Analyzing the Impact of Financial Stability and Family Structure on Elder Health: A Machine Learning Approach Using the RAND HRS Longitudinal File 2020

Authors : Yuzhen Zhou, Zeyuan Pang

Department : McKelvey School of Engineering

Major : Engineering Data Analytics and Statistics

Advisor : Patricio S. La Rosa

Part 1. Executive Summary

The primary goal of our research is to analyze the impact of financial stability and family structure on the health outcomes. This investigation is crucial as nations globally grapple with the ramifications of an aging demographic, including heightened demands on healthcare systems, pension sustainability, and the overall welfare of the elderly population. Our study aims to provide evidence-based insights to guide policymaking in healthcare, social security, and family support initiatives, thereby facilitating more effective strategies to aid older adults.

For this analysis, we utilize the RAND HRS Longitudinal File 2020, which encompasses 15 waves of interview data collected over two decades. This comprehensive dataset is invaluable for research on health, family dynamics, retirement planning, employment history, and includes imputations for income, assets, and healthcare spending.

By examining the connections between economic status, family structure, and the health of the elderly, we aim to recommend targeted interventions that could improve life quality, reduce medical costs, and promote the sustainability of aging populations.

This project is of particular importance as it confronts a pressing challenge faced by East Asia: its rapidly aging population. This demographic transformation poses significant challenges for the social and economic progress of these countries, necessitating innovative approaches to ensure the well-being of the elderly and their families in a changing societal landscape.

Part 2. Data Exploration and Preprocessing

RAND HRS Longitudinal File 2020 is a huge dataset. It took weeks to read and understand the document and organize the data from different waves into a single dataset.

2.1 Data exploration

After reading the document, several variables of interest were selected for the initial analysis. The response variables are shown in Table 1:

VARIABLE CODE	CODE MEANING
SHLT	Self-rated Health Level
COGTOT	Cognitive Level
MSTOT	Mental Status Level

Table 1

The input variables are shown in Table 2:

VARIABLE CODE	CODE MEANING	VARIABLE CODE	CODE MEANING
BMI	Body Mass Index	**PRPCNT**	Number of Private Insurance Plans
INHPFN	Total Number of Helpers Ever Helped	***INHPE***	*Any employee of institution ever helped*
HHHRES	Number of People in Household	***HINPOV***	*Live in poverty*
HCHILD	Number of Children	***PENINC***	*Current receiving pension income*
LIVSIB	Number of Living Siblings	***HIGOV***	*Covered by government health insurance plan*
HAIRA	Individual Retirement Account Assets	***RETMON***	*Retirement Status*
HATOTB	Total Asset Amount	***SLFEMP***	*Self-Employment Status*
IEARN	Individual Income		
HITOT	Total Household Income		

Table 2

** Italic items on right side stand for binary categorical variables*

Table 3 in Appendix shows the statistical description of the features. Figure 1 and 2 in Appendix show the histogram of features.

According to the table and figures presented, it is observed that most ordinal predictive features exhibit a rightward skew. This trend is both normal and understandable when considering the distributions of earnings and assets. The imbalance distributions of categorical features also project the real society that a few people live in poverty, many elders are still working, and most are covered by government insurance plan.

2.2 Data Preprocessing

Group the data by categorical features

The whole dataset is separated into 32 different groups according to the different combinations of the seven binary categorical features. The five categorical features are then replaced by a single group feature.

Outliers

Isolate Forest algorithm is applied to each group separately to detect and remove 10% of the total points as outliers. This method ensures a focused approach towards outlier detection and removal, allowing for a cleaner and more accurate analysis of the data within each group.

Isolation Forest is an efficient and specialized algorithm for anomaly detection, leveraging a tree-based approach that excels in identifying outliers with minimal assumptions about data distribution in high dimensional situation.

After removing the outliers, 22 groups of data with less than 500 samples are dropped to ensure the performance of machine learning and statistical accuracy of research result. 10 groups remained.

Feature and Group Engineering

The column "INHPE" is dropped because its values are identical.

To ensure the selected groups provide meaningful insights, a Multivariate Analysis of Variance (MANOVA) was conducted across each pair of groups. The findings are presented in Table 4. Adopting a significance level of 0.1, the analysis revealed that the three pairs of groups without statistically significant differences are:

(0,0,1,0,1), (0,1,1,0,0) with p-value of 0.103
(0,0,1,0,1), (0,1,1,1,0) with p-value of 0.109
(0,1,1,0,0), (0,1,1,1,0) with p-value of 0.648

Subsequently, these groups were amalgamated. There are 8 groups in total.

The summary of cleaned groups after group merging is shown below in Table 5

HINPOV	PENINC	HIGOV	RETMON	SLFEMP	COUNT
0	0	0	0	0	11859
0	0	1	0	1	6138
0	1	1	0	0	
0	1	1	1	0	
0	0	1	0	0	4554
0	0	1	1	0	4067
0	0	1	1	1	2967
0	0	0	0	1	1969
0	1	1	1	1	1418
1	0	0	0	0	566

Table 5

** The groups in the box are combined*

To examine the interrelationship between each pair of target features, correlation matrices for the three designated target features were constructed for each group. Additionally, the Pearson correlation coefficient, along with its corresponding p-value for each pair of target features within each group, was calculated. The findings from these analyses are presented in Figures 3 and 4 of the Appendix. These results reveal that MSTOT and COGTOT exhibit a medium to strong correlation across the eight groups. However, the accompanying p-values are exceedingly small, denoting a statistically significant difference between MSTOT and COGTOT across all groups. Consequently, based on this statistical significance, the decision was made not to amalgamate the three target features.

