COS 424 Final Project: Analyszing Worker's Earnings via Descriptive Components

Frances Ling
Electrical Engineering
Princeton University
fling@princeton.edu

Raymond Sheng
Computer Science
Princeton University
rsheng@princeton.edu

Yang Yu
Computer Science
Princeton University
yangy@princeton.edu

Abstract

Research has found that height and earnings are correlated, with taller workers earning more than their shorter co-workers, even when controlling for sex. Taller people are found to have higher IQs, thus earning more in the American workplace. In this final project, we look at worker data from the 1994 U.S. National Health Interview Survey. This dataset collects information from 17,870 workers, identifying 11 different features, such as age, weight, sex, etc., for each worker. We hypothesize that our machine learning methods will perform equivalently when trained on the full dataset versus only married workers' data. To fit the appropriate machine learning models to this dataset, we used mutual information regression method to perform feature selection on the sparse data matrix. Then, we used Gamma Generalized Linear Model, Lasso Regression and Elastic Net machine learning models on part of training data and evaluate the performances on the part of testing data by using the metric such as RMSE. Lasso models gives lower RMSE than the baseline of predicting the mean and Lasso and Elastic Net have similar RMSE when only using married couples vs all the data. We were able to reproduce a similar effect of the original study that on average taller people have higher earnings.

1 Introduction

Previous research has shown that there is a correlation between height and job status and earnings. On average, taller individuals hold jobs of more prestige and earn more than shorter workers. In Western countries, an increase in four or five inches (a jump from the 25th percentile of height to the 75th) correlates with an increase in salary between 9 and 15 percent [1].

To explain this trend, researchers hypothesized that taller people have better non-cognitive skills, namely more social skills and higher self-confidence [2], while others have proposed that taller people earn more because they are smarter [3].

For our final assignment, we take a look at a subset of the 1994 US National Health Interview Survey, and we hypothesize that we can use married worker's data to produce models that are of similar predictive accuracy, yielding similar results to using the entire dataset. We argue that these results can be useful in determining which subset of the population to survey in the future. If the working population can be accurately represented by using only married workers' data, such information can be obtained when workers obtain their marriage license, thus saving the government time and resources as there is less overhead when collecting this information during marriage license applications rather than surveying the entire populace.

The dataset consists of the information collected from 17,870 workers, both male and female. 11,641 workers are married, representing more than half of the surveyed workers. The interview 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 categorical features, we use one-hot-coding to transform them into binary variables that results in a feature space of 41 features.

Machine learning models enable researchers to make sense of various survey questions collected per worker, learning the relationship between various factors and earnings. Here, we aim to predict the earnings of these workers using a subset of the dataset, namely using the data of the married workers. 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. Likewise, we hope to reproduce the results of the Case and Paxon study showing that adjusting for inflation and using 2012 dollars on average a person an inch taller than average would be predicted to make $707.7 more than the average person.

2 Related Work

The 1994 U.S. National Health Interview Survey is a popular dataset that has been widely analyzed and reported on. Most famously, researchers found that taller workers earn more than their shorter co-workers which is an effect seen in both sexes.

The National Health Interview Survey data is collected through hour-long in-person interviews at the household of the interviewee [8]. If our hypothesis proves correct, the government can potentially narrow down the pool of potential interview candidates and focus on obtaining information from only married workers. This could prove to be less of a strain on government time and resources if workers are asked to complete a survey when they obtain their marriage license.

3 Methods

3.1 Data Processing and Feature Selection

The dataset consists of 11 features with varying type (e.g. continuous, count, ordered categorical, etc.) for 17,870 workers. 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.

Since sex, marital status, job category, region of the U.S., race, and occupation are categorical data, we use one-hot encoding to represent these features. This expanded the number of features from 11 to 41. We are using a subset of the national survey that only includes workers without missing data, so we do not need to remove any of the data from our current dataset. We set aside 30% of the dataset to be used for testing, and use 70% of the dataset to train our machine learning models.

