https://doi.org/10.29013/YSJ-22-1.2-23-29

Qinglan Luo,
11th-grade student at Rutgers Prep

AUTISM AMONG CHILDREN IN2019 NATIONAL SURVEY OF CHILDREN'S HEALTH

Abstract. Autism is a broad range of complex developmental disabilities in social communication and interaction coincided with restricted, repetitive behaviors. According to research from the National Survey of Children's Health (NSCH), approximately 28.5% of children aged from 0 to 17 have autism, and the percentage has been increasing in recent years. To control the increasing trend of autism, this study aims to examine the predictors of autism and build a predictive model for autism using the logistic regression model.

We used the 2019 NSCH dataset in this report. After feature normalization, we built a logistic regression model to predict whether a child is likely to develop autism. The predictive model is further validated by an overall evaluation of the model and has achieved an AUROC score of 0.68. By investigating the correlation among variables and the logistic regression coefficients, we found the dependent variable is most positively correlated with the financial situation of the household and most negatively correlated with the gender of the child. The results imply that older children are more likely to develop autism ($p = 0.0086$, OR = 1.093), and children in the family whose income is not able to cover basic living are more likely to develop autism ($p < 0.001$, OR = 1.77). Besides, children do not have a low birth weight ($p = 0.033$, OR = 0.47) and female children are less likely to develop autism ($p < 0.001$, OR = 0.82).

Keywords: autism, financial situation, predictive model, machine learning, logistic regression.

1. Introduction

"It is like getting an internet error when you are playing a game. The little avatar on the screen is disconnected from the outside world. He was just trapped there, unable to move or get out." This is a heartbreakingly true account given by a child's mother with autism. Before the age of two and a half, Liu had always been a bright and lively child. While his growth trajectory was no different from that of a normal child, Liu's symptoms were suddenly detected by his parents – he became muted just overnight. Then, he was diagnosed with autism. The abrupt diagnosis shocked Liu and his family, signaling a cascade of questions and inviting an ever-evolving emotional burden with it. Luckily for Liu, he was not the first autistic child to be clinically assessed in the United States. Nowadays, there have been more analogous cases in the United States. In order to control the growth of such tragedies, scientists have done plenty of research on the causes of ASD. It has been clear now what genetic factors play an essential role in leading to this disease. On the other hand, the environmental factors are still implicitly demonstrated.

Autism, also known as the autism spectrum disorder (ASD), is still an undeveloped area of understanding. It is a broad range of complex developmental disabilities in social communication and interaction coincided with restricted, repetitive behaviors. According to DSM-5, social deficits are defined as "deficits in social-emotional reciprocity, deficits in nonverbal communicative behaviors used for social interaction, and deficits in developing, maintaining, and understand relationships". (CDC, 2020) Repetitive behaviors include "stereotyped

or repetitive motor movements, use of objects, or speech, insistence on sameness, inflexible adherence to routines, or ritualized patterns of verbal or nonverbal behavior, highly restricted, fixated interests that are abnormal in intensity or focus, hyper- or hyporeactivity to sensory input or unusual interest in sensory aspects of the environment." (CDC, 2020) According to research from the National Survey of Children's Health (NSCH), there were approximately 28.5% of children aged from 0 to 17 having autism (CDC, 2021), increasing by 9.7 in percentage compared to the number of autisms in 2012 (188 per 1000 children). Indeed, there has been a consistent increase in the number of diagnosed autisms across various data sources. Although people are still not sure to what extent the changes in the clinical definitions of ASD and more people being consciously involved in ASD diagnosis has stimulated the growth, it is dangerous to assume that the real number of ASD has been stable and in control in the recent years.

The statistics have shown that this disorder transcends race, gender, and SES and challenges teenagers in different degrees. According to DSM-5, there are three levels of autism. Level 1 refers to any diagnosed children having trouble initiating social interactions and requiring support; level 2 includes children whose social interactions are limited to narrow special interests and have frequent restricted or repetitive behaviors; level 3 refers to children with severe deficits in verbal and nonverbal social communication skills, requiring substantial support from others (CDC, 2020).

As the number of people with autism rises each year, more families are suffering from the disease, and many talented children cannot utilize their gifts because of autism. Although we are still unclear about the treatment yet, we can go into the causative factors of this disorder and effectively prevent its increase after fully understanding both the genetic and environmental factors. This study serves this end by aiming to examine the predictors of autism and building a predictive model for autism using the logistic regression model.

