14	0.01	0.11
Birth rate	0.23	−0.07	−0.13	−0.16	0.18	−0.20	−0.05	−0.13
GED test completion rate	0.13	0.13	−0.18	−0.17	0.18	0.07	−0.22	−0.17
English as a second language enrollment rate	−0.15	−0.33	0.32	0.20	−0.23	−0.18	0.33	0.22
Traffic fatalities per 100 million vehicle miles	0.31	0.25	−0.40	−0.35	0.35	0.09	−0.39	−0.35
Women 15–50 years old who gave birth in the past 12 months	0.10	−0.05	−0.04	−0.06	0.04	−0.09	0.02	−0.03
Average amount of grant and scholarship aid received	−0.26	−0.01	0.20	0.24	−0.26	0.09	0.19	0.24
Graduation rate of postsecondary institutes	−0.01	−0.18	0.12	0.06	−0.08	−0.12	0.15	0.09
Homeownership rate	−0.18	0.20	0.02	0.12	−0.14	0.17	0.02	0.13
Infant mortality rate per 1000 live birth	0.21	0.22	−0.30	−0.24	0.31	0.05	−0.32	−0.29
Average 4th-grade math composite scale scores	−0.23	0.01	0.17	0.22	−0.24	0.06	0.19	0.24
Average 8th-grade math composite scale scores	−0.37	−0.06	0.32	0.35	−0.41	0.07	0.34	0.39
Energy consumption per person	0.23	−0.06	−0.14	−0.18	0.20	−0.13	−0.11	−0.18
State government general revenue per capita	−0.11	−0.25	0.25	0.19	−0.19	−0.08	0.23	0.20
Federal aid to state and local governments per capita	−0.01	−0.07	0.05	0.07	−0.02	−0.06	0.05	0.06
Average 4th-grade reading composite scale scores	−0.22	0.13	0.09	0.18	−0.19	0.12	0.11	0.20
Average 8th-grade reading composite scale scores	−0.41	0.10	0.26	0.35	−0.42	0.19	0.28	0.39
Average weighted monthly salary for full-time instructional staff	−0.34	−0.29	0.33	0.23	−0.25	−0.12	0.32	0.25
Annual college cost (tuition and fees)	−0.25	−0.02	0.21	0.23	−0.24	0.03	0.21	0.24
Violent crime rate per 100,000 population	0.21	−0.15	−0.07	−0.19	0.16	−0.11	−0.09	−0.19

P1: proportion at or below level 1; P2: proportion at level 2; P3: proportion at or above level 3; SNAP: Supplemental Nutrition Assistance Program. Source: U.S. Department of Education, National Center for Education Statistics, U.S. Program for the International Assessment of Adult Competencies (PIAAC), 2012/2014/2017.

Table A4. PIAAC county- and state-level variable LASSO selection results (regression coefficients corresponding to standardized covariates) for models with literacy/numeracy proficiency outcomes: 2012/2014/2017.

Variable	**Literacy**						**Numeracy**					
	λ **= 0.02**		λ **= 0.03**		λ **= 2**	λ **= 3**	λ **= 0.02**		λ **= 0.03**		λ **= 2**	λ **= 3**
	P1	**P3**	**P1**	**P3**	**Avg.**	**Avg.**	**P1**	**P3**	**P1**	**P3**	**Avg.**	**Avg.**
Percentage of population aged 25+: with education less than high school	0.62	−0.51	0.57	−0.53	−108.97	−107.03	0.48	−0.26	0.44	−0.28	−86.67	−88.65
Percentage of population aged 25+: with education more than high school	−0.13	0.45	−0.13	0.38	27.54	22.77	−0.24	0.47	−0.21	0.39	41.17	31.47
Percentage of population below 100 percent poverty line	0.26	−0.27	0.26	−0.28	−31.84	−34.53	0.47	−0.35	0.52	−0.39	−45.39	−59.07
Percentage of Blacks	0.03	0.02	†	†	−1.71	†	0.10	0.05	0.04	0.02	−12.76	−5.30
Percentage of foreign-born people who entered United States after year 2010	−0.01	0.02	†	†	1.64	†	†	†	†	†	†	†
Percentage of civilian non-institutionalized population with no health insurance coverage	0.07	−0.04	†	†	−11.68	−6.44	0.21	−0.14	0.11	−0.07	−38.52	−33.75
Birth rate	<0.001	<0.001	†	†	†	†	†	†	†	†	†	†
Average amount of grant and scholarship aid received	<0.001	<0.001	†	†	<0.001	†	†	†	†	†	<0.001	†
Percentage of population born outside of United States	<0.001	<0.001	†	†	†	†	†	†	†	†	†	†
Unemployment rate	†	†	†	†	−0.03	†	†	†	†	†	−0.33	−0.26
Percentage of population aged 16+: service occupation	†	†	†	†	−16.52	−0.70	†	†	†	†	−17.44	−0.81
Percentage of population aged 16+ and did not work at home: 60+ minutes to work	†	†	†	†	−0.31	†	†	†	†	†	†	†
Percentage of Hispanics	†	†	†	†	†	†	†	†	†	†	−1.95	†

† Not applicable. P1: proportion at or below level 1; P3: proportion at or above level 3; Avg.: average score.