In order to avoid overfitting we apply feature selection via mutual information regression. This is a feature selection method implemented by the sklearn libraries [4] which separately calculates the mutual information (MI) between each nonmissing response variable and each of the features in our training set. Mutual information between variables measures the dependence of one to another. For example, if two features are completely independent, meaning no information about one can be obtained by knowing the other and vice versa, their mutual information is 0 [5].

We can remove features that have low dependence with the response, namely earnings, defined by MI $<$ 0.005 and MI $<$ 0.015. We then filtered the remaining features in our training set to only include responses from married workers.

		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

Table 1: Feature Space: Description of each training set X_{ij} input into each of our 3 machine learning models. $||i||$ is the sample size, or the number of non-missing responses. $||j||$ is the number of features remaining after applying the mutual information thresholding and/or married worked-only response filtering.

3.2 Machine Learning Methods

We used three machine learning methods were used to predict our responses. Using the Python sklearn libraries, we implemented the following two linear regression methods:

- LASSO Regression (normalize=True)
- Elastic Net (normalize=True, L1 ratio = 0.5)

Cross-validation was used to choose the optimal scaling parameter λ and/or the mixing parameter α that minimizes the mean-squared error between the model and the validation responses.

Given the non-Gaussian nature of our data, we also tested a Gamma generalized linear model with an inverse link function using the Statsmodels libraries.

4 Methods

4.1 Lasso Regression

Lasso (least absolute shrinkage and selection operator) regression is an extension on regularized linear regression [7]. To prevent overgeneralization and overfitting that results from simple linear regression, we use lasso regression to reduce model complexity.

The key difference can be found in the loss function which is defined as follows:

$$\sum_{i=1}^{M}(y_i - \hat{y_i})^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p}|w_j|$$

Instead of just taking the square of the coefficients, magnitudes are also taken into account. This type of L1 regularization can lead to zero coefficients, meaning that some features are ignored when evaluating the response. Thus, Lasso regression aids in reducing overfitting and in feature selection.

4.2 Elastic Net Regression

Elastic Net Regression combines the L1 and L2 penalties of Lasso and Ridge regression. Regularization is used to prevent overfitting the machine learning models to training data by adding noise to the objective function before optimizing the model.

$$\frac{\sum_{i=1}^{n}(y_i - x_i^J\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right)$$

The l_1 portion of the penalty generates a sparse model, and the quadratic part removes the limitation of the number of selected variables, encourages a grouping effect, and stabilizes the l_1 regularization path.

4.3 Gaussian generalized linear model

A generalized linear model can be described by a probability distribution, linear predictor, and a link function. Let Y be a N-vector of response variables and X be a NxJ matrix of corresponding features. Discriminative models assume that the joint probability of each response factors as $p(X_i, Y_i) = p(X_i) * p(Y_i, X_i)$ and estimate the latter term for prediction. For Gamma GLMs, $p(Y_i|X_i) \sim Gamma(k, \theta)$ for some shape k and scale θ. Here, we will derive the relationship between the $Gamma$ parameters and X_{ij}.

Gaussian GLMs are best used for data that are continuous, positive, right-skew and when variance is near-constant on the log-scale.

5 Results and Evaluation

5.1 Evaluation Metrics

5.1.1 RMSE

To compare the performance of the three machine learning methods trained on the data types listed in Figure 1, we calculated the RMSE which is the square root of the mean square error (MSE) between the predicted and true outcomes. MSE measures the average of the squares of errors. In particular, when we want to calculate the error over the training or test data which are used to make an estimated response. The corresponding deviations are called residuals or prediction errors when they are calculated out-of-sample. RMSE is used to measure the accuracy and to compare predicted errors of various machine learning models that are fitted to some typical dataset instead of different datasets.

Using this metric allows us to compare the predictive accuracy for a given response across different regression methods and data subsetting. We note that the magnitude of the MSE is dependent on the range of each response v, making it a poor measure of between-response comparisons.

