STAT 151A Fall 2021 Project Proposal

Isaac Cheong, Sam Tan, Max Zhang

Introduction

What makes some second generation Americans more adaptive to mainstream society than others? In this report, we set out to investigate this question by predicting an important factor in overall happiness: income. Using survey data provided by the Center for Migration and Development at Princeton University, we will predict the income level of 2500+ second generation Americans residing in San Diego and Miami in 2005 using baseline information on their families and their own demographic characteristics, such as their language use, self-identities, and academic attainment.

Data Description

We have 400+ variables on information obtained during three different survey rounds. The first survey was conducted in 1991, when respondents were in eighth and ninth grade. This survey established baseline information on immigrant families and had a sample size of 5562 respondents. The second survey was conducted in 1995, as respondents were about to graduate high school. This survey studied the development of respondents through their adolescent years; 4288 respondents were able to be reached. At the same time as this survey was conducted, half of parents were interviewed to directly establish characteristics of immigrant parents. Finally, the last survey was conducted in 2005, when respondents had reached adulthood. 3613 respondents were reached, and asked questions about their adult life, including their monthly income.

We will use data from the first and second survey to predict their reported income in the third survey. The prediction will be meaningful because there is a 10-years-gap between the second and the third survey.

Highlighted examples of potential variables from the first and the second survey are listed below:

- **English knowledge 1995-1996** (`C5`): Respondent's self-reported knowledge of English, on a scale of 1 to 4.
- **Family current economic situation in 1991** (`V44`): "What do you think your family's economic situation is?". On a scale of 1 (Wealthy) to 5 (Poor).
- **Grade point average in 1991** (`V139`): GPA from 0.0 to 5.0.
- **Stanford reading achievement total score in 1991** (`V135`): Total score on Stanford Reading Achievement Test, maximum of 830.
- **Child raising customs** (`P140`): "Do you want your child to be raised according to the customs of your own country or according to American customs?", possible answers include "Own Country", "American Customs", "Both/Mix of the Two", and "Other".
- **People still discriminate regardless of education** (`V322`): "Please answer how true each statement is for you. No matter how much education I get, people will still discriminate against me." On a scale of 1 (Very true) to 4 (Not true at all).

EDA

In our Exploratory Data Analysis, we will examine each response and explanatory variable to check if these variables fit the basic assumptions of linear regression to provide an accurate and concise context for our future analysis, including model selection and model diagnostic.

If we look at the histogram of our response variable of income (see Appendix), we observe that it is heavily skewed to the right, violating our normality assumption of the response variable. To fix this issue, we will apply a log transformation. The log-transformed income histogram has an approximately normal distribution. We are not excluding the maximum point of the histogram because it is reasonable that someone earns $17,000 per month, and it is hard to consider this point as an error from the data collection process.

Our data analysis on explanatory variables will be more complicated since we have many explanatory variables. We will first use LASSO to help us narrow down the variables that are highly predictive, and then examine those variables in detail. To identify any unusual data point or outliers in our explanatory variables, we will do univariate visualization on each explanatory variable. We will then use a correlation matrix with the goal of checking the correlation between income and our explanatory variables as well as between explanatory variables themselves. This will also provide us information about whether we need to re-group our categorical variable to dichotomous variables like "Yes" or "No" or changing our categorical variable to a numerical variable. We will also create coplots that help us identify any possible interaction variables.

Model Selection

There are 3 major methods for selecting models: shrinkage (ridge and LASSO), greedy selection (stepwise/forward/backward selection) and optional (exhaustive) subset selection. Since we have roughly around 400 categorical variables (each with roughly 4 values), we will have around 1600 individual dummy variables. Exhaustive subset selection will be far too computationally intensive, but greedy selection will only explore a small number of possible models relative to all the possible model options (2^{1600}) out there. Shrinkage methods like LASSO will have good performance for models with many possible features like ours because it remains effective regardless of the number of explanatory variables. In particular, we will use LASSO to make predictions because it will provide sparser coefficients, which allows us to use those remaining variables to do inference in later sections.

To do LASSO, we will create a new design matrix that excludes the intercept, and then standardize all the variables, since the scale of each variable is different and all weights will be treated equally during the shrinkage process. By standardizing the variables, only the important ones will be kept after the shrinkage. We will choose the penalty size λ based on cross validation.

Model Diagnostics

Because our data is mostly categorical data, e.g. bounded between 1 and 5. So the possibility of having high-leverage/influential points is small. Therefore, we won't put much focus into analyzing them. However, we will check the four major assumptions for the linear model. To check whether the mean of the errors is centered at zero, we will plot residuals against all the exploratory variables and do feature engineering if necessary to transform specific variables that don't have a zero mean. To check whether the errors have constant variance, we will plot fitted values of y against the residuals. If larger values of y correspond to a large variance of residuals, we will do bootstrap by cases during inference. To check normality for our residuals, we will make a normal Q-Q plot of standardized residuals plotted against our theoretical quantiles, checking if the points line up across the $y = x$ line. If it's not, then we will rely on bootstraps by cases to do inference.