2. Data and Methods:

2.1 Data

This report uses data from the National Survey of Children's Health (NSCH) in 2019, which is a population-based survey established by the Health Resources and Services Administration (HRSA) Maternal and Child Health Bureau (MCHB) to monitor the prevalence of the children health condition in the United States and to evaluate their access to quality health care (NSCH – Questionnaires 2019). The whole survey mainly encapsulates family composition, race/ethnicity, income, type of health insurance, and a variety of other important demographic and health status characteristics related questions. The data is collected by telephone surveys to random households across the United States. The 2019 NSCH dataset is used in this report. Before the data-cleaning process, the NHIS dataset has 67,625 valid observations.

The table below shows all the variables that have been chosen in this report to examine the relationship between independent variables and the dependent variable:

Table 1.– Variables used for analysis

Item Code	Question
1	*2*
HHCOUNT	How many people are living or staying at this address?
A1_SEX	What is your sex?
A1_BORN	Where were you born?
A1_GRADE	What is the highest grade or level of school you have completed?
A1_MARITAL	What is your marital status?
A1_AGE	What is your age?

1	*2*
A1_PHYSHEALTH	In general, how is your physical health?
A1_MENTHEALTH	In general, how is your mental or emotional health?
SC_AGE_YEARS	Is this child 3 years old or older?
SC_RACE_R	What is this child's race/ethnicity?
SC_HISPANIC_R	Is this child Hispanic?
BIRTHWT_L	Is this child born with low birth weight?
ACE1	How often has it been very hard to cover the basics, like food or housing, on your family's income?
ACE3	To the best of your knowledge, has this child EVER experienced: parent or guardian divorced?
ACE4	To the best of your knowledge, has this child EVER experienced: parent or guardian died?
ACE5	To the best of your knowledge, has this child EVER experienced: parent or guardian served time in jail?
ACE6	To the best of your knowledge, has this child EVER experienced: saw or heard parents or adults slap, hit, kick punch one another in the home?
ACE7	To the best of your knowledge, has this child EVER experienced: was a victim of violence or witnessed violence in his or her neighborhood?
ACE8	To the best of your knowledge, has this child EVER experienced: lived with anyone who was mentally ill, suicidal, or severely depressed?
ACE9	To the best of your knowledge, has this child EVER experienced: lived with anyone who had a problem with alcohol or drugs?
ACE10	To the best of your knowledge, has this child EVER experienced: treated or judged unfairly because of his or her race or ethnic group?
AUTISMMED	Is this child CURRENTLY taking medication for Autism, ASD, Asperger's Disorder or PDD?

This report uses the variable "AUTISMMED" as the dependent variable. Responses to the question "AUTISMMED" is dichotomous, meaning that the respondents either answer "yes", indicating that the child needs treatment for autism, or "no", indicating that the child does not need such treatment or does not have such behavior.

2.2 Statistical Models

2.2.1 Pre-processing

The data set is pre-processed in this step to improve both the training speed and accuracy. Since there is inevitably missing data, imputation is required to better analyze and extrapolate the missing data. As most machine learning algorithms are not able to deal with missing values, we replaced the missing values with the mean value of the entire feature column (mean value imputation). Some machine learning algorithms, such as artificial neural networks, require a specific technique called feature scaling which transforms different features into comparable scales for better training speed and accuracy. In this report, we will use the min-max scalar for this purpose. For each feature, its minimum and maximum value are first computed as X_{min} and X_{max}. Then each data point X with respect to that feature is replaced by X_{sc} calculated as:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Using this formula, X_{sc} is the ultimate value that is going to be analyzed in this report.

2.2.2 Logistic Regression Model

A logistic regression model refers to a model that is used to predict the probability of an incidence to happen. The probability varies from 0 to 1, with zero indicating not likely to happen and one indicating very likely to happen. Instead of a linear relationship, the logistic regression model fits an "S" shape which can be expressed using the formula below:

$$\ln\left(\frac{y}{y\text{-}1}\right) = a_0 + a_1x_1 + a_2x_2 + L + a_nx_n$$

In the above equation, a_0 is the intercept, xn represents the independent variables, and a_1 to a_n, are their corresponding coefficients (weights). In this report, our goal is to find the coefficients $(a_0, \ldots, a_n)$ minimizing the sum of squared errors (SSE) so that our predicted values will deviate the least from the real values.

