Using Demographic Information to Predict Income Class

Allie Palko, Daniel Detrick, Jane Wakefield, Megan Schlag

Introduction

The 1990s were a time of economic growth and job creation. With this in mind, we wanted to look to see if the income levels reflected the economy. We decided to do this by exploring the Current Population Survey of 1994 given by the US Census Bureau. The data gave us 15 variables to explore in order to see if there are predictors that make a person more likely to make $50,000 a year. These variables include education level, race, sex, age, etc. The individuals in the data were a randomly selected group of adults that were over the age of 16.

Upon searching Ohio State library databases and Google Scholar, very limited literature was available. None of this available literature was applicable to our scientific question and dataset.

With that, our question is the following: *Since this was a time of economic prosperity, we want to know, economy aside, what are good predictors for American income exceeding $50,000 per year?*

EDA

The original variables in the data set we looked at for each person were: age, education level, marital status, race, sex, capital gain, capital loss, hours working per week, native country, and whether or not income is greater than $50,000.

We decided to turn education level into a continuous variable and described it to be the years the person was in school. This variable looked appropriately linear with logit of probability (Figure 1) of making over $50,000 a year.

We then examined age (Figure 2), we saw that it had a curve, so we decided it was likely that age would

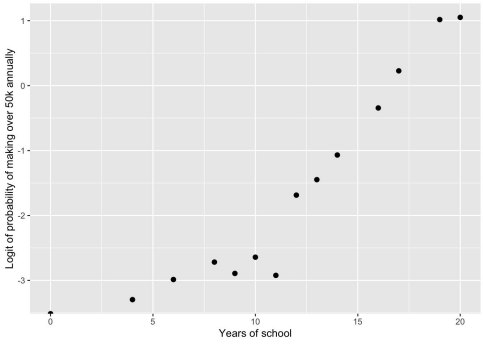

Figure 1: The logit of probability vs years of school

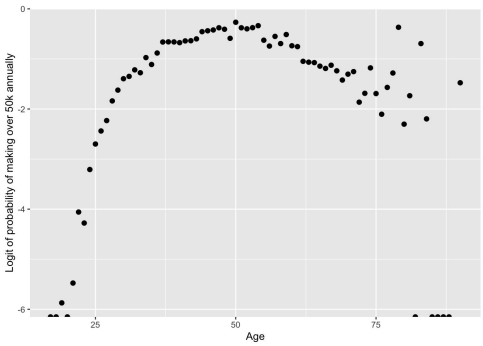

Figure 2: The logit of probability vs age

need a transformation (we estimated a transformation to a fourth power).

We also looked at the native country variable and decided to group it instead by continent, to lessen the complexity of the model. We also made a single variable to encompass capital gain and capital loss, which was capital gain minus capital loss, or net capital change.

Looking at the marriage status variable (Figure 3), people that are married to a civil spouse or an armed forces spouse had similar logits of the probability of making over $50,000 annually, while the rest of the marriage statuses had lower, similar logits. Because of this, we grouped the two similar, higher logit probabilities, as well as grouping the rest of the similar lower ones.

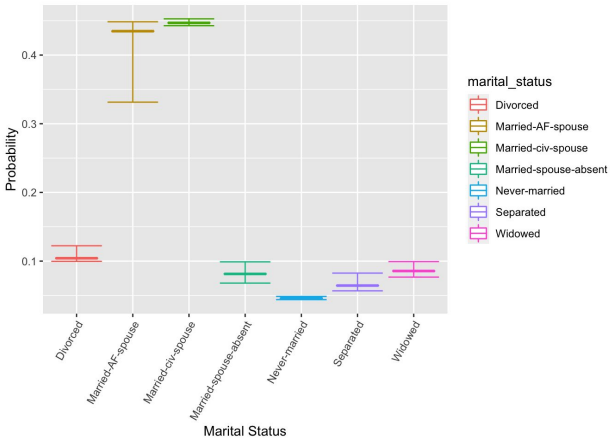

Figure 3: Variation in marital status

We then looked into any possible interactions. First we looked at age and sex (Figure 4). It looked as though the logit of the probability of making over $50,000 annually increases more quickly for males than females. We therefore decided to examine this interaction in our models.

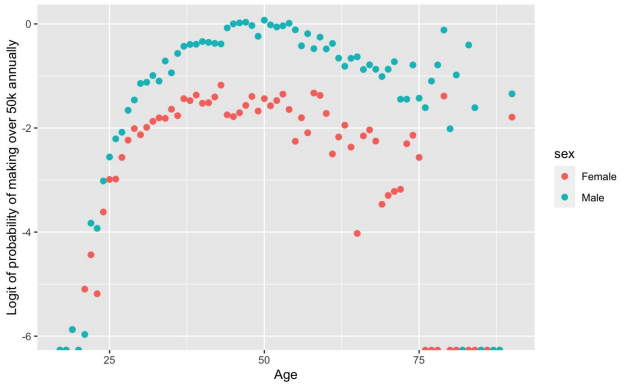

Figure 4: Logit of probability vs. age, grouped by sex

The other interaction we examined was age and years of school (Figure 5). The logit of the probability of making over $50,000 annually grows much slower with age for people with less schooling than for people with more schooling.

