Predicting Language Outcomes for 12-month-old Infants From Low-Income Families: A Machine Learning Approach based on Demographics Data at Birth

Hsuan-Wei (Isaac) Chen[1*] and William R. Doyle[2]

[1*]Department of Psychology and Human Development, Vanderbilt University, Nashville, USA, TN.
[2]Department of Leadership, Policy and Organizations, Vanderbilt University, Nashville, USA, TN.

*Corresponding author(s). E-mail(s): hsuan-wei.chen@vanderbilt.edu;

Abstract

Purpose: The abstract serves both as a general introduction to the topic and as a brief, non-technical summary of the main results and their implications. The abstract must not include subheadings (unless expressly permitted in the journal's Instructions to Authors), equations or citations. As a guide the abstract should not exceed 200 words. Most journals do not set a hard limit however authors are advised to check the author instructions for the journal they are submitting to.

Keywords: language, early childhood development, low-income familites, machine learning

Introduction

- Problem statement: The object of this study is to predict ASQ communication language scores based on demographic variables collected at birth in low-income families
- Understanding early language development is crucial because it is strongly linked to their reading proficiency later in school. Socioeconomic status (SES) may be a factor that affects a child's early language abilities because they may not be exposed to an environment filled with literacy interactions or opportunities to listen and use the language constantly. A child who enters school without essential building blocks for learning how to read may be at risk for falling behind their peers.

Methods

Dataset Description

The Baby's First Years (BFY) project is first randomized controlled trial (RCT) in the U.S. designed to evaluate the causal impact of poverty reduction on a child's early development. Since its initiation in 2018, the BFY has recruited 1,000 mothers of infants with incomes below the federal poverty line across four diverse communities: New York City, New Orleans, the greater Omaha metropolitan area, and the Twin Cities. Mothers were recruited from postpartum wards shortly after giving birth and received a monthly cash gift by debit card for the first 76 months of their child's life. Mothers were randomly assigned to one of two groups: (1) an experimental group (n = 400) receiving $333 per month ($3,996 per year) and (2) a control group (n = 600) receiving $20 per month ($240 per year). Importantly, participants did not lose eligibility to public benefits (e.g. Supplemental Nutrition Assistance Program, Head Start, or Medicaid) due to the cash reward (Noble et al (2021)).

The inclusionary criteria was the following: (1) mother's self-reported income was below the federal poverty threshold in the previous calendar year; (2) mother was of legal age for informed consent; (3) infant was admitted to the newborn nursery and not requiring admittance to the intensive care unit; (4) mother was residing in the state of recruitment; (5) mother reported not being "highly likely" to move to a different state or country in the next 12 months; (6) infant was discharged in the custody of the mother; and (7) mother was either English or Spanish speaking (necessary for instruments of some child outcomes) (Noble et al (2021)).

Families in the BFY study were involved in four waves of data collection. First, baseline data was collected in the hospital shortly after birth. Afterwards, in-person home visits were conducted when the child was 12 and 24 months of age. Lastly, a university-based laboratory visit was conducted when the child was 36 months of age. This analysis used self-reported surveys data collected at baseline including mother demographics, mother-father relationship, and public assistance as predictors of language outcome (**Table 1**). The language outcome of interest was the communication subtest of the Ages and Stages Questionnaire (ASQ) collected at 12 months of age. The ASQ is a developmental screening tool designed to assess young children's progress across five key domains: Communication, Gross Motor, Fine Motor, Problem Solving, and Personal-Social. The Communication domain specifically evaluates a child's ability to understand and use of both expressive and receptive language.

