Analyzing worker’s earnings via descriptive components

Frances Ling, Raymond Sheng, Yang Yu

Introduction

- Our final project focuses on exploring the relationship between earnings and various worker characteristics
- We hypothesize that machine learning models built using only married workers provides similar models to surveying all workers
- We justify this approach as there could be less overhead by only surveying married couples (e.g. collect earnings information when they apply for a marriage license)
- We use a subset of the 1994 U.S. National Health Interview Survey
- We will compare the predictive accuracy of machine learning models trained on all available data with the accuracy of the same models using only married workers' data

Background and Motivation

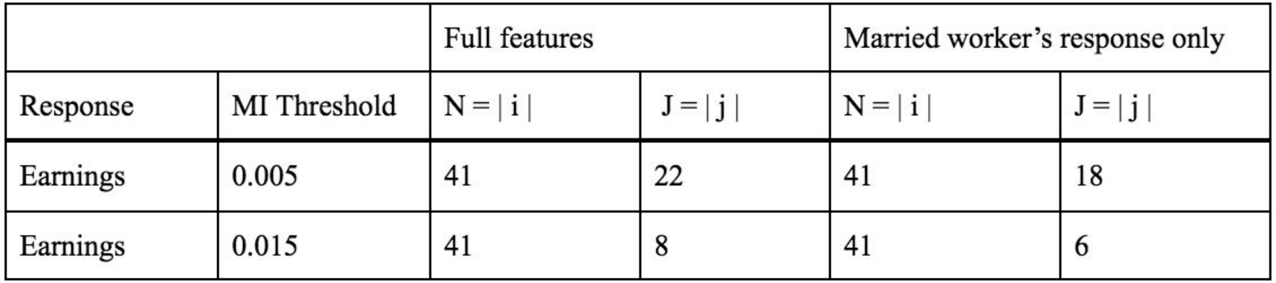
- Original analysis conducted used data set to see relation between earnings and worker features (e.g. height, age, education level, geographic location, race, gender, etc…)
- There is a correlation between height and job status/earnings
  - On average, taller individuals hold more prestigious jobs and earn more than shorter workers
  - An increase of 4-5 inches correlates with an increase in salary between 9 and 15 percent
- Possible explanations
  - Better non-cognitive skills (more social skills and higher self-confidence)
  - Better cognitive skills (at age 3, taller children perform better on tests)
- We want to test whether models built only using married workers produces similar predictive accuracy to models built using the entire data set.

Dataset - 1994 U.S. National Health Interview Survey

- Consists of the information collected from 17,870 workers, both male and female
- Identifies 11 characteristics of each worker: *age, class of worker, earnings, education, height, marital status, occupation, race, region, sex*, and *weight*.
- Since there are 6 non-numerical features, we use one-hot-coding to transform them into binary variables.
  - Expands feature-space from 11 variables to 419
- We set aside 70% of the dataset for training and 30% of the dataset for testing

Data Processing and Feature Selection

- We find that the average worker has 13.5 years of education, earns around $46,874, weighs 170 pounds, and is about 5 foot 7 inches tall.
- To avoid overfitting and poor predictions, we use *mutual information regression*
    - Mutual information between variables measures the dependence of one to another.
    - We can remove features that have low dependence with the response, namely earnings, defined by MI < 0.005 and MI < 0.015.