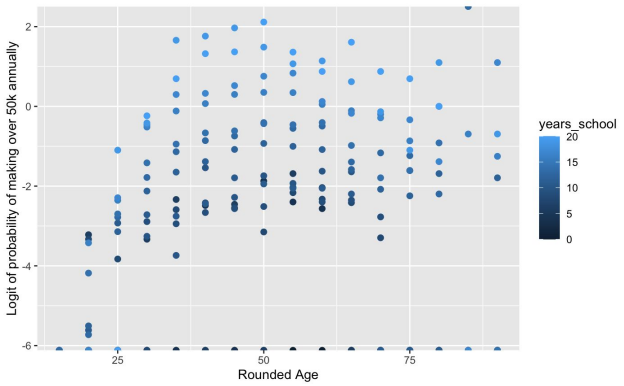

Figure 5: Logit of probability vs.rounded age, grouped by years of school

Model Building & Selection & Diagnostics

Initially, we began building our model by focusing in on the variables age, years of school, the binary variable determining marriage status, race, sex, hours worked per week, net capital change, and the individuals' native continent, each variable was deemed significant at an alpha level .05, and the AIC for this model was 22,639. The best linear model for this is logistic regression, as our response is a binary variable. We chose to focus on AIC as our measure of overall model fit as we wanted to minimize model complexity. We did not use any residual plots in our diagnostics because the residuals prove meaningless in logistic regression.

Next, we decided to explore possible variable transformations, and chose to start with the variable age. Initially seeing the plot of this variable in our EDA, we thought age would require a fourth power transformation. Yet, when adding these terms to our model, we found only age and age squared were statistically significant based on the deviance residuals and the model summary. Thus, we added the age squared term to our model, as it lowered the AIC to 22,011.

After exploring transformations, we began looking into interactions. Upon identifying an interesting relationship between age and sex in the EDA, we added that interaction term to our model. After adding this term, we found that this interaction was not statistically significant based on the deviance residuals and model summary (p-value = 0.064 in model summary). It is also important to note that any interaction that included age_sq was not statistically significant, and therefore was not included in the model.

The next model that we explored included our base model, age_sq, and an interaction with age and years of school. The deviance residuals confirmed that this interaction term was statistically significant (p-value = 0.003). This lowered our AIC to 22,004, and thus we decided to include it in further model building.

For our next model, we looked into adding an interaction between age and marital status. This brought the AIC to 21,966, which is the lowest AIC that we discovered. With all of the terms being statistically significant based on both the

deviance residuals and the model summary, and the model lending to a fairly low AIC, we decided that this would be our final model. This final model includes our base model, an age_squared term, an interaction with years school and age, and an interaction with age and marital status.

Discussion & Results

Based on our final model we chose the following variables as the most important to answering our overarching question, the interpretation for each of those variables is as follows (after adjusting for all other variables):

- People that are married to a civil spouse or armed forces spouse have 1.17 times the odds of making over $50,000 annually. Although income was for an individual, we saw that marriage status varied greatly and included this term in our final model.
- Males have odds of making over $50,000 that are 1.05 times higher than females. Overall, sex is one of the more interesting questions in the income discussion so we included this term.
- For every ten hours more a week a person works, their odds of making over $50,000 a year increase by a multiplicative change of 1.02. We decided to include this term because we saw the difference between a part-time worker and a full-time worker had an effect on the final model.
- For every $1,000 increase in net capital, the odds of making over $50,000 annually increase by a multiplicative factor of 1.01. Capital gains and capital losses are extremely indicative of financial health, so we thought net cap was a very important term to include to answer our question.
- People originally from the Carribean have odds of making over $50,000 annually that are 1.44 times higher than for people originally from Asia. People originally from Central America have odds of making over $50,000 annually that are 1.25 times higher than for people originally from Asia. People originally from Europe have odds of making over $50,000 annually

that are 1.24 times higher than for people originally from Asia. People originally from the Middle East have odds of making over $50,000 annually that are 1.56 times higher than for people originally from Asia. People originally from North America have odds of making over $50,000 annually that are 1.19 times higher than for people originally from Asia. People originally from South America have odds of making over $50,000 annually that are 1.57 times higher than for people originally from Asia. People whose native region of the world is unknown have odds of making over $50,000 annually that are 1.21 times higher than for people originally from Asia. We thought this was interesting because the native region of the world where people are from seems to have a strong impact on their income.