To mitigate the potential adverse effects of disparate ranges across continuous features on machine learning models and performance metrics, standardization was employed for these features. This process ensures that each feature contributes equally to the model's prediction capability by normalizing their ranges. The statistical characteristics of the standardized features are comprehensively detailed in Table 6 of the Appendix, illustrating the distribution and scale uniformity post-standardization.

Part 3. Model

3.1 Baseline Models

In this phase of the analysis, four baseline models were evaluated: the MLP (Multi-Layer Perceptron) regressor, the KNN (K-Nearest Neighbors) regressor, the Random Forest regressor, and the Linear regressor, the latter of which utilizes features generated through the application of the K-means algorithm. The cleaned dataset was divided into training and testing subsets following a 3:1 ratio, facilitating the training of the baseline models and the subsequent evaluation of their performance. The outcomes of these performance evaluations are detailed in Table 7.

Model	Test R-Squared	MSE
MLP regressor with default parameters	-49868.69	265301.27
KNN regressor with k = 2	0.220	4.92
Linear regressor & features generated by K-means with clusters = 50	0.073	5.79
Random Forest regressor with default parameters	0.656	2.18

Table 7

The Random Forest regressors exhibit superior performance compared to other baseline models, as evidenced by achieving the highest R-squared values. This algorithm's distinctive advantage lies in its robustness and efficiency in managing complex, unstructured data across multiple dimensions.

Figures 5, 6, and 7 in the Appendix, which present the Kernel Density Estimate (KDE) plots of errors for each data point during the Random Forest regressor fitting, further support this observation. The KDE plots approximate normal distributions, indicating that the Random Forest regressors have a commendable capability in accurately predicting the target feature.

3.2 Random Forest Regressor on Grouped Data

Random Forest Regressor is trained separately for each group in the cleaned data. For each group, the data is separated into training data and testing data with the ratio 9:1, and a grid search with 10-fold cross validation was performed on the training data. After finding the best parameters, the performance of Random Forest Regressor is evaluated based on testing data. The best parameters for each group are shown in Table 8.

Three metrics are used to evaluate the performance—R-Squared, RMSE, and MAPE. R-Squared is best for understanding the proportion of variance explained by the model, RMSE is valuable for capturing the average error magnitude while penalizing large errors, and MAPE is useful for comparing the accuracy of models in terms of percentage errors, making it intuitive for expressing how large the errors are relative to actual values.

The performance of Random Forest Regressor for each group is shown in Table 9 of

Appendix. The test R-squared values for all groups are generally around 0.7, and are a little higher than their training R-squared values, showing high accuracy and great generalization of Random Forest Regressor on this dataset. RMSE is also relatively consistent between the train and test sets. However, the MAPE values tend to be higher on the test set, which could indicate that the models are not predicting as accurately on the test data when considered in terms of percentage errors.

3.3 Machine Learning Morphisms

ML1

ML1 is the unsupervised Isolation Forest Algorithm that maps each row of dataset to 0 or 1. It do not need any estimation on prior distribution, and it has no loss function. Isolation Forest use isolation score to decide whether a point is outlier.

$$ML_1 = (X \in \mathbb{R}^{18}, Y \in \{0,1\}, F(x, \Theta) = \text{Isolation Score Function}, P_\theta(\theta) = 1, L = \text{None})$$

ML2

ML2 is the process of Standardization. By subtracting the mean and dividing by the standard deviation for each value of each feature, the range of all features are transformed to same scale.

$$ML_2 = \left(X \in \mathbb{R}^{10}, Y \in \mathbb{R}^{10}, F(x, \Theta) = \frac{\mathrm{X} - \mu}{\sigma}, P_\theta(\theta) = 1, L = \text{None}\right)$$

ML3

ML3 is the Random Forest Regressor. It takes predictive features and predict response features by the average of every tree's output. It does not require a prior distribution but do need MSE as loss function to optimize the result.

$$ML_3 = \left(X \in \mathbb{R}^{15}, Y \in \mathbb{R}^3, F(x, \Theta) = \frac{1}{k}\sum_{i=1}^{k} T_i(x), P_\theta(\theta) = 1, L = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - F(x_i, \Theta)\right)^2\right)$$

Part 4. Next Steps

4.1 Model Accuracy

The model's predictive accuracy falls short of being ideal on some aspect. To enhance the performance of the Random Forest Regressor on the test data, more strategies will be explored and implemented with the objective of reducing the test MAPE.

4.2 Solving Heteroscedasticity

Heteroscedasticity is found on almost all the prediction of target features, methods like Box-Cox transformation or log transformation would be tried to deal with this problem. Alternative models like AdaBoost would be considered if transformation does not work well.

4.3 Diagnose and prescriptive analysis

We plan to delve into the reasons behind the specific characteristics and distributions of the data and weights observed within each group. This investigation aims to uncover underlying patterns or issues that may inform our recommendations for policies or business strategies. Additionally, by conducting a comparative analysis of the differences across groups, we intend to deepen our understanding of the dataset. This enhanced insight will not only shed light on the unique attributes of each group but also potentially reveal broader trends or anomalies that could inform more targeted and effective interventions.