The general formula for MSE can be defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where n is the total number of points, and y_i is the target response and $\tilde{y_i}$ is the predicted response.

RMSE is a non-negative value. An RMSE value of 0 indicates a perfect fit to the dataset. Therefore, in general, a lower RMSE indicates more accurate results given by the machine learning model. Additionally, the measurement depends on the scale of the number of values we use. In fact, a larger size of squared error indicates better effect of each error given by RMSE, which also means that RMSE is sensitive to outliers.

The mathematics formula of RMSE is defined as following:

$$\mathbf{RMSE}(\hat{\theta}) = \sqrt{\mathbf{MSE}(\hat{\theta})} = \mathbb{E}((\hat{\theta} - \theta)^2)$$

where $\hat{\theta}$ is the estimator with respect to an estimated parameter θ.

We evaluated and compared the RMSE values with respect to predicted earnings from three machine learning models we use, and for each model, there are two RMSE values with different size of dataset (full dataset and the dataset containing only married people). The main results with RMSE is showed in the following graph:

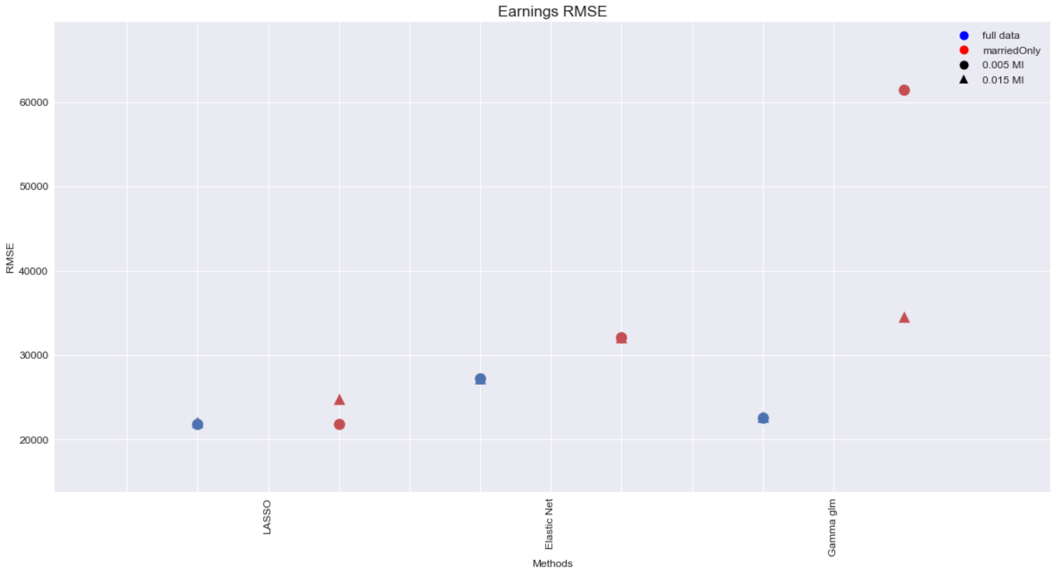


Figure 1: RMSE with respect to predicted earnings

5.1.2 R squared

Another metric we use to evaluate our results is called coefficient of determination (R^2), which measures how closely the data fits the regression. R^2 is defined to be the proportion of the variance in the dependent parameters by predicting them from given independent variables.

The mathematical formula of R^2 is defined as following:

$R^2 = 1 - \frac{S_{res}}{S_{tot}}$

where R_{res} is the residual sum of the squares, and S_{tot} is the total sum of the squares.