We wanted to find a way to test our data to see its accuracy level in predicting who would make more than $50,000 a year depending on certain traits. To do this we decided to take a random sample of individuals from the data provided and test our model using those values. We decided that we would count any probability greater than 0.50 as the individual makes over $50,000/year and any probability less than or equal to 0.50 would mean that the individual does not make $50,000/year. We took a random sample of 20% of the people in our dataset, created a model with the other 80%, and found that we correctly guessed 5,474 of the 6,512 people sampled, giving us about an 83% accuracy rate.

Overall, our model showed that factors such as age, marital status, hours per week, years of education and continent of origin are all good predictors for American income exceeding $50,000 per year. With the age predictor, we found that increasing age increased the probability of earning more than $50,000/year until about the age of 50. However, after the age of 50, an increase of age begins to decrease the probability of making more than $50,000. The same pattern follows the age squared variable when increasing.

One shortcoming of this analysis is that the dataset is very large, with about 32,000 records. With this, just about any predictor could be deemed statistically significant. Thus, other statistics that are beyond our knowledge and the scope of this course might be helpful in future analyses to understand the true significance of the predictors.

Ideas For Further Analysis

Understanding that this data is a bit outdated and gives a view of only one point in time, we think it would be interesting to explore more recent census data at varied states of the overall health of the economy. We could do this by exploring census data from 2007 in a poor economy and then census data from 2019 in a strong economy. Additionally, we think it would be interesting to have the data for which region of the country each person resides in so that we could explore differences along the geographic locations. If we were able to break down the regional data by city, we would be able to standardize the $50,000 to each city's standard of living and then from there see if each individual makes more or less than the flat figure of $50,000. We also wanted to explore the differences between those living in rural, suburban, and urban areas and look at the high-income areas versus low-income areas. These differences may allow us to find interesting correlations that we cannot see with our initial data. We understand that if we wanted to investigate these different analyses we would have to join different datasets and verify their accuracy before starting our analysis.

Appendix

Code for variable transformations:

```
adult$over_50 = case_when((adult$income == '>50K') ~ 1, TRUE ~ 0)
adult$years_school = case_when(adult$education == 'Preschool' ~ 0,
                  adult$education == '1st-4th' ~ 4,
                  adult$education == '5th-6th' ~ 6,
                  adult$education == '7th-8th' ~ 8,
                  adult$education == '9th' ~ 9,
                  adult$education == '10th' ~ 10,
                  adult$education == '11th' ~ 11,
                  adult$education == '12th' ~ 12,
                  adult$education == 'HS-grad' ~ 12,
                  adult$education == 'Some-college' ~ 13,
                  adult$education == 'Assoc-acdm' ~ 14,
                  adult$education == 'Assoc-voc' ~ 14,
                  adult$education == 'Bachelors' ~ 16,
                  adult$education == 'Masters' ~ 17,
                  adult$education == 'Prof-school' ~ 19,
                  adult$education == 'Doctorate' ~ 20
)
adult$isMarried = ifelse(adult$marital.status == "Married-civ-spouse" |
adult$marital.status == "Married-AF-spouse",1,0)
adult$isMarried = as.factor(adult$isMarried)
adult$isUS = ifelse(adult$native.country == "United-States",1,0)
adult$net_cap = adult$capital.gain - adult$capital.loss
adult$continent = case_when(adult$native.country == 'United-States' ~ 'North America',
                     adult$native.country == 'Greece' ~ 'Europe',
                     adult$native.country == 'Taiwan' ~ 'Asia',
```

```
adult$native.country == 'Trinadad&Tobago' ~ 'Carribean',
adult$native.country == 'Holand-Netherlands' ~ 'Europe',
adult$native.country == 'Iran' ~ 'Middle East',
adult$native.country == 'Italy' ~ 'Europe',
adult$native.country == 'Honduras' ~ 'Central America',
adult$native.country == 'Cambodia' ~ 'Asia',
adult$native.country == 'Dominican-Republic' ~ 'Central America',
adult$native.country == 'Hungary' ~ 'Europe',
adult$native.country == 'Jamaica' ~ 'Carribean',
adult$native.country == 'Yugoslavia' ~ 'Europe',
adult$native.country == 'Laos' ~ 'Asia',
adult$native.country == 'Vietnam' ~ 'Asia',
adult$native.country == 'India' ~ 'Asia',
adult$native.country == 'Canada' ~ 'North America',
adult$native.country == 'Puerto-Rico' ~ 'Central America',
adult$native.country == 'England' ~ 'Europe',
adult$native.country == 'Japan' ~ 'Asia',
adult$native.country == 'Cuba' ~ 'Central America',
adult$native.country == 'Peru' ~ 'South America',
adult$native.country == 'Haiti' ~ 'Carribean',
adult$native.country == 'Columbia' ~ 'South America',
adult$native.country == 'Ecuador' ~ 'South America',
adult$native.country == 'Scotland' ~ 'Europe',
adult$native.country == 'Thailand' ~ 'Asia',
adult$native.country == 'Mexico' ~ 'Central America',
adult$native.country == 'China' ~ 'Asia',
adult$native.country == 'Philippines' ~ 'Asia',
adult$native.country == 'South' ~ 'Unknown',
adult$native.country == 'Poland' ~ 'Europe',
```

```
                adult$native.country == 'Germany' ~ 'Europe',
                adult$native.country == 'Hong' ~ 'Asia',
                adult$native.country == 'Ireland' ~ 'Europe',
                adult$native.country == 'Nicaragua' ~ 'Central America',
                adult$native.country == 'El-Salvador' ~ 'Central America',
                adult$native.country == 'Guatemala' ~ 'Central America',
                adult$native.country == 'France' ~ 'Europe',
                adult$native.country == 'Portugal' ~ 'Europe',
                adult$native.country == 'Outlying-US(Guam-USVI-etc)' ~ 'Asia',
                adult$native.country == '?' ~ 'Unknown',
                TRUE ~ 'Yes')
adult$age_sq = adult$age^2
```