Part 5. Answers to Questions from Peers

1. Why you decided to use these baseline models?

   We tried different classes of model to study which model fits this problem best. The baseline models are just random choices.

2. How will you adjust for the overfitting?

   Solved. See report for details.

3. Is there a time series component in the data?

   No.

4. I agree that the slide where your scores are presented indicates overfitting. Have you considered why this might be the case.

   The method of calculating training score was wrong. Fixed.

5. Some of your data seems to be skewed, how do you plan to deal with this in training and testing?

   See part 4.2.

6. Is there any better model will fit?

   ANN could be a better model, but it is too time consuming to train and tune, so we picked a model that's faster to train.

7. How are you going to tune the model, so that the model generalizes well in the real world.

   We used grid search with cross validation. The performance is shown on Table 9 of Appendix.

8. In your model you use SHLT as target feature, why you choose this feature?

   We are analyzing the influential factor of elder's health.

Part 6. Source Code

https://github.com/YuzhenZhou1327/ESE527_Project_HRS

Part 7. Appendix

Figure 1 Histogram of Numerical Ordinal Features

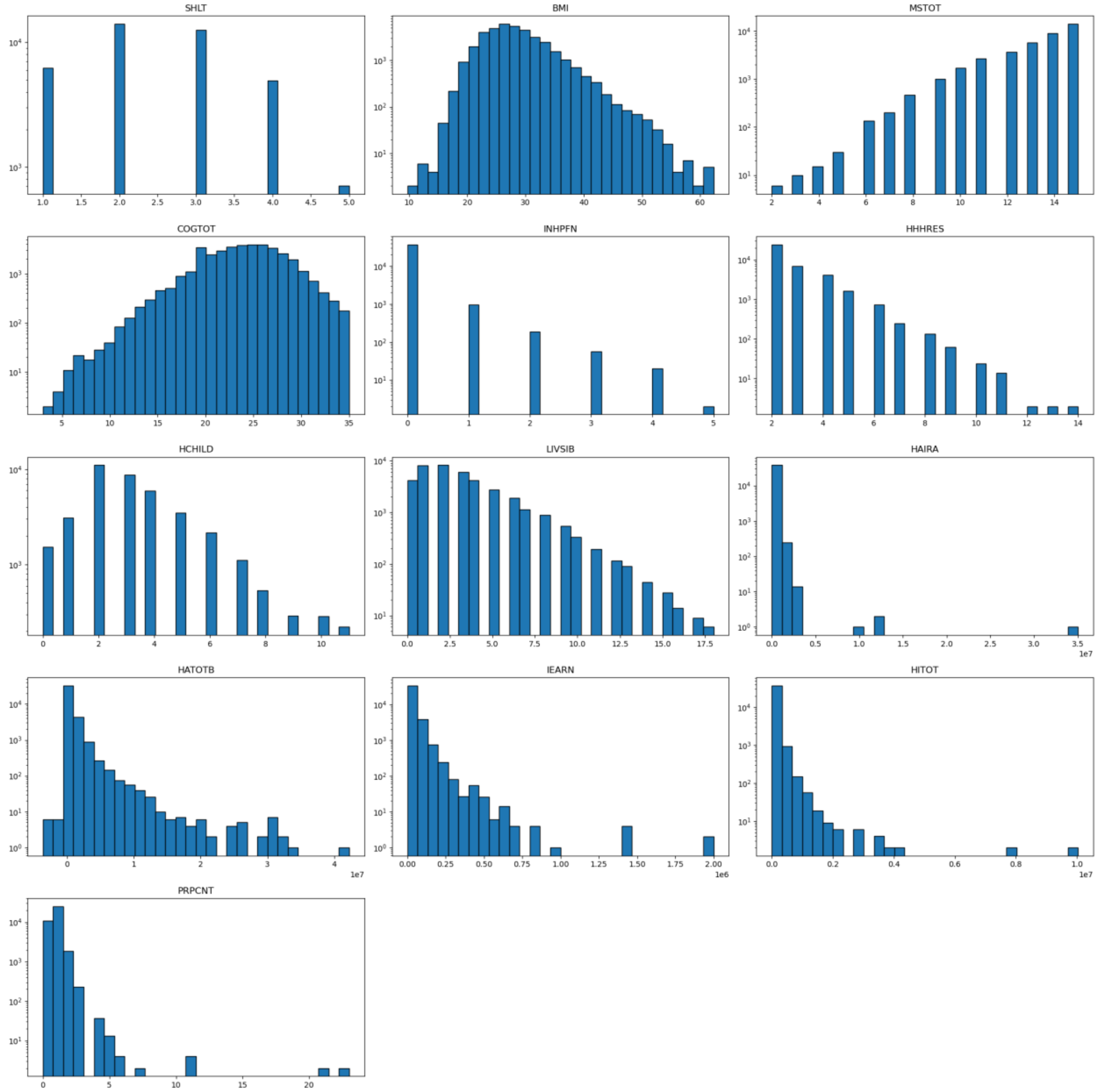


Figure 2 Histogram of Binary Categorical Features

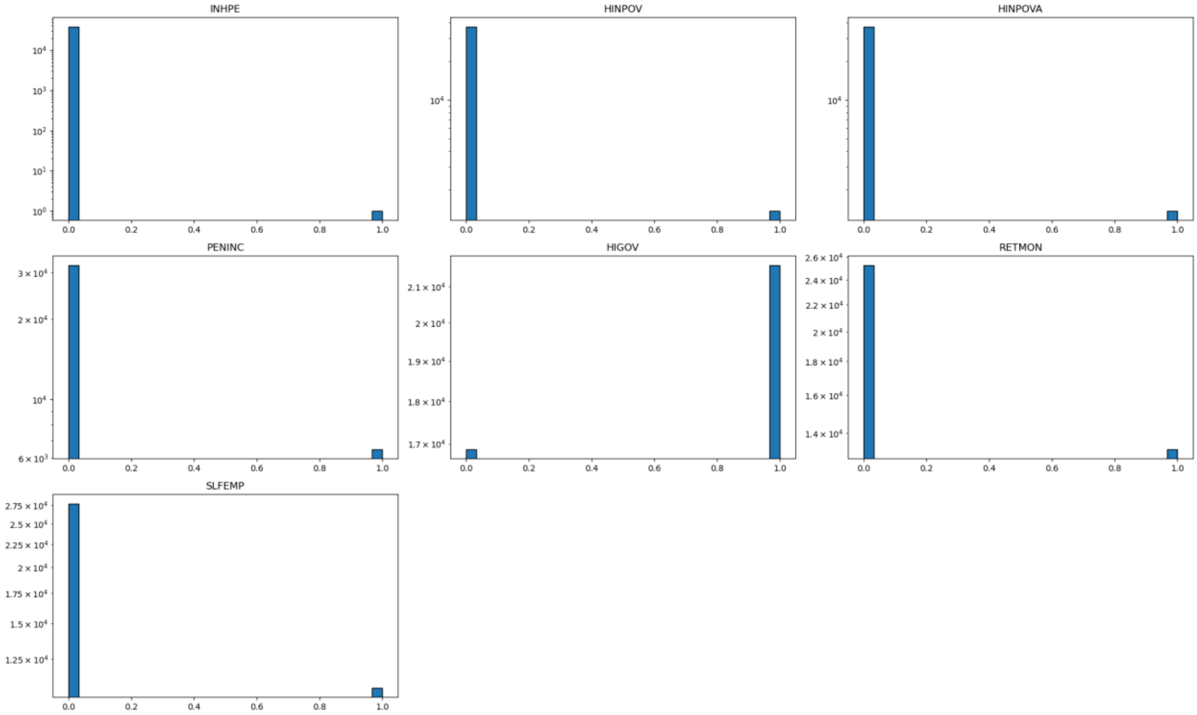
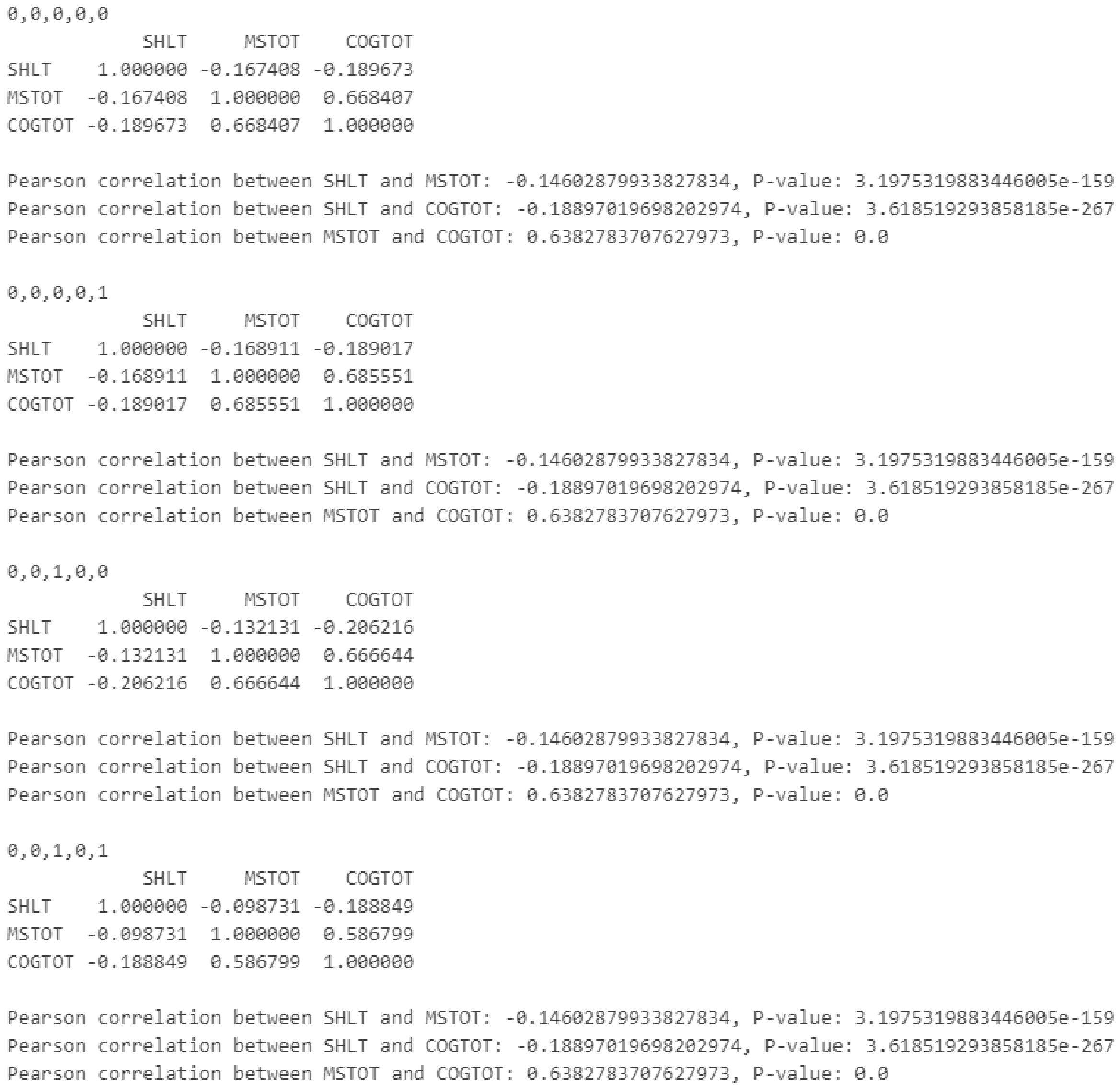
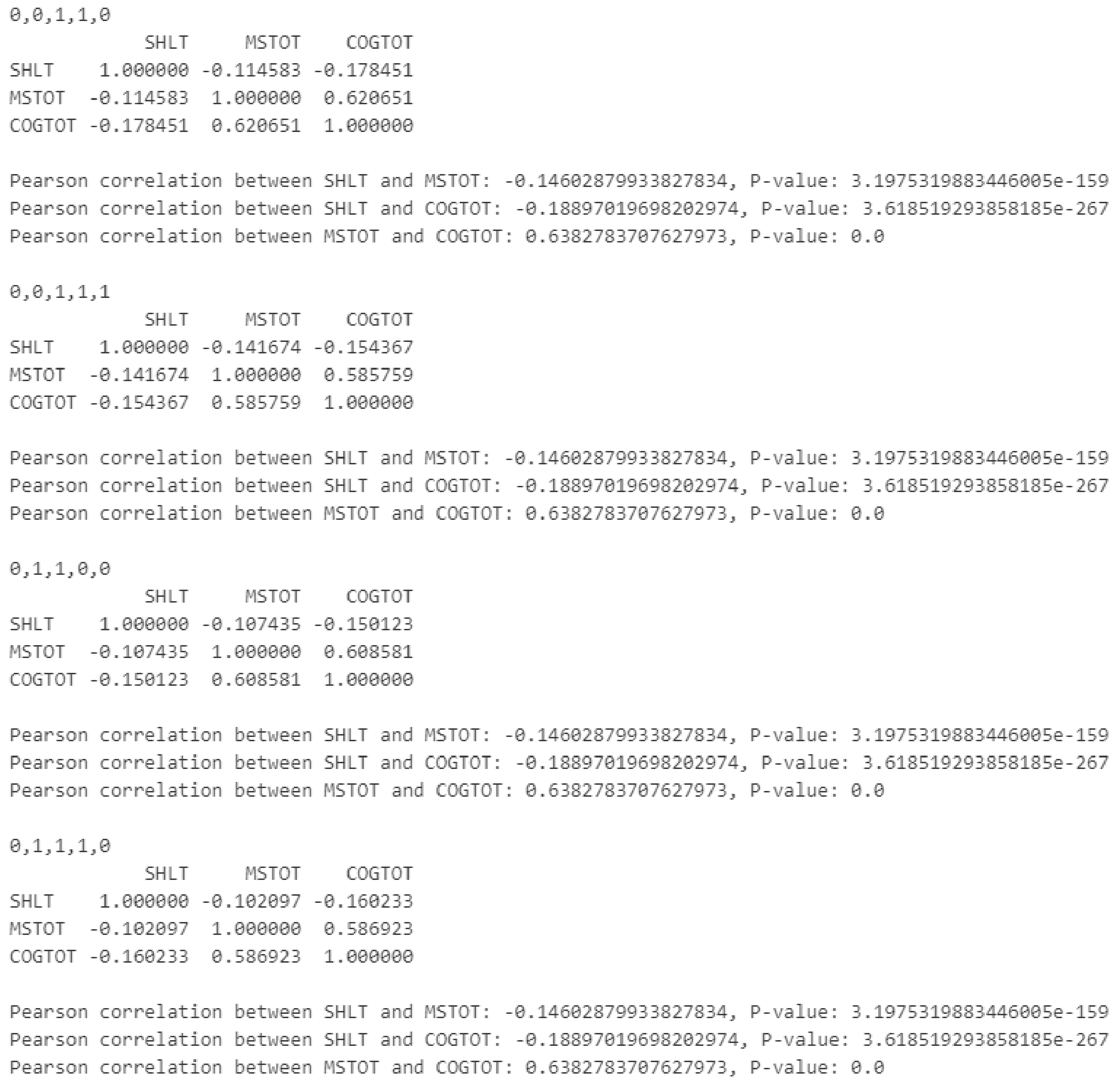