We also evaluated and compared the R^2 values with respect to the three machine learning models we used, and for each model, we calculate two R^2 values with different size of dataset (full dataset and the dataset containing only married people). The main results are showed in the following graph:

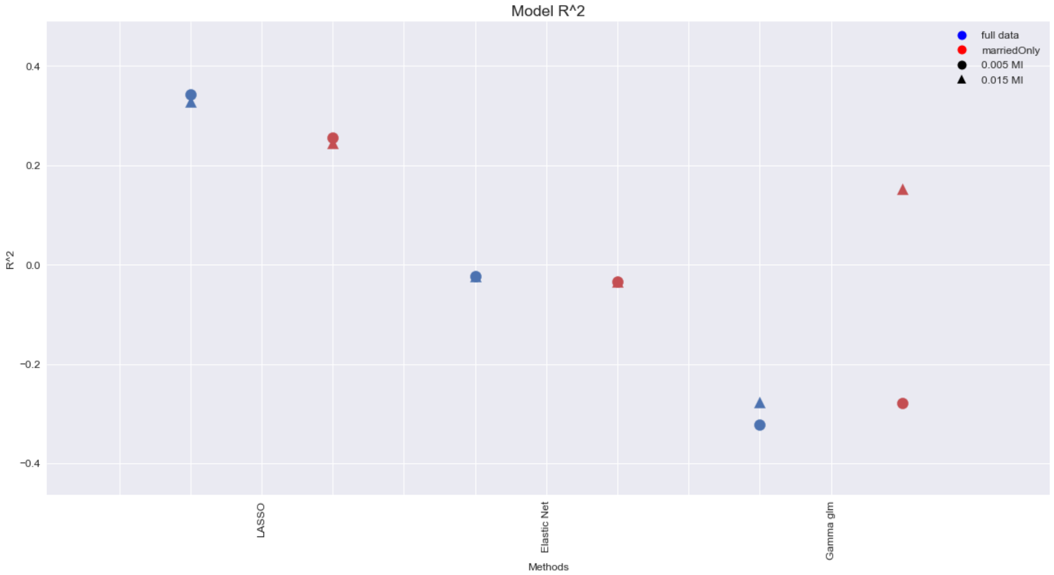


Figure 2: R^2 with respect to predicted earnings

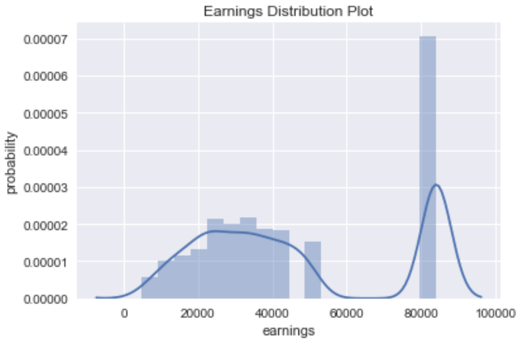
5.2 Results Analysis

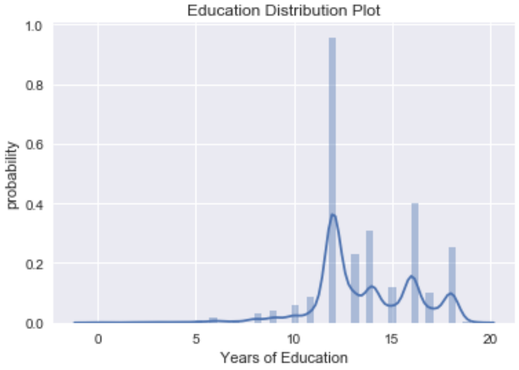
We split the full dataset into two sections, 70% of the dataset for the training set and 30% for the test set. We categorize our training set into two parts: full dataset and married couples only dataset, but the test set is based on the full dataset. To fit the machine learning models to our training data, we do the feature selection, using mutual information regression method. We use two different thresholds to select features, 0.005 and 0.015, and we calculate the test error with respect to those two MI thresholds. Specifically, MI with 0.005 selects 22 features for the dataset with all people included and 8 features for the dataset with only married people. On the other hand, MI with 0.015 selects 16 features for the dataset with all people included and 6 features for the dataset with only married couples.