Table 1: Description of self-report survey measures and examples

Measure	Survey Question Example
Child Information	Child is female (Yes/No)
Mother Demographics	Mother's has unpaid maternity leave (Yes/No)
Father Demographics	Father's highest level of educaion attained (Multinomial)
Mother-Father Relationship	Biological dad put money towards baby's arrival (Yes/No)
Household Roster	Number of adults in the household including mother (Continuous)
Income/Net Worth	Household combined calculated income (Continuous)
Public Assistance	Household receives child care subsidy (Yes/No)
Maternal Health	Average alcohol drinks per week during pregnancy (Continous)

Dataset Access and Cleaning

- Baby's First Years (BFY) Data Access
- SPSS data file were download via package *Haven*

- Variables in the baseline data file are of two types – **raw** and **generated**. The first type of variables is considered raw because they are direct outputs from self-reported surveys. They are unprocessed.
- The second (**generated**) type of variables in the Baseline_Clean_Data_BFY data file are generated by BFY analysts in preparation for analyses of the data. These variables are re-coded (e.g., yes/no responses are coded yes=1 and no=0). In addition to simple recoding of values, a number of quality checks were conducted to create complicated generated variables, such as income, that required analytic decisions.
- The user guide recommends analysts to use the generated variables

Cleaning

- Raw variables were removed based on variable labels from SPSS file
- Variables were converted to numeric

Elastic net Pre-processing

1. step_other(all_nominal_predictors(), threshold = 0.01) - any categories that constitute less than 1% of the data will be lumped into the category "other"
2. step_dummy(all_nominal_predictors()) - converts categorical variables into dummy variables
3. step_filter_missing(all_predictors(), threshold = 0.1) - removes any predictor variables that have more than 10% missing values
4. step_impute_mean(all_numeric_predictors()) - substitute missing values of numeric variables by the training set mean of those variables
5. step_naomit(all_outcomes()) - Removes cases where the otucome has missing values
6. step_zv(all_predictors()) - Removes predictor variables that have a zero variance, meaning they have the same value for all observations
7. step_corr(all_predictors(), threshold = 0.95) - Identifies and removes predictor variables that have a correlation higher than 0.95 with any other predictor
8. step_normalize(all_predictors()) - Normalizes all predictor variables so they have a mean of 0 and a standard deviation of 1.

Random Forest Pre-processing

1. step_other(all_nominal_predictors(), threshold = 0.01) - any categories that constitute less than 1% of the data will be lumped into the category "other"
2. step_dummy(all_nominal_predictors()) - converts categorical variables into dummy variables
3. step_filter_missing(all_predictors(), threshold = 0.1) - removes any predictor variables that have more than 10% missing values
4. step_impute_mean(all_numeric_predictors()) - substitute missing values of numeric variables by the training set mean of those variables
5. step_naomit(all_outcomes()) - Removes cases where the otucome has missing values
6. step_zv(all_predictors()) - Removes predictor variables that have a zero variance, meaning they have the same value for all observations