2.3 Model Validation

Consider a two-class prediction problem, where the outcomes are labeled either as positive or negative. There are four possible outcomes from a binary classifier. If the outcome from a prediction is positive and the actual value is also positive, then it is called a true positive (TP); however, if the actual value is negative then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are negative, and a false negative (FN) is when the prediction outcome is negative while the actual value is positive. In this way, the true positive rate (TPR) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN}$$

And the false positive rate (FPR) can be calculated as:

$$FPR = \frac{FP}{TN + FP}$$

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied (Google, 2020). The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The best possible prediction method would yield a point in the upper left corner of the ROC space. A random guess would give a point along a diagonal line from the left bottom to the top right corners. Points above the diagonal represent better than random classification results, while points below the line represent worse than random results. In general, ROC analysis is one tool to select possibly optimal models and to discard suboptimal ones independently from the class distribution. Sometimes, it might be hard to identify which algorithm performs better by directly looking at ROC curves. Area Under Curve (AUC) overcomes this drawback by finding the area under the ROC curve, making it easier to find the optimal model.

3. Results

3.1 Chorogram

A chorogram is a graphical representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes by using visual thinning and correlation-based variable ordering. Moreover, the cells of the matrix can be shaded or colored to show the correlation value. The positive correlations are shown in blue, while the negative correlations are shown in red; the darker the hue, the greater the magnitude of the correlation.

According to the chorogram above, children's chance for developing autism has the strongest positive correlation with the variable "ACE1", whether the household is able to cover basics on family's income, and has the strongest negative relationship with "SC_SEX", which is the gender of the child.

3.2 Logistic Regression Results

The results of logistic regression analysis of children ever need medical treatment for emotional and behavioral disorder are listed in the figure below.