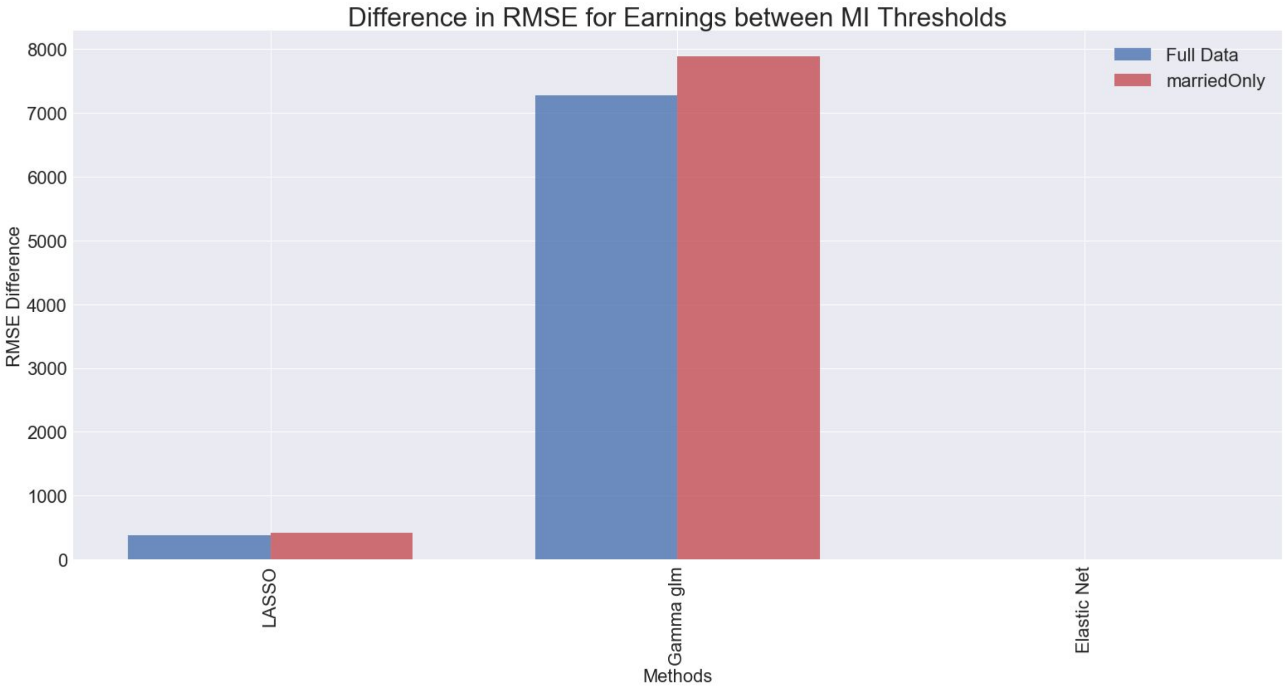
		Full features		Married worker's response only	
Response	MI Threshold	N = | i |	J = | j |	N = | i |	J = | j |
Earnings	0.005	41	22	41	18
Earnings	0.015	41	8	41	6