Final model coefficients output:

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.767e+01  5.695e-01 -31.032  < 2e-16 ***
age                        2.980e-01  1.354e-02  22.010  < 2e-16 ***
age_sq                    -2.406e-03  1.097e-04 -21.923  < 2e-16 ***
years_school               4.794e-01  2.993e-02  16.016  < 2e-16 ***
isMarried1                 3.252e+00  1.611e-01  20.183  < 2e-16 ***
raceAsian-Pac-Islander     5.919e-01  2.572e-01   2.301  0.02139 *
raceBlack                  4.071e-01  2.251e-01   1.808  0.07060 .
raceOther                  1.826e-01  3.432e-01   0.532  0.59455
raceWhite                  6.731e-01  2.153e-01   3.126  0.00177 **
sexMale                    1.550e-01  4.808e-02   3.223  0.00127 **
hours.per.week             2.411e-02  1.515e-03  15.922  < 2e-16 ***
net_cap                    2.461e-04  8.203e-06  29.998  < 2e-16 ***
continentCarribean        -3.895e-03  3.643e-01  -0.011  0.99147
continentCentral America -3.988e-01  2.201e-01  -1.812  0.06995 .
continentEurope            3.506e-01  2.120e-01   1.654  0.09813 .
continentMiddle East       1.523e-02  4.471e-01   0.034  0.97283
continentNorth America     2.403e-01  1.737e-01   1.383  0.16651
continentSouth America    -1.036e+00  4.484e-01  -2.311  0.02082 *
```

```
continentUnknown          -2.104e-01  1.925e-01  -1.093  0.27445
age:years_school          -2.587e-03  6.374e-04  -4.059 4.92e-05 ***
age:isMarried1            -2.338e-02  3.656e-03  -6.393 1.62e-10 ***
```

Final Model ANOVA Deviance Residuals:

```
              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                             32560      35948
age            1   1746.6     32559      34202 < 2.2e-16 ***
age_sq         1   2281.0     32558      31921 < 2.2e-16 ***
years_school   1   3387.9     32557      28533 < 2.2e-16 ***
isMarried      1   4527.9     32556      24005 < 2.2e-16 ***
race           4     57.7     32552      23947 8.974e-12 ***
sex            1     36.0     32551      23911 1.959e-09 ***
hours.per.week 1    322.5     32550      23589 < 2.2e-16 ***
net_cap        1   1567.9     32549      22021 < 2.2e-16 ***
continent      7     47.8     32542      21973 3.839e-08 ***
age:years_school 1    8.5     32541      21964  0.003461 **
age:isMarried  1     40.2     32540      21924 2.315e-10 ***
```

Code for calculating model accuracy:

```
# THIS SHOWS OUR FINAL MODEL HAS A 83% (5474 OUT OF 6512) HIT RATE
(ASSUMING PROB >50% IS A 1)
data_rows_saved = adult[sample(nrow(adult), nrow(adult)*.2), ]
adultnew = setdiff(adult, data_rows_saved)
final_model = glm(over_50 ~ age + age_sq + years_school + isMarried + race + sex +
hours.per.week + net_cap  + continent + age:years_school + age:isMarried, data =
adult, family= binomial)
actual = data_rows_saved$over_50
logits = predict(final_model, data_rows_saved)
probs = exp(logits)/(1+exp(logits))
guess = case_when(probs > .5 ~ 1, TRUE ~ 0)

predic_vals = data.frame(actual, guess, probs)
```

```
sum(predic_vals$actual == predic_vals$guess)
#predic_vals = cbind(predic_vals,data_rows_saved)
predic_vals
```