Figure 3

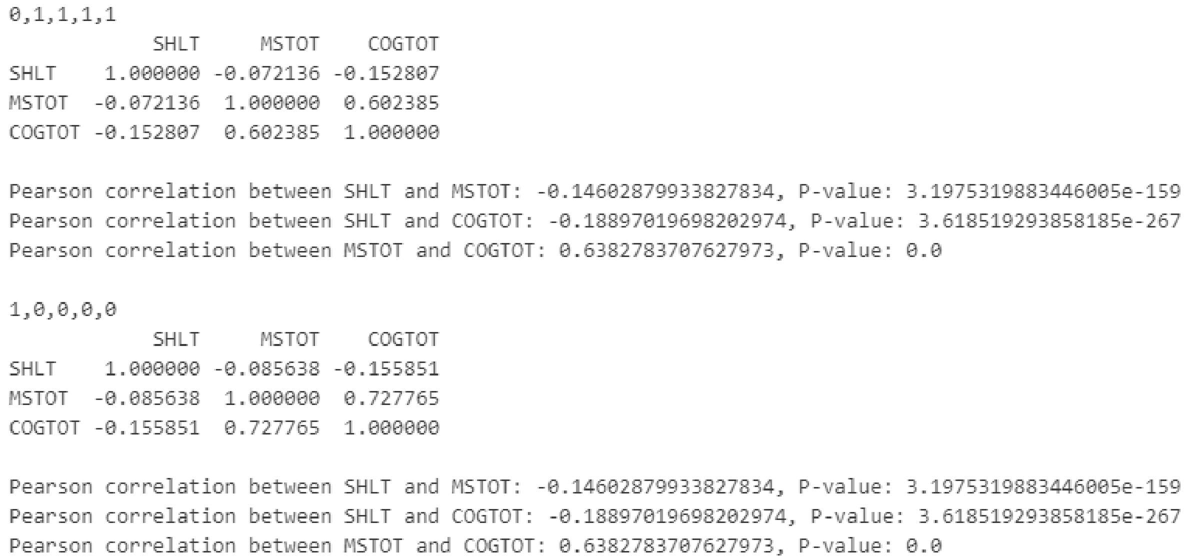


```
0,0,0,0,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.167408 -0.189673
MSTOT  -0.167408  1.000000  0.668407
COGTOT -0.189673  0.668407  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,0,0,0,1
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.168911 -0.189017
MSTOT  -0.168911  1.000000  0.685551
COGTOT -0.189017  0.685551  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,0,1,0,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.132131 -0.206216
MSTOT  -0.132131  1.000000  0.666644
COGTOT -0.206216  0.666644  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,0,1,0,1
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.098731 -0.188849
MSTOT  -0.098731  1.000000  0.586799
COGTOT -0.188849  0.586799  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0
```

Figure 4



```
0,0,1,1,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.114583 -0.178451
MSTOT  -0.114583  1.000000  0.620651
COGTOT -0.178451  0.620651  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,0,1,1,1
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.141674 -0.154367
MSTOT  -0.141674  1.000000  0.585759
COGTOT -0.154367  0.585759  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,1,1,0,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.107435 -0.150123
MSTOT  -0.107435  1.000000  0.608581
COGTOT -0.150123  0.608581  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,1,1,1,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.102097 -0.160233
MSTOT  -0.102097  1.000000  0.586923
COGTOT -0.160233  0.586923  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0
```



```
0,1,1,1,1
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.072136 -0.152807
MSTOT  -0.072136  1.000000  0.602385
COGTOT -0.152807  0.602385  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

1,0,0,0,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.085638 -0.155851
MSTOT  -0.085638  1.000000  0.727765
COGTOT -0.155851  0.727765  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0
```