In general, we find out that the difference in the RMSE on the results from two different kinds of datasets is small, which also suggests that surveying only married workers can be used to predict the earnings. This result is clearly shown in Figure 2.

We will focus our analysis on RMSE results, and we use the mean square error on test sets to calculate the baseline RMSE: 26923. This baseline RMSE was calculated by estimating the mean. This establishes how the model will perform before we apply any feature selection or machine learning algorithms. By seeing the performance of only estimating the mean, the performance of the other models is put into perspective.

Gamma GLM Results Since the dataset is sparse, as showed in the following two plots, we use Gamma generalized linear model to fit the dataset.





To apply the Gamma GLM, we first train the dataset with all workers included, encompassing the first 12,508 rows of the original dataset, and we predict the earnings of the people corresponding to the rest of the rows in the dataset. Then we evaluate the mean squared error on our predicted results, and find that the RMSE of the test dataset after feature selection with MI threshold of 0.005 is around 22,597, while the RMSE of the test dataset after feature selection with MI threshold of 0.015 is around 22,728.

Since we need to find the relationship between marriage status and earnings, we then train the dataset with only married couples included, which consists of the first 8148 rows of the dataset with only rows corresponding to the married people, and we then make predictions on the earnings of the married people corresponding to the rest of the rows. Similarly, we evaluated the mean squared error on our results, and it shows that the RMSE of the test dataset after feature selection with MI threshold of 0.005 is around 61427, while the RMSE of the test dataset after feature selection with MI threshold of 0.015 is around 34511.

From the RMSE we got, we find out that MI threshold of 0.015 gives closer errors, and the RMSE on the test set with only married people is 50% higher than the RMSE on the test set with all people included.

Overall, the Gamma GLM model performed worse than the baseline when using only the married couples subset of the data.

5.2.1 Lasso Regression Results

To apply the Lasso regression machine learning model, we set the normalize parameter to be true. Similar to the Gamma GLM, on one hand, we train the dataset with only married couples included, and we then make predictions on the earnings of the married people corresponding to the rest of the data. Similarly, we evaluated the mean squared error on our results, and it turns out that the RMSE of the test dataset after feature selection with MI threshold of 0.005 is around 21811, while the RMSE of the test dataset after feature selection with MI threshold of 0.015 is around 24800. On the other hand, when we train the dataset with all people included, the results to the evaluation of the mean squared error on our predicted results indicates that the RMSE of the test dataset after feature selection with MI threshold to be 0.005 is around 21800, while the RMSE of the test dataset after feature selection with MI threshold to be 0.015 is around 22034.

From the RMSE we got, we find out that the MI threshold of 0.005 gives lower errors compared with MI threshold of 0.015, which is consistent with the relative results we get from Gamma GLM with MI threshold of 0.005. However, the RMSE with different thresholds are closer when we fit the Lasso regression model. In addition, the four RMSE values derived from Lasso regression are all closer to each other. We can also discover that the RMSE with MI threshold of 0.005 are close whether we use Gamma GLM or Lasso regression. Therefore, when we do feature selection with MI threshold of 0.005, we can see that predicting earnings on the basis of only marriage status gives similar results when we predict earnings with consideration about all other aspects such as sex, races, ages, weights etc.

Overall, the Lasso model performed better than the baseline when using only the married couples subset of the data.

5.2.2 Elastic Net Results

To apply the Elastic Net machine learning model, we set the normalize parameter to be true and the L_1 ratio to be 0.5. Similar to the Gamma GLM and Lasso regression, we train both the dataset with only married couples included and the dataset with all people included. We then make predictions on the earnings of the married people corresponding to each test dataset. By using the metrcis of mean squared error, we report that the RMSE of the test dataset with only married couples after feature selection with MI threshold of 0.005 is around 32091, while the RMSE of the test dataset after feature selection with MI threshold of 0.015 is around 32093. When we train the dataset with all people included, it reveals that the RMSE of the test dataset after feature selection with MI threshold to be 0.005 is around 27275, and the RMSE of the test dataset after feature selection with MI threshold to be 0.015 is also around 27275.

