STAT 451
Analyzing OCD: A Data-Driven Approach

12/7/ 2023

Group 17: Brendon Chen Cheng-HSU NEE Faye Jiang Xin Lin Ziming Xu

Dataset analysis

- **1500** patients list
- **Quantitative standard:** o_score and c_score
- **Chosen variables**:
- Sex, gender, duration, and family history

	age	sex	duration	f_h	o_score	c_score
0	32	Female	203	No	17	10
1	69	Male	180	Yes	21	25
2	57	Male	173	No	3	4
3	27	Female	126	Yes	14	28
4	56	Female	168	Yes	39	18
5	32	Female	46	No	26	11
6	38	Female	110	No	12	16
7	57	Male	197	No	31	4
8	36	Male	84	No	37	24
9	72	Female	47	Yes	28	36

- Dataset Source: Kaggle OCD Patient Dataset.

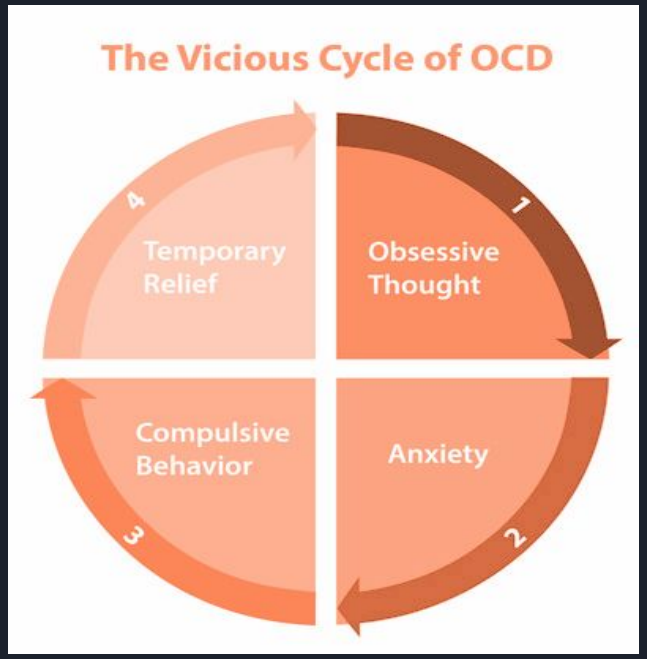
Why Obsessive-Compulsive Disorder (OCD)?

- **OCD:** Obsessive-Compulsive Disorder, a mental health disorder characterized by persistent
- **Focus**: Association between demographic and clinical factors with the onset and severity of OCD
- **Motivation**: Our experiences with OCD; challenges in managing diverse symptoms and treatment responses.
- **Goal:** To identify predictors of symptom severity, informing better treatment strategies.

Our OCD Analytical Methods

- **Predictive Variables:** Age, gender, duration, marital status, family history of OCD, Y-BOCS scores
- **Objective:** Predicting Y-BOCS score rankings to determine OCD severity.
- **Data Encoding:**

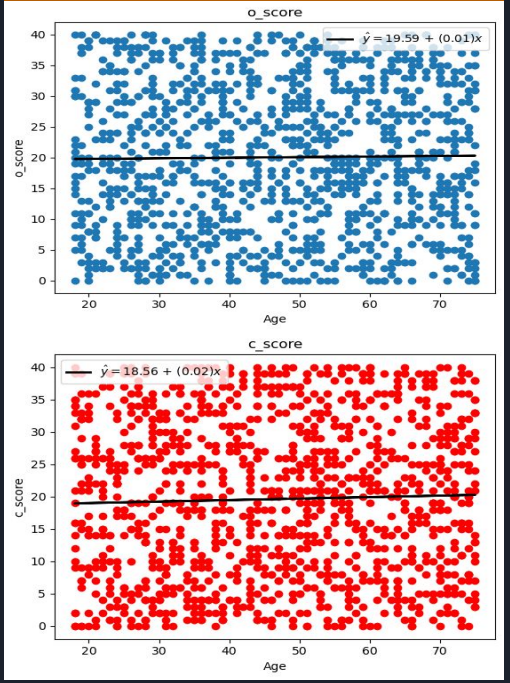
  Application of one-hot coding to reevaluate variables for model compatibility.



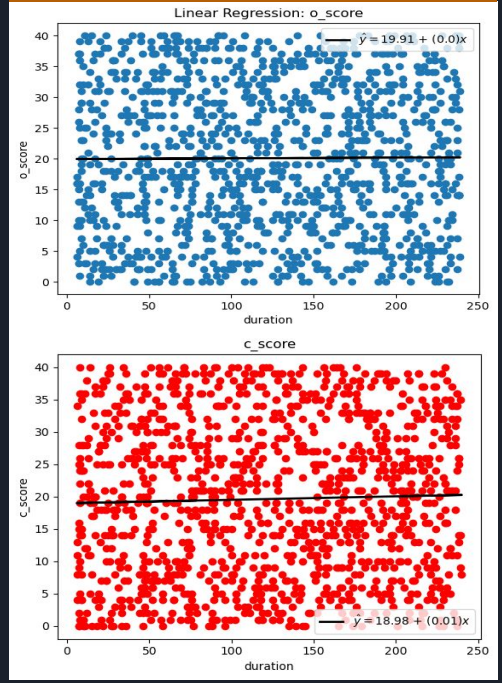
- **Model Building:** Regression methods
- **Dataset Allocation:**
- 80% for Model Training
- 10% for Model Testing
- 10% for Model Validation

Insights from Analysis: Key Findings & Impac

Age vs OCD level



Duration vs OCD level



As pre-programming session: we plotted linear relationship between age, duration and o_score, c_score to evaluate the correlation

Outcome: **Almost no correlation at all**

One-hot encoding data processing

After finishing the logistic regression analysis of age, duration and OCD level, we did some on-hot encoding data process, making family history, sex to 0 and 1.