Figure 5

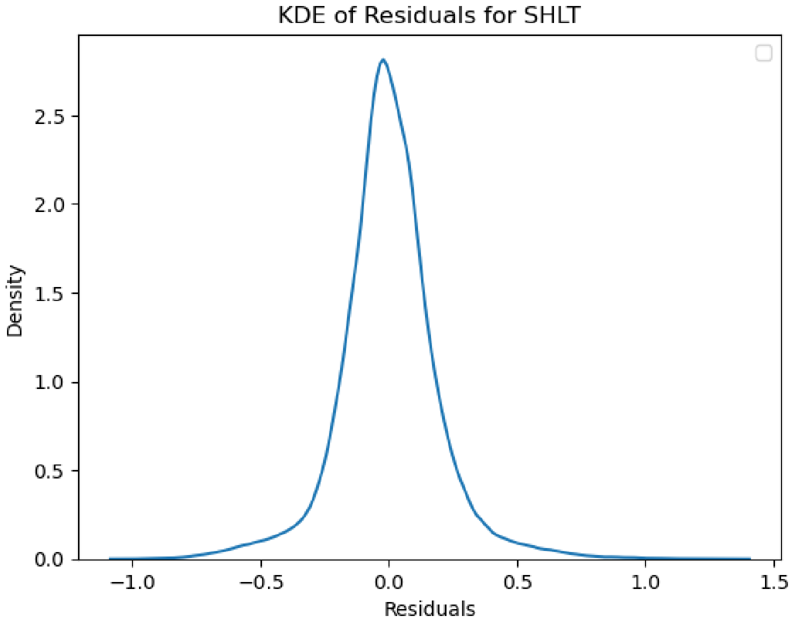


Figure 6

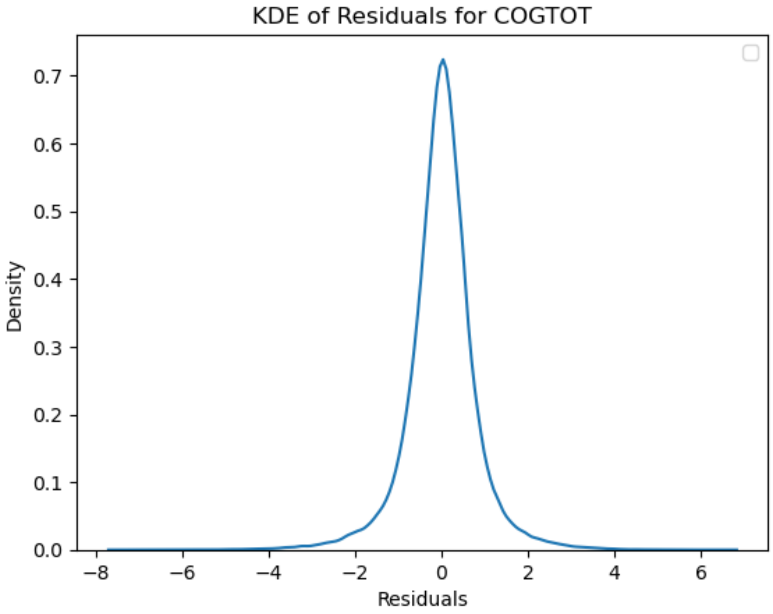


Figure 7

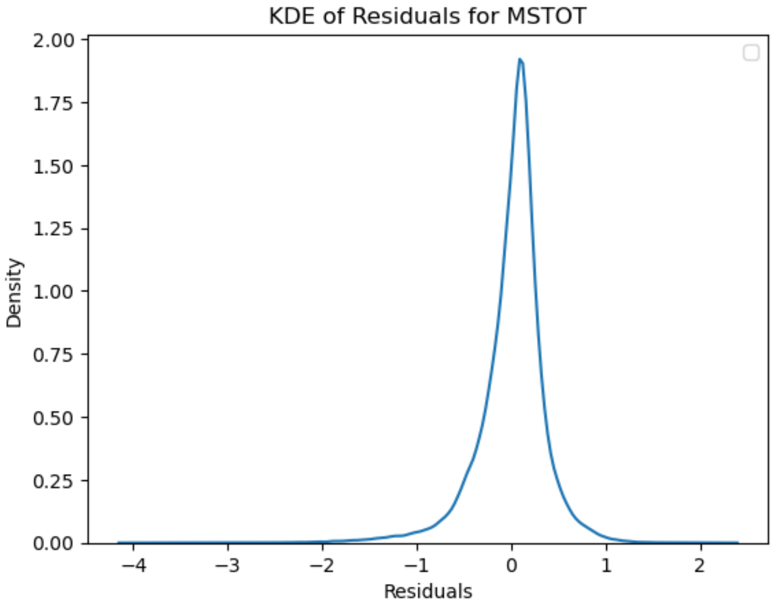

KDE of Residuals for MSTOT
Density
Residuals
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−4
−3
−2
−1
0
1
2

Table 3

	count	mean	std	min	25%	50%	75%	max
SHLT	38487	2.475251	0.970384	1	2	2	3	5
BMI	38487	28.25911	5.320587	9.7	24.6	27.4	31.1	62.3
MSTOT	38487	13.36553	1.874137	2	12	14	15	15
COGTOT	38487	23.94676	4.143787	3	21	24	27	35
INHPFN	38487	0.041287	0.255348	0	0	0	0	5
INHPE	38487	2.60E-05	0.005097	0	0	0	0	1
HHHRES	38487	2.678879	1.140705	2	2	2	3	14
HCHILD	38487	3.26661	1.933677	0	2	3	4	11
LIVSIB	38487	2.944813	2.451244	0	1	2	4	18
HINPOV	38487	0.035518	0.185089	0	0	0	0	1
HINPOVA	38487	0.035544	0.185154	0	0	0	0	1
HAIRA	38487	78742.64	283976.1	0	0	0	60000	35027000
HATOTB	38487	579882.2	1330807	-3624527	76000	228400	588500	42226312
IEARN	38487	31068.2	52357.43	0	0	15000	42000	2000000
HITOT	38487	102512.5	159141.1	0	41812	70880	119400	10036000
PENINC	38487	0.167953	0.373829	0	0	0	0	1
HIGOV	38487	0.561618	0.496195	0	0	1	1	1
PRPCNT	38487	0.786214	0.620732	0	0	1	1	23
SLFEMP	38487	0.280484	0.449242	0	0	0	1	1
RETMON	38487	0.343285	0.474812	0	0	0	1	1