Exploratory Analyses

Univariate Analyses

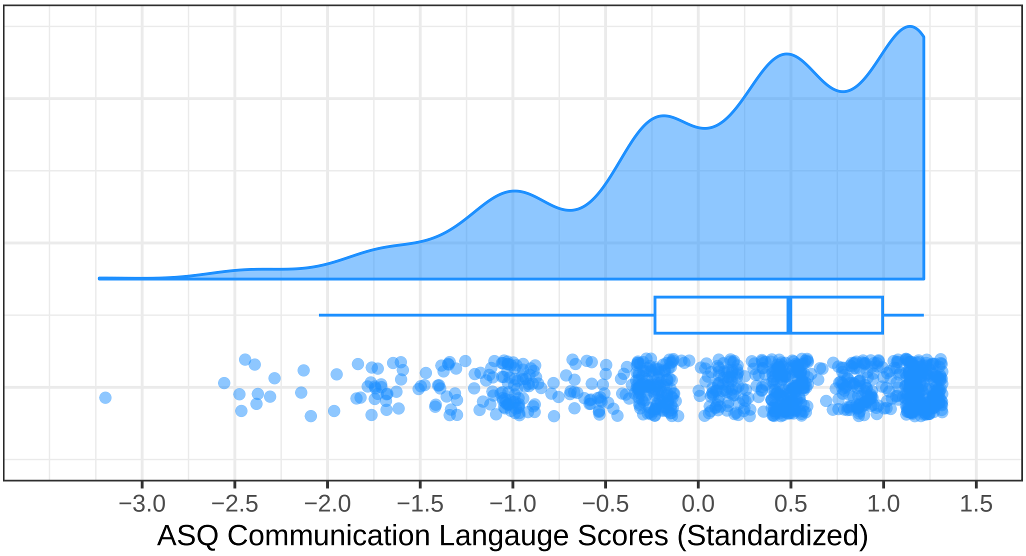

Fig. 1 Distribution of ASQ communication language scores

Bivariate Analyses

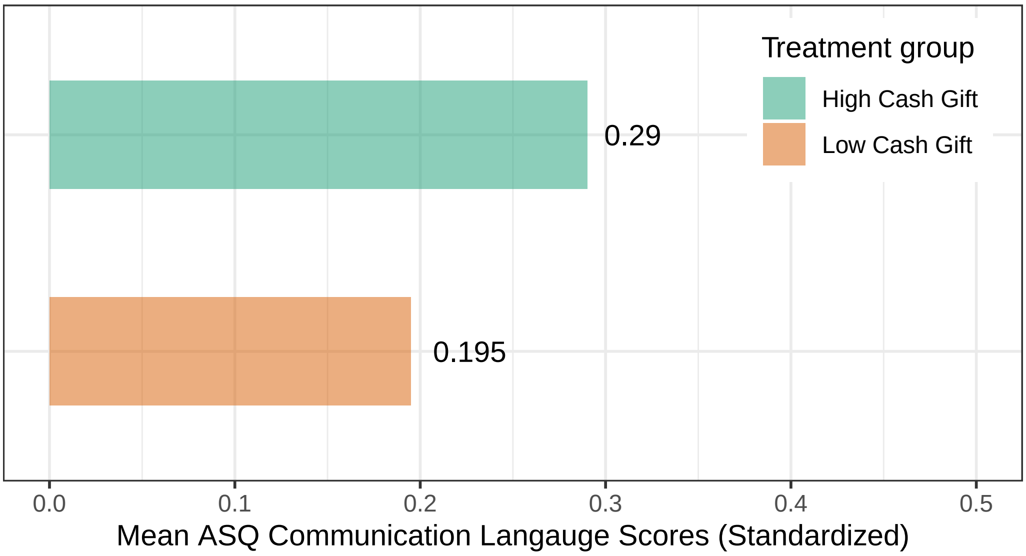

Fig. 2 Mean ASQ communication language scores by Treatment Group

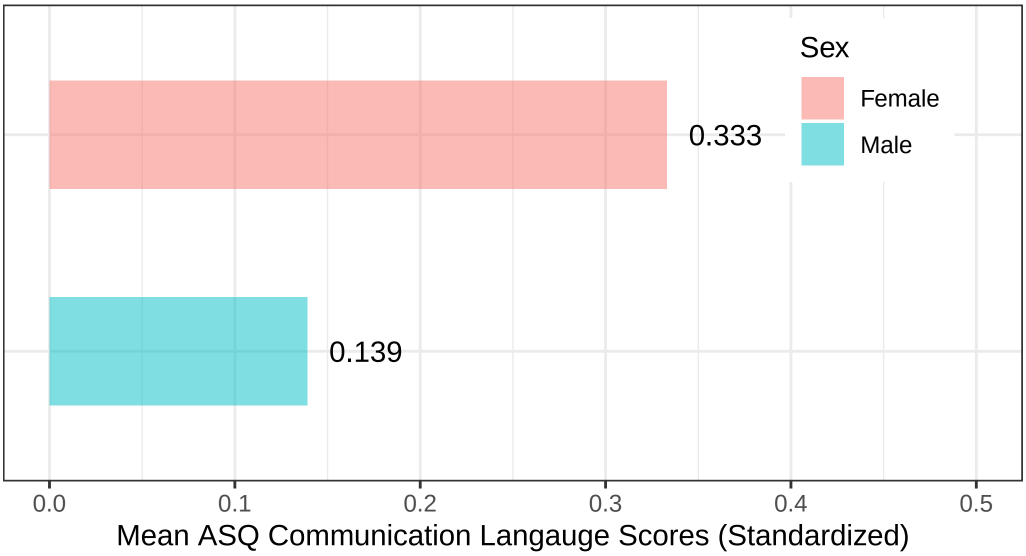

Fig. 3 Mean ASQ communication language scores by Sex

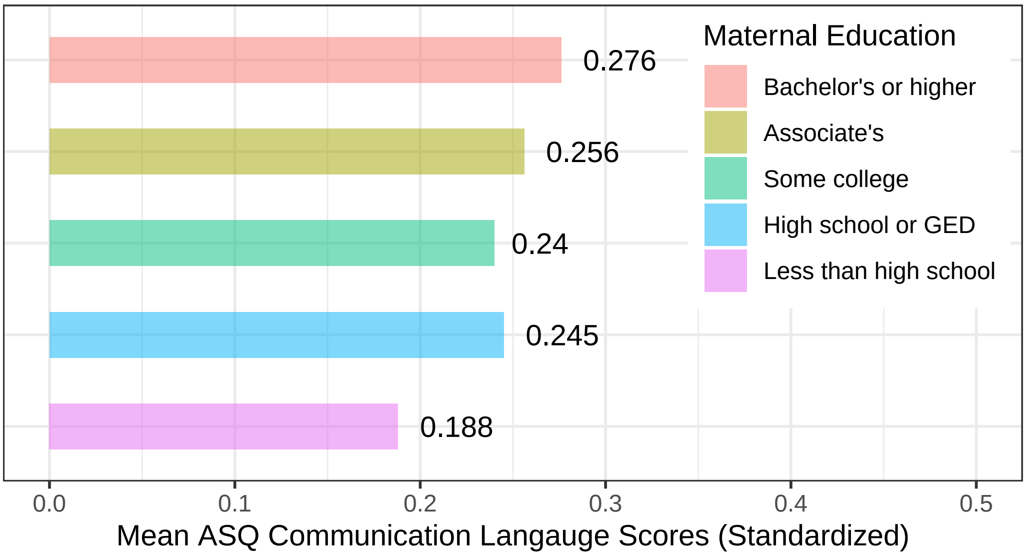

Fig. 4 Mean ASQ communication language scores by Maternal Education

Model Development

Model Selection

- Elastic net was chosen because of number of cases was low (<2,000).
- Random forest was also chosen because high number of features (˜170)

Hyperparameter Tuning

- Elastic net
  - Monte Carlo resamples
  - Regular grid research with 500 levels for both mixture and penalty
  - Evaluation metric was RMSE
- Random Forest
  - Monte Carlo resamples
  - 1000 trees

  - Regular grid research with 50 levels for mtry (between 10 to 100) and min_n
  - Evaluation metric was RMSE

Model Performance

- RMSE was used as the performance metric

Evaluation on Testing Set

Elastic net final model performance on testing set:

```
## # A tibble: 2 x 4
##   .metric .estimator .estimate .config
##   <chr>   <chr>          <dbl> <chr>
## 1 rmse    standard     0.841   Preprocessor1_Model1
## 2 rsq     standard     0.00237 Preprocessor1_Model1
```

(Optional) Comparison Across Models

- Compare the elastic net model results with a random forest model.

Discussion

Implications and Use Cases Revisited

Limitations and Future Work

Appendix

References

Noble KG, Magnuson K, Gennetian LA, et al (2021) Baby's First Years: Design of a Randomized Controlled Trial of Poverty Reduction in the United States. Pediatrics 148(4):e2020049702. https://doi.org/10.1542/peds.2020-049702, URL https://publications.aap.org/pediatrics/article/148/4/e2020049702/181277/Baby-s-First-Years-Design-of-a-Randomized