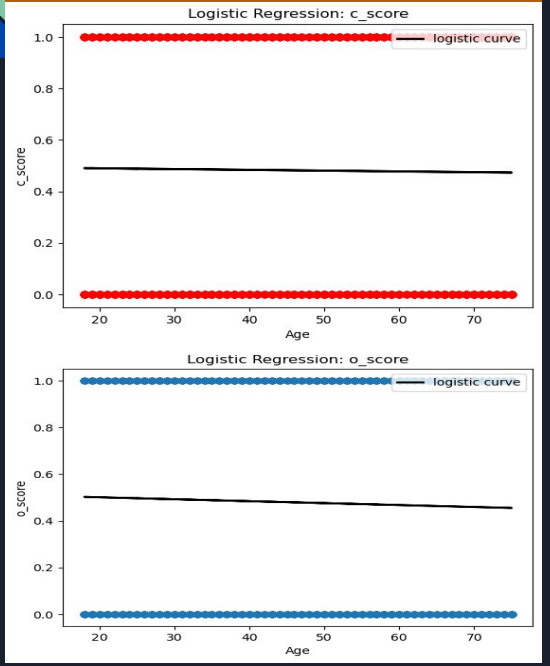
And dividing o_score and c_score into low level and high level base on ranking 0-20 and 20-40

	age	sex	duration	f_h	o_score	c_score
0	32	Female	203	No	17	10
1	69	Male	180	Yes	21	25
2	57	Male	173	No	3	4
3	27	Female	126	Yes	14	28
4	56	Female	168	Yes	39	18
5	32	Female	46	No	26	11
6	38	Female	110	No	12	16
7	57	Male	197	No	31	4
8	36	Male	84	No	37	24
9	72	Female	47	Yes	28	36

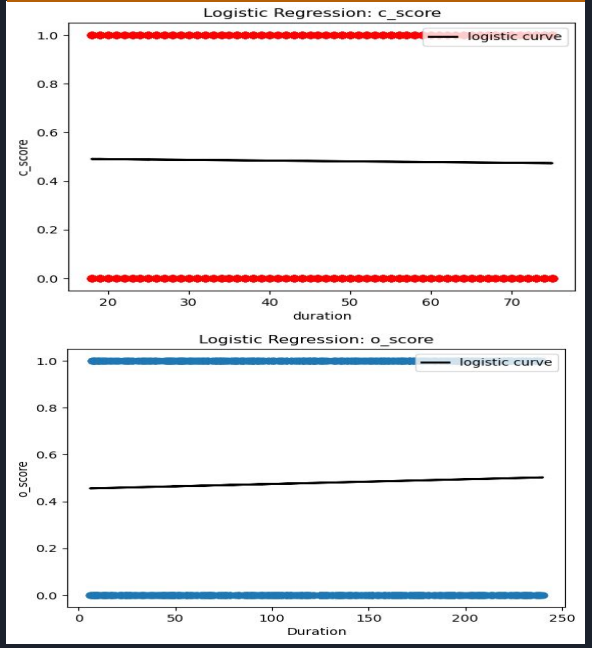
	age	duration	f_h	male	o_score	c_score
0	32	203	0	0	0	0
1	69	180	1	1	1	1
2	57	173	0	1	0	0
3	27	126	1	0	0	1
4	56	168	1	0	1	0
...	...	...	...	...	...	...
1495	38	53	0	1	1	1
1496	19	160	1	0	1	0
1497	40	100	1	1	0	0
1498	37	210	1	0	0	0
1499	18	91	1	1	1	1

Insights from Analysis: Key Findings & Impac

Age vs OCD level



Duration vs OCD level



Outcome: After putting Logistic Regression as analysis model, they clearly shows that there are very low correlation between age and OCD level, o_score is a little bit more relate to age compare to c_score, which may shows mental factor is a bit more significant than behavior.

Insights from Analysis: Key Findings & Impac

Random Forest Analysis-4 variables

Statistical Methods: We use Random Forest method to calculate the accuracy of Age, Family History, Duration, and Gender.

Result: After running code, the accuracy of Age, Family History, Duration, Male is 0.54, 0.53, 0.52, and 0.53.

Low accuracy analysis: As the 4 variables are all have very similar accuracy through the same model, it means that they have all low correlation with OCD level.

Insights from Analysis: Key Findings & Impac

Accuracy comparison— 4 Variables and OCD level

	Classifer	accuracy of o_score	accuracy of c_score
0	KNN	0.53	0.52
1	RandomForest	0.55	0.52
2	SVM	0.54	0.51
3	LogisticRegression	0.54	0.51

As for the accuracy between 5 variables and o_score and c_score (OCD level), 3 models are used to test the accuracy

Result: All accuracies are around 0.50, which means all models correctly predicts half of the instances.

Why accuracy is low?

Statistical factor

1. Data distribution
2. Sample size
3. Nonlinear relationship

Social factor

1.Working environment

2.Education level

3.Financial situation

Impact and Conclusion: Implications and Future Directions

- Our goal is to predict the OCD level (o_score and c_score) through 4 variables, and we take the database to build the model and test the accuracy of them.
- Linear regression : No correlation (Age/Duration & OCD)
- Logistic regression: No correlation (Age/Duration & OCD)
- Random Forest: No correlation (4 variables & OCD)

Future Directions: Expand the Set of Variables, Focus on Subtypes of OCD,Refine Measurement Techniques, and Longitudinal Studies.


Q
&
A