Problem Set 3: Income Prediction Challenge

Due: 5pm on Wednesday, March 6.

Student identifer: 5956

- Use this .qmd template to complete the problem set
- In Canvas, you will upload the PDF produced by your .qmd file
- Put your identifier above, not your name! We want anonymous grading to be possible

This problem set is connected to the PSID Income Prediction Challenge from discussion.

Income Prediction Challenge

Collaboration note. This question is an individual write-up connected to your group work from discussion. We expect that the approach you tell us might be the same as that of your other group members, but your answers to these questions should be in your own words.

1.1 (5 points) How did you choose the predictor variables you used? Correct answers might be entirely conceptual, entirely data-driven, or a mixture of both.

I created 4 models for this. The first two models were created based on my interpretation of what would impact a person's income. The first model focused more on the analysis of the parent's status (income and education) on the child's income. I included both income and education, despite the conceptual dependence of income of education, because I did a proportion table to see the probability that a parent earns at a specific quantile given their education and the percentages did not seem high enough. I then did another model that replaced the parent's education with the child's education because conceptually, that seemed to have a more direct impact on the child's income. In both of these models, I included the child's race and sex because as we previously analyzed during assignments, gender inequality and racial inequality is still very prevalent in the workforce, The third and fourth models are the all variables and no variables model.

Using these models, I ran to find the MSE values for the training and testing data multiple times. Of every trial I did, which generated a different set of training data and testing data, model 2 (Income ~ Parent Income + Sex + Race + Education) and model 3 (g3_log_income

~ g2_log_income + g1_log_income + g3_educ + g2_educ + g1_educ + race + sex),seemed to do the best in predicting the child's log income. In addition, when testing for the AIC (Akaike Information Criterion), which is a good tool of measuring how much information is lost and penalized the use of too many variables, model 2 and model 3 had very similar AIC values. Given, this I decided to use a model with less variables (model 2) to reduce the amount of noise that extra variables can include.

1.2 (5 points) What learning algorithms or models did you consider, and how did you choose one? Correct answers might be entirely conceptual, entirely data-driven, or a mixture of both.

I used the linear models to narrow down the select of variables I wanted to use in my final prediction– model 2 (Income ~ Parent Income + Sex + Race + Education) and model 3 (g3_log_income ~ g2_log_income + g1_log_income + g3_educ + g2_educ + g1_educ + race + sex). then I used the penalized regression to see if the few children with very low log_income that I observed in my data exploration part impacted the results at all. The MSE values for the penalized regression versions of model 2 and model 3 still turned out to be lower than that of model 2 for every one of the 20 run that I did with my code. Therefore, for my final prediction, I selected the linear model of model 2 (Income ~ Parent Income + Sex + Race + Education), which is the model that yielded consistently lowest MSE values for the training and testing data.

1.3 (20 points) Split the `learning` data randomly into `train` and `test`. Your split can be 50-50 or another ratio. Learn in the `train` set and make predictions in the `test` set. What do you estimate for your out-of-sample mean squared error? There is no written answer here; the answer is the code and result.

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(rsample)
```

```r
train_df <- read.csv("learning.csv")
# Used a 70 30 split for train and test to ensure model is more accurate.
split <- initial_split(train_df, prop = .7)
test_df <- read.csv("holdout_public.csv")

# Model: Income ~ Parent Income + Sex + Race + Education
lm2 <- lm(
  g3_log_income ~ as.factor(g3_educ) + as.factor(sex) +
    as.factor(race) + g2_log_income,
  data = training(split)
)

# No Model: Income ~ 1
lm4 <- lm(g3_log_income ~ 1, data = training(split))

# MSE with model and without model
testing(split) |>
  mutate(
    predicted_model = predict(lm2, newdata = testing(split)),
    predicted_no_model = predict(lm4, newdata = testing(split)),
    squared_error_model = ((g3_log_income - predicted_model) ^ 2),
    squared_error_no_model = ((g3_log_income - predicted_no_model) ^ 2),
  ) |>
  select(starts_with("squared")) |>
  summarize_all(.funs = mean)
```

```
  squared_error_model squared_error_no_model
1           0.3690934              0.4849923
```

Create a new task

The predictability of life outcomes is not likely to be the same in every setting. Imagine you were designing a challenge like this one in a new setting, to study how outcomes change over the life course or across generations.

2.1 (5 points) From what population would you draw your sample?

I want to draw from two populations, those born in the 1990s - 2000s and 1960s - 1980s in the top 10 countries that has the highest GDP as of 2024 (in terms of USD). These two cohorts would allow me to capture the child's generation and parent's generation to analyze the change in marriage trends across generations.

Those countries were we would look to sample from is US, China, Japan, Germany, India, UK, France, Brazil, Canada, Italy (Forbes).

2.2 (5 points) What outcome would you study?

The response variable that will be studied in this task would be whether or not the individual is married.

I want to study the marriage trends of people born in the 1990s - 2000s and see if factors that impact marriage remained the same as their parent's generation. Specifically, I am focusing on countries with highest GDP in the world to analyze the change in marriage rates given that those countries have a good economy because a better economy may mean better access to channels that can empower individuals to pursue their dream life and introduce more modern ideas. In many traditional views, marriage is essential to life and happiness and thus marriage is more prevalent in the parent's generation. However, over the course of these years, marriage rates has decreased in many countries such as US and China. I am curious to see if the reason for that drop in marriage rates can be attributed to the change in factors considered when an individual is deciding whether or not they are getting married.

2.3 (5 points) What predictors would you include?

Predictors I would include is whether or not their parent had divorce, age (20-55), income, education, race, country, whether or not the individual wants children, whether or not their parents encouraged marriage/ arranged marriage, and religion.

2.4 (5 points) Why would it be interesting in your setting if predictions were accurate? Why would it be interesting if predictions were inaccurate?

For this analysis, I would train models using the parent's generation data and use the train-test split to select the better model. Then, I would use the model to predict the child's generation. If the child's marriage status could be predicted accurately with a high F1 and recall scores, this means that factors that influence marriage has remained relatively the same. However, if the child's marriage status could not be predicted accurately, this will support the claim that the factors that influence an individual born in the 1990s - 2000s is different from that of those born in the 1960s - 1980s.

Grad. Machine learning versus statistics

> This question is required for grad students. It is optional for undergrads, and worth no extra credit.

20 points. This question is about the relative gain in this problem as we move from no model to a statistical model to a machine learning model.

First, use your `train` set to estimate 3 learners and predict in your `test` set.

a) No model. For every `test` observation, predict the mean of the `train` outcomes
b) Ordinary Least Squares. Choose a set of predictors $\vec{X}$. For every `test` observation, predict using a linear model `lm()` fit to the `train` set with the predictors $\vec{X}$.
c) Machine learning. Use the same set of predictors $\vec{X}$. For every `test` observation, predict using a machine learning model fit to the `train` set with the predictors $\vec{X}$. Your machine learning model could be a Generalized Additive Model (`gam()`), a decision tree (`rpart()`), or some other machine learning approach.

Report your out-of-sample mean squared error estimates for each approach. How did mean squared error change from (a) to (b)? From (b) to (c)?

Interpret what you found. To what degree does machine learning improve predictability, beyond what can be achieved by Ordinary Least Squares?