Table 4

Group Pair	P-Value	Group Pair	P-Value
0,0,0,0,0 and 0,0,0,0,1	1.01E-12	0,0,1,0,0 and 0,1,1,1,1	1.95E-38
0,0,0,0,0 and 0,0,1,0,0	1.15E-36	0,0,1,0,0 and 1,0,0,0,0	1.27E-60
0,0,0,0,0 and 0,0,1,0,1	7.52E-73	0,0,1,0,1 and 0,0,1,1,0	9.75E-08
0,0,0,0,0 and 0,0,1,1,0	2.08E-81	0,0,1,0,1 and 0,0,1,1,1	7.05E-05
0,0,0,0,0 and 0,0,1,1,1	7.99E-71	0,0,1,0,1 and 0,1,1,0,0	0.103445
0,0,0,0,0 and 0,1,1,0,0	5.51E-20	0,0,1,0,1 and 0,1,1,1,0	0.108941
0,0,0,0,0 and 0,1,1,1,0	6.44E-61	0,0,1,0,1 and 0,1,1,1,1	0.005394
0,0,0,0,0 and 0,1,1,1,1	4.42E-52	0,0,1,0,1 and 1,0,0,0,0	2.12E-118
0,0,0,0,0 and 1,0,0,0,0	1.50E-109	0,0,1,1,0 and 0,0,1,1,1	0.014511
0,0,0,0,1 and 0,0,1,0,0	7.15E-39	0,0,1,1,0 and 0,1,1,0,0	6.76E-06
0,0,0,0,1 and 0,0,1,0,1	4.37E-44	0,0,1,1,0 and 0,1,1,1,0	6.66E-09
0,0,0,0,1 and 0,0,1,1,0	2.01E-57	0,0,1,1,0 and 0,1,1,1,1	2.80E-12
0,0,0,0,1 and 0,0,1,1,1	2.91E-53	0,0,1,1,0 and 1,0,0,0,0	3.35E-98
0,0,0,0,1 and 0,1,1,0,0	1.71E-19	0,0,1,1,1 and 0,1,1,0,0	0.000504
0,0,0,0,1 and 0,1,1,1,0	6.02E-42	0,0,1,1,1 and 0,1,1,1,0	0.000103
0,0,0,0,1 and 0,1,1,1,1	2.23E-42	0,0,1,1,1 and 0,1,1,1,1	1.33E-06
0,0,0,0,1 and 1,0,0,0,0	2.16E-101	0,0,1,1,1 and 1,0,0,0,0	9.56E-104
0,0,1,0,0 and 0,0,1,0,1	9.11E-41	0,1,1,0,0 and 0,1,1,1,0	0.647902
0,0,1,0,0 and 0,0,1,1,0	3.19E-23	0,1,1,0,0 and 0,1,1,1,1	0.069745
0,0,1,0,0 and 0,0,1,1,1	1.78E-29	0,1,1,0,0 and 1,0,0,0,0	3.94E-79
0,0,1,0,0 and 0,1,1,0,0	7.88E-17	0,1,1,1,0 and 0,1,1,1,1	0.021763
0,0,1,0,0 and 0,1,1,1,0	4.71E-39	0,1,1,1,0 and 1,0,0,0,0	7.94E-130
0,1,1,1,1 and 1,0,0,0,0	8.52E-117		

Table 6

	count	mean	std	min	25%	50%	75%	max
BMI	33538	1.41E-14	1.000015	-3.56846	-0.69774	-0.15828	0.535316	6.353814
INHPFN	33538	5.82E-15	1.000015	-0.12617	-0.12617	-0.12617	-0.12617	32.9837
HHHRES	33538	-4.92E-14	1.000015	-0.60029	-0.60029	-0.60029	0.366366	9.066287
HCHILD	33538	7.75E-15	1.000015	-1.72947	-0.65153	-0.11256	0.426414	4.199211
LIVSIB	33538	2.06E-15	1.000015	-1.21745	-0.79643	-0.3754	0.466642	6.360959
HAIRA	33538	-1.18E-14	1.000015	-0.41891	-0.41891	-0.41891	-0.05202	14.42363
HATOTB	33538	-1.00E-15	1.000015	-3.4569	-0.49971	-0.31643	0.089569	16.44927
IEARN	33538	-4.36E-14	1.000015	-0.77812	-0.77812	-0.36534	0.322617	10.22924
HITOT	33538	-1.34E-16	1.000015	-1.13878	-0.60941	-0.26853	0.285929	15.36991
PRPCNT	33538	8.69E-17	1.000015	-1.34082	-1.34082	0.358474	0.358474	37.74293
SHLT	33538	-4.80E-15	1.000015	-1.54866	-0.48538	-0.48538	0.577894	2.704447
MSTOT	33538	-2.33E-15	1.000015	-5.53478	-0.27574	0.308599	0.892937	0.892937
COGTOT	33538	-1.87E-16	1.000015	-4.66151	-0.54534	-0.03082	0.740961	2.799045

Table 8

Group	*Best Parameters*
0,0,0,0,0	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
0,0,0,0,1	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
0,0,1,0,0	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
0,0,1,0,1 *+ 0,1,1,0,0* *+ 0,1,1,1,0*	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
0,0,1,1,0	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
0,0,1,1,1	{'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
0,1,1,1,1	{'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
1,0,0,0,0	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}

Table 9

Group	*Train R^2*	*Test R^2*	*Train RMSE*	*Test RMSE*	*Train MAPE*	*Test MAPE*
0,0,0,0,0	0.68934439	0.774382961	0.182254221	0.474177552	0.233304608	0.650197365
0,0,0,0,1	0.647113296	0.776677687	0.186345084	0.447670429	0.245029458	0.623240062
0,0,1,0,0	0.709024677	0.763115367	0.18920196	0.525944047	0.228331541	0.64904841
0,0,1,0,1 *+ 0,1,1,0,0* *+ 0,1,1,1,0*	0.68802007	0.766886165	0.173116743	0.45229543	0.220104437	0.578865638
0,0,1,1,0	0.677786012	0.756692127	0.179725473	0.491717105	0.230380117	0.61393154
0,0,1,1,1	0.685561318	0.734663996	0.183206614	0.482640112	0.227754199	0.71461593
0,1,1,1,1	0.661402102	0.753137565	0.178808815	0.417472335	0.248393304	0.652643497
1,0,0,0,0	0.488510025	0.519843747	0.287365396	0.859031587	0.389015753	0.752216606