There are several interesting viewpoints we can draw from our results by Elastic Net. If we compare the results from Elastic Net with the results from other two machine learning models, we can find out that when we do the feature selection with MI threshold of 0.005 and use the full dataset Elastic Net gives higher RMSE then Lasso. Additionally, when we use the dataset processed by feature selection of MI threshold of 0.015, the RMSE with Elastic Net is still higher than the RMSE with Lasso regression but lower than the RMSE with Gamma GLM if we use the data with only married couples. Moreover, the RMSE with Elastic Net on the same dataset indicates no difference between MI thresholds. Overall, the general results given by those three models are consistent with each other in the way that prediction on earnings by using only marriage status provides similar results when we consider all other factors.

Overall, the Elastic Net model performed slightly worse than the baseline when using only the married couples subset of the data.

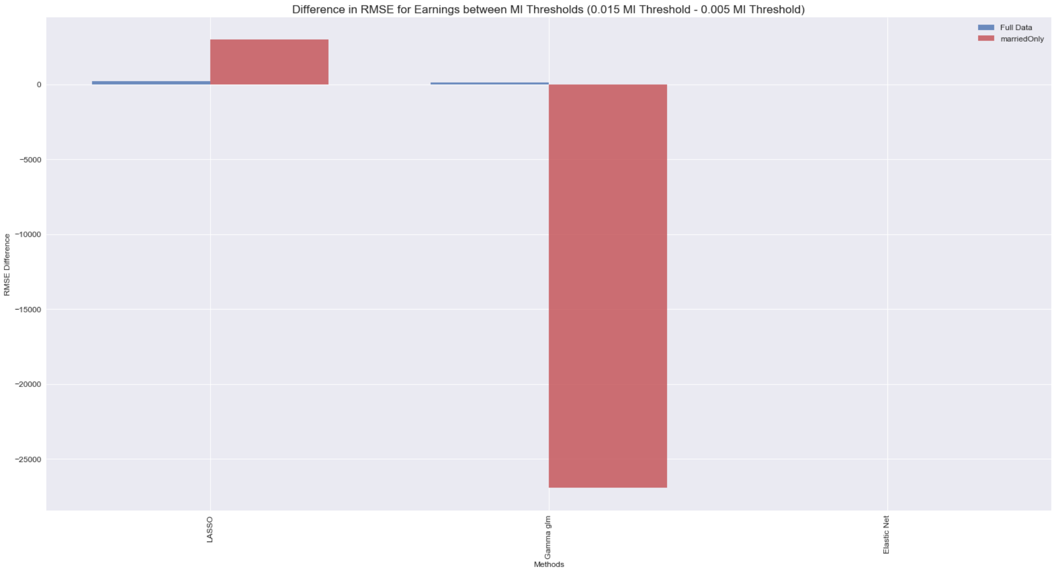


Figure 3: Difference between RMSE for the different models, on the two datasets, with different mutual information thresholds

5.2.3 R^2 results

The following graph shows that the difference in R^2 results with respect to different models is also small for the Lasso and Elastic Net models. This supports our hypothesis that we can predict worker's earning by only fitting a model from the data of marital status.