Appendix

Loading in Dataset

```
load('../ICPSR_20520/DS0001/20520-0001-Data.rda')  # creates df called da20520.0001
cils <- da20520.0001
```

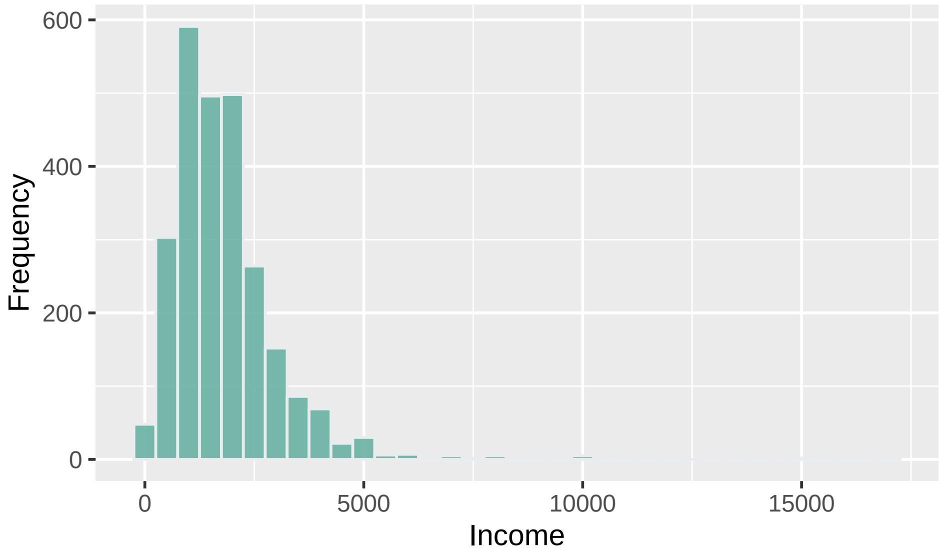
Sample of Dataset

```
cils[1:10,1:5]
```

```
##    CASEID   V1                 V2             V4                V5
## 1       1  257          (1) Miami  (07) School 7 (08) Eighth grade
## 2       2 2347          (1) Miami (13) School 13  (09) Ninth grade
## 3       3  860          (1) Miami (12) School 12  (09) Ninth grade
## 4       4 5178 (3) Ft. Lauderdale (20) School 20  (09) Ninth grade
## 5       5 1984          (1) Miami (12) School 12  (09) Ninth grade
## 6       6 1067          (1) Miami  (07) School 7 (08) Eighth grade
## 7       7 2344          (1) Miami  (07) School 7 (08) Eighth grade
## 8       8 1647          (1) Miami  (09) School 9 (08) Eighth grade
## 9       9 2355          (1) Miami  (09) School 9 (08) Eighth grade
## 10     10 1666          (1) Miami  (05) School 5 (08) Eighth grade
```

Distribution of Income

```
ggplot(cils, aes(x = V421)) +
  geom_histogram(na.rm = TRUE, binwidth = 500,
                 fill="#69b3a2", color="#e9ecef", alpha=0.9) +
  xlab("Income") +
  ylab("Frequency")
```



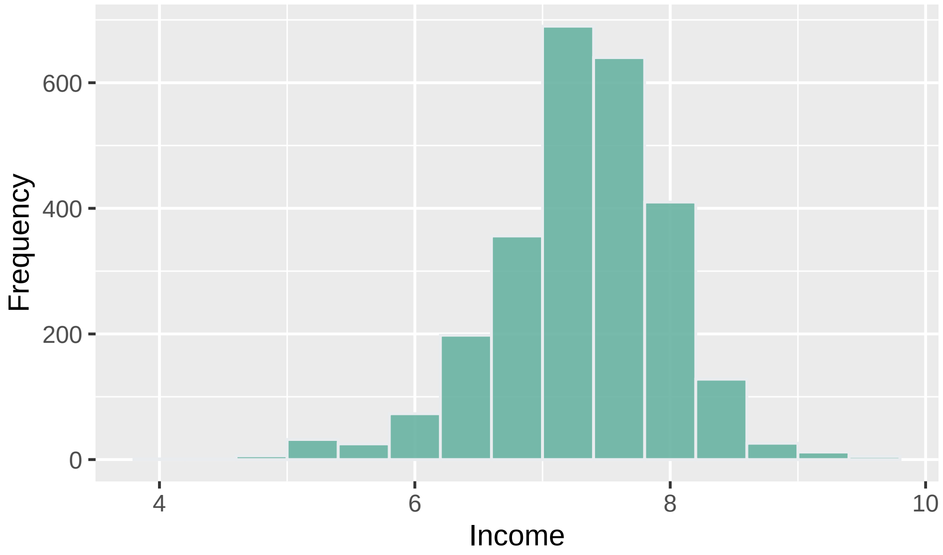
Maximum Income

```
max(cils$V421, na.rm = TRUE)
```

```
## [1] 17000
```

Histogram of Log Transformed Income

```
ggplot(cils, aes(x = log(V421))) +
  geom_histogram(na.rm = TRUE,binwidth = 0.4,
                 fill="#69b3a2", color="#e9ecef", alpha=0.9) +
  xlab("Income") +
  ylab("Frequency")
```


Frequency
Income