References

1. Rao, J.N.K.; Molina, I. *Small Area Estimation*, 2nd ed.; Wiley Series in Survey Methodology; Wiley**:** Hoboken, NJ, USA, 2015.
2. Fay, R.E.; Herriot, R.A. Estimates of income for small places: An application of James-Stein procedures to census data. *J. Am. Stat. Assoc.* **1979**, *74*, 269–277.
3. Battese, G.E.; Harter, R.M.; Fuller, W.A. An error-components model for prediction of county crop areas using survey and satellite data. *J. Am. Stat. Assoc.* **1988**, *83*, 28–36.
4. Tibshirani, R. The Lasso Method for Variable Selection in the Cox Model. *Stat. Med.* **1997**, *16*, 385–395.
5. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*, 1st ed.; Wadsworth Statistics/Probability; Routledge: New York, NY, USA, 1984.
6. Shao, J. Linear Model Selection by Cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.
7. Pfeffermann, D. New Important Developments in Small Area Estimation. *Stat. Sci.* **2013**, *28*, 40–68.
8. Van den Brakel, J.A.; Buelens, B. Covariate Selection for Small Area Estimation in Repeated Sample Surveys. *Stat. Transit. New Ser. Surv. Methodol. Jt. Issue Small Area Estim.* **2014**, *16*, 523–540.
9. Erciulescu, A.L.; Berg, E.J.; Cecere, W.; Ghosh, M. A bivariate hierarchical Bayesian model for estimating cropland cash rental rates at the county level. *Surv. Methodol.* **2019**, *45*, 199–216.
10. Cai, S.; Rao, J.N.K.; Dumitrescu, L.; Chatrchi, G. Effective Transformation-based Variable Selection under Two-Fold Subarea Models in Small Area Estimation. *Stat. Transit. New Ser.* **2020**, *21*, 68–83.
11. Erciulescu, A.L.; Opsomer, J.D. A model-based approach to predict employee compensation components. In *Proceedings of the Joint Statistical Meetings*; Government Statistics Section, American Statistical Association: Alexandria, VA, USA, 2019. Available online: https://ww2.amstat.org/MembersOnly/proceedings/2019/data/assets/pdf/1199560.pdf (accessed on 22 July 2022).
12. Hogan, J.; Thornton, N.; Diaz-Hoffmann, L.; Mohadjer, L.; Krenzke, T.; Li, J.; Van De Kerckhove, W.; Yamamoto, K.; Khorramdel, L.U.S. *Program for the International Assessment of Adult Competencies (PIAAC) 2012/2014: Main Study and National Supplement Technical Report (NCES 2016-036REV)*; U.S. Department of Education, National Center for Education Statistics: Washington, DC, 2016.
13. Wolter, K.M. Taylor Series Methods and Generalized Variance Functions. In *Introduction to Variance Estimation. Statistics for Social and Behavioral Sciences*; Springer: New York, NY, USA, 2007.
14. Krenzke, T.; Mohadjer, L.; Li, J.; Erciulescu, A.; Fay, R.; Ren, W.; VanDeKerckhove, W.; Li, L.; Rao, J.N.K. *Program for the International Assessment of Adult Competencies (PIAAC): State and County Estimation Methodology Report*; National Center for Education Statistics: Washington, DC, USA, 2020.
15. Särndal, C.E.; Hidiroglou, M. Small Domain Estimation: A Conditional Analysis. *J. Am. Stat. Assoc.* **1989**, *84*, 266–275.
16. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **2011**, *21*, 137–146.
17. Rampey, B.D.; Finnegan, R.; Goodman, M.; Mohadjer, L.; Krenzke, T.; Hogan, J.; Provasnik, S. *Skills of U.S. Unemployed, Young, and Older Adults in Sharper Focus: Results from the Program for the International Assessment of Adult Competencies (PIAAC) 2012/2014: First Look (NCES 2016-039rev)*; U.S. Department of Education, National Center for Education Statistics: Washington, DC, USA, 2016.
18. Goodman, M.; Finnegan, R.; Mohadjer, L.; Krenzke, T.; Hogan, J. *Literacy, Numeracy, and Problem Solving in Technology-Rich Environments among U.S. Adults: Results from the Program for the International Assessment of Adult Competencies 2012: First Look (NCES 2014-008)*; U.S. Department of Education, National Center for Education Statistics: Washington, DC, USA, 2013. Available online: https://nces.ed.gov/pubs2014/2014008.pdf (accessed on 22 July 2022).
19. Kirsch, I.S.; Jungeblut, A.; Jenkins, L.; Kolstad, A. *Adult Literacy in America: A First Look at the Results of the National Adult Literary Survey (NALS)*; U.S. Department of Education, National Center for Education Statistics: Washington, DC, USA, 2002.
20. Greenberg, E.; Macias, R.F.; Rhodes, D.; Chan, T. *English Literacy and Language Minorities in the United States: NCES 2001-464*; U.S. Department of Education, National Center for Education Statistics: Washington, DC, USA. Available online: http://www.nces.ed.gov/pubs2002/2002382.pdf (accessed on 22 July 2022).
21. Coley, R.J. International adult literacy. *ETS Policy Notes* **1996**, *7*, 1–12.
22. Vaish, A.K. Small area estimation with data from multiple sources. Presented at the 61st ISI World Statistics Congress Satellite Meeting on Small Area Estimation, Paris, France, 10–12 July 2017.
23. Harrell, F.E.; Lee, K.L.; Califf, R.M.; Pryor, D.B.; Rosati, R.A. Regression modeling strategies for improved prognostic prediction. *Stat. Med.* **1984**, *3*, 143–152.
24. Harrell, F.E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*; Springer: New York, NY, USA, 2015.
25. Austin, P.C.; Allignol, A.; Jason, P.F. The number of primary events per variable affects estimation of the subdistribution hazard competing risks model. *J. Clin. Epidemiol.* **2017**, *83*, 75–84.
26. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
27. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*; Springer: New York, NY, USA, 2013.
28. Lahiri, P.; Suntornchost, J. Variable selection for linear mixed models with applications in small area estimation. *Indian J. Stat.* **2015**, *77*, 312–320.