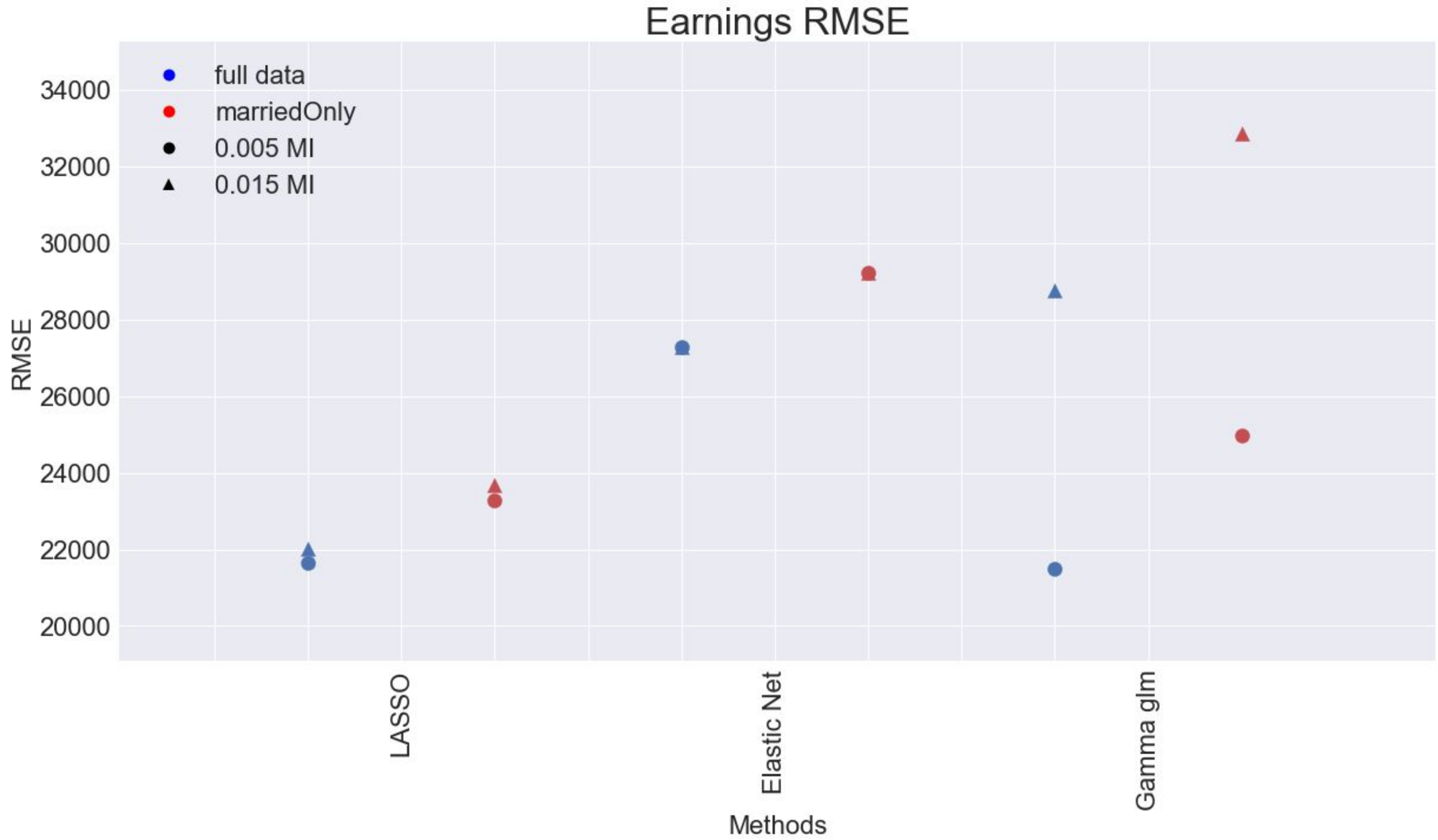
Methods

- We use three machine learning methods to predict our resposnes:
  - Lasso regression
  - Elastic net regression
  - Gamma generalized linear model

Results

In general, the difference in the RMSE on the results from two different kinds of datasets(full vs married couples) is small, which also reveals the fact that marriage status can possibly be the only metric to predict the earnings.




Earnings RMSE
full data
marriedOnly
0.005 MI
0.015 MI
34000
32000
30000
28000
26000
24000
22000
20000
RMSE
LASSO
Elastic Net
Gamma glm
Methods

Evaluation

Machine Learning Models	MI threshold of 0.005		MI threshold of 0.015	
Data	Full	Married	Full	Married
Lasso Regression	21634	23268	22011	23686
Elastic Net	27274	29233	27275	29234
Gamma GLM	21493	24968	28773	32857

Discussion

MI threshold of 0.005:

Full dataset - RMSE with Elastic Net is 26% higher than the RMSE with Lasso regression and 27% higher than the RMSE with Gamma GLM.

Only married couples - RMSE with Elastic Net is 25.6% higher than the RMSE with Lasso and 17% higher than the RMSE with Gamma GLM.

MI threshold of 0.015:

Only married couples - RMSE with Elastic Net is still higher than the RMSE with Lasso regression but lower than the RMSE with Gamma GLM.

Conclusion

Data

- 1994 US National Health Interview Survey
- information collected from 17,870 workers
- identifies several characteristics which have predictive power on the earning of a worker

Show

- marriage status can be an effective predictive tool on a worker's earning

Method

- Gamma generalized linear model, Lasso Regression model and Elastic Net models

Results

- **Earnings can be predicted by factors other than height**
  Marital status is correlated with earnings

- Elastic Net model gives almost the same RMSE results with respect to the full dataset and the dataset with only married couples.