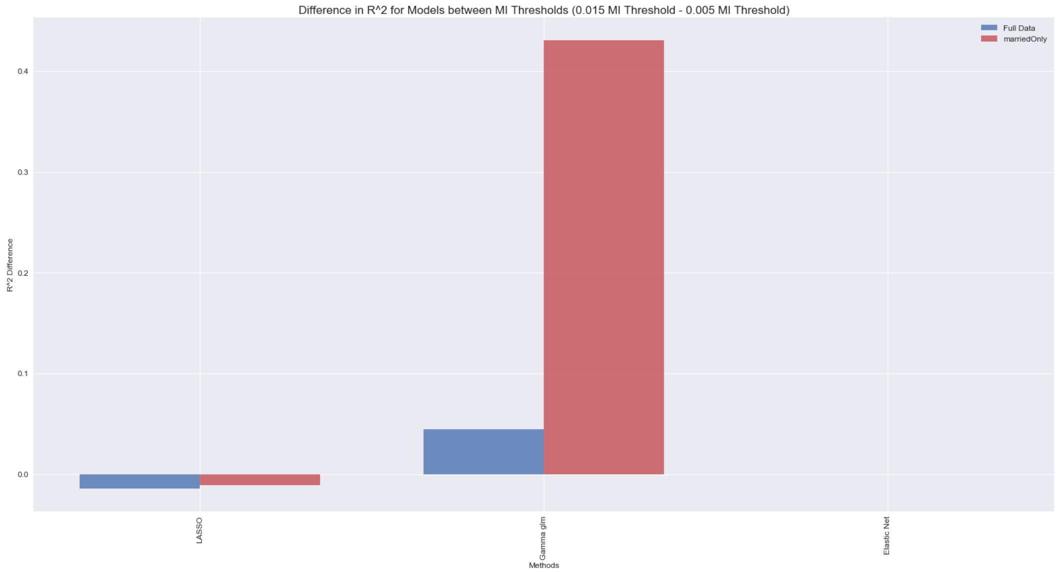


Figure 4: Difference between R^2 for the different models, on the two datasets, with different mutual information thresholds

6 Discussion and Conclusion

In this final assignment, we use a subset of the 1994 US National Health Interview Survey to be our dataset, which consists of the information collected from 17,870 workers. The questions in the survey identifies several characteristics which have predictive power on the earnings of a worker, such as age, class of worker, race, region, sex, weight etc. The main goal of this project is to find if we can have the same predictive accuracy building models just on the subset of married couples, and reproduce conclusions from well-known studies on this dataset using a smaller but still representative subset of the data. To achieve that, we use machine learning models to prove that using the dataset of only married couples can be an effective predictive tool on a worker's earning. We use Gamma generalized linear model, Lasso Regression model and Elastic Net model to fit the dataset. By evaluating and comparing the RMSE, we hypothesize we can see similar if not better RMSE using only married couples and more advanced models rather than just the baseline of predicting the mean. In fact, we find out that the Lasso model performs better than the baseline on just married couples. Specifically, we discover that the Lasso and Elastic Net model gives almost the same RMSE results with respect to the full dataset and the dataset with only married couples. Lastly, we were able to very closely reproduce the results of the original study. They found that on average a person on inch taller than average would be predicted to make $707.7 more than the average person which they attributed to taller people having higher IQ. Using the trimmed down dataset of just the married couples, feature selection of mutual information threshold at 0.015, the best performing model of Lasso regression reproduced on average a person on inch taller than average would be predicted to make $531.637 more than the average person.

Acknowledgments

We would like to thank Professor Barbara Engelhardt for her engaging lectures and the COS 424 TAs for their support and help during this project.

References

[1] The Financial Perks of Being Tall,
`https://www.theatlantic.com/business/archive/2015/05/the-financial-perks-of-being-tall/393518/`

[2] Personal Factors Associated with Leadership: A Survey of the Literature,
`https://www.tandfonline.com/doi/abs/10.1080/00223980.1948.9917362`

[3] Stature and Status: Height, Ability, and Labor Market Outcomes, `https://www.nber.org/papers/w12466.pdf`

[4] SKlearn API Reference, `https://scikit-learn.org/stable/modules/classes.html/`

[5] Why, How and When to apply Feature Selection, `https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adfabf2`

[6] The 19941995 National Health Interview Survey on Disability (NHIS-D): A Bibliography of 20 Years of Research, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4666019/`

[7] Ridge and Lasso Regression: A Complete Guide with Python Scikit-Learn, `https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b`

[8] What is the National Health Interview Survey?, `https://www.cdc.gov/nchs/nhis/participant.htm`