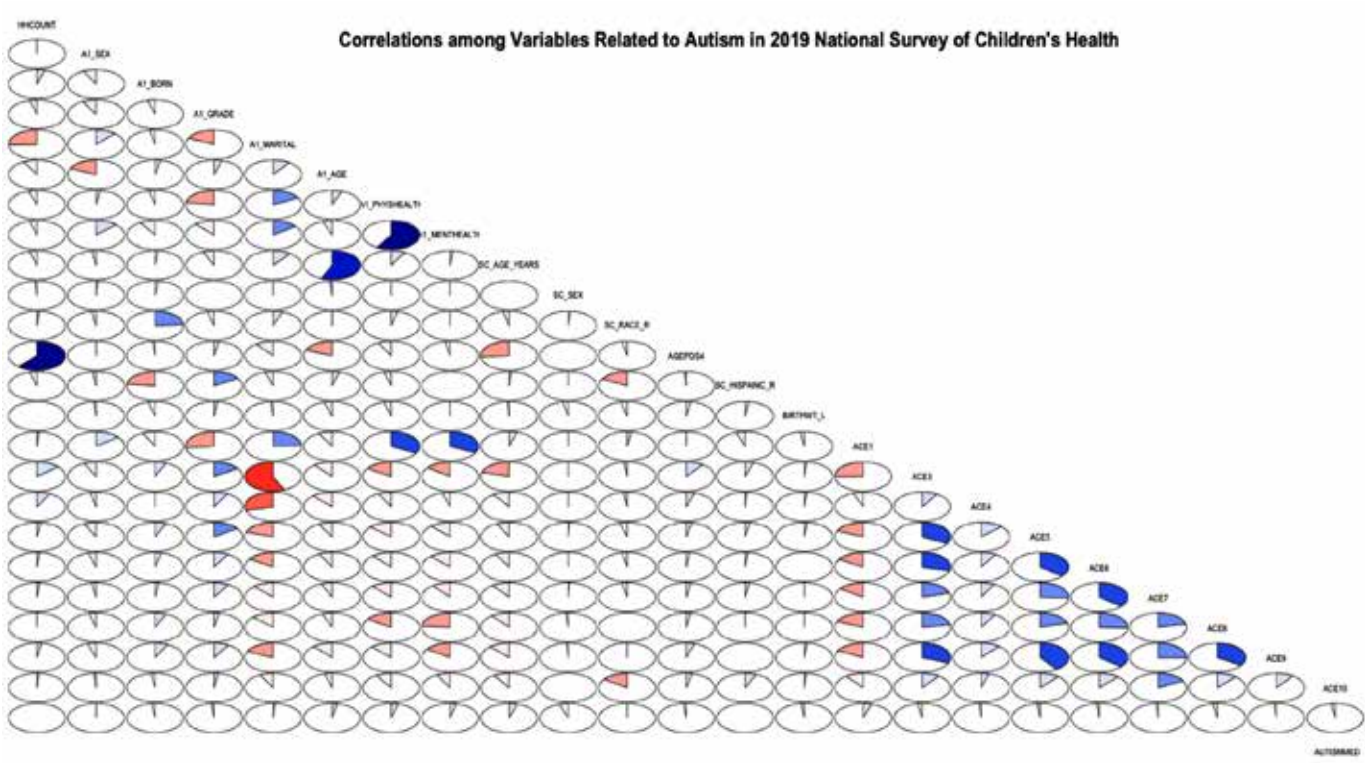

Figure 1. Correlation among variables

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.07213    2.66074  -0.779 0.436109
HHCOUNT        0.08447    0.13388   0.631 0.528043
A1_SEX        -0.19398    0.28884  -0.672 0.501855
A1_BORN       -0.67205    0.54575  -1.231 0.218170
A1_GRADE       0.01166    0.07533   0.155 0.877036
A1_MARITAL    -0.15882    0.12512  -1.269 0.204328
A1_AGE         0.01486    0.01638   0.907 0.364441
A1_PHYSHEALTH  0.31368    0.17159   1.828 0.067535 .
A1_MENTHEALTH  0.01666    0.17218   0.097 0.922911
SC_AGE_YEARS   0.08895    0.03384   2.628 0.008579 **
SC_SEX        -1.29141    0.31505  -4.099 4.15e-05 ***
SC_RACE_R     -0.08852    0.08788  -1.007 0.313782
AGEPOS4       -0.00577    0.18131  -0.032 0.974614
SC_HISPANIC_R -0.38219    0.38751  -0.986 0.324007
BIRTHWT_L     -0.76067    0.35645  -2.134 0.032839 *
ACE1           0.57006    0.14735   3.869 0.000109 ***
ACE3          -0.21313    0.33652  -0.633 0.526521
ACE4          -0.88675    0.49241  -1.801 0.071730 .
ACE5           0.17044    0.52198   0.327 0.744031
ACE6           0.01877    0.55622   0.034 0.973085
ACE7           0.67314    0.65041   1.035 0.300701
ACE8           0.13867    0.41484   0.334 0.738178
ACE9          -0.06939    0.43282  -0.160 0.872623
ACE10         -0.66089    0.51839  -1.275 0.202349
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                [,1]
(Intercept)   0.1259168
HHCOUNT       1.0881461
A1_SEX        0.8236755
A1_BORN       0.5106616
A1_GRADE      1.0117235
A1_MARITAL    0.8531473
A1_AGE        1.0149668
A1_PHYSHEALTH 1.3684558
A1_MENTHEALTH 1.0168010
SC_AGE_YEARS  1.0930289
SC_SEX        0.2748817
SC_RACE_R     0.9152824
AGEPOS4       0.9942471
SC_HISPANIC_R 0.6823669
BIRTHWT_L     0.4673508
ACE1          1.7683658
ACE3          0.8080552
ACE4          0.4119904
ACE5          1.1858233
ACE6          1.0189436
ACE7          1.9603741
ACE8          1.1487411
ACE9          0.9329604
ACE10         0.5163909
```

Figure 2. Logistic regression results (left), odds ratio for each variable (right)

From the logistic regression results, it is not hard to find that, taking a 90% confidence level, children's age, sex, physical health, birth weight, whether the parents or guardians are dead, and whether the household is able to cover daily basic life are all significant predictors of the dependent variable. More specifically, according to the logistic regression coefficients, children's sex, birth weight, and whether the parents or guardians are dead are significant negative predictors, while physical health, age, and whether a family's income can cover basic living are significant positive predictors. The results of the logistic regression model corroborate with the findings from the correlogram shown above.

In addition, combing with the odds ratio table, we could identify those older children are more likely to develop autism ($p = 0.0086$, $OR = 1.093$), children with a worse physical health condition are more likely to develop autism ($p = 0.067$, $OR = 1.093$), and children in the family whose income is not able to cover basic living are more likely to develop autism ($p < 0.001$, $OR = 1.77$).

Besides, children who do not have a low birth weight ($p = 0.033$, $OR = 0.47$), whose parents or guardians are not dead are less like to develop autism ($p = 0.071$, $OR = 0.41$), and female children are less likely to develop autism ($p < 0.001$, $OR = 0.82$).

3.3 Model Validation

The figure below displays the ROC curve for the logistic regression model with an AUROC score 0.68. It can be concluded that the model has achieved a relatively good performance much better than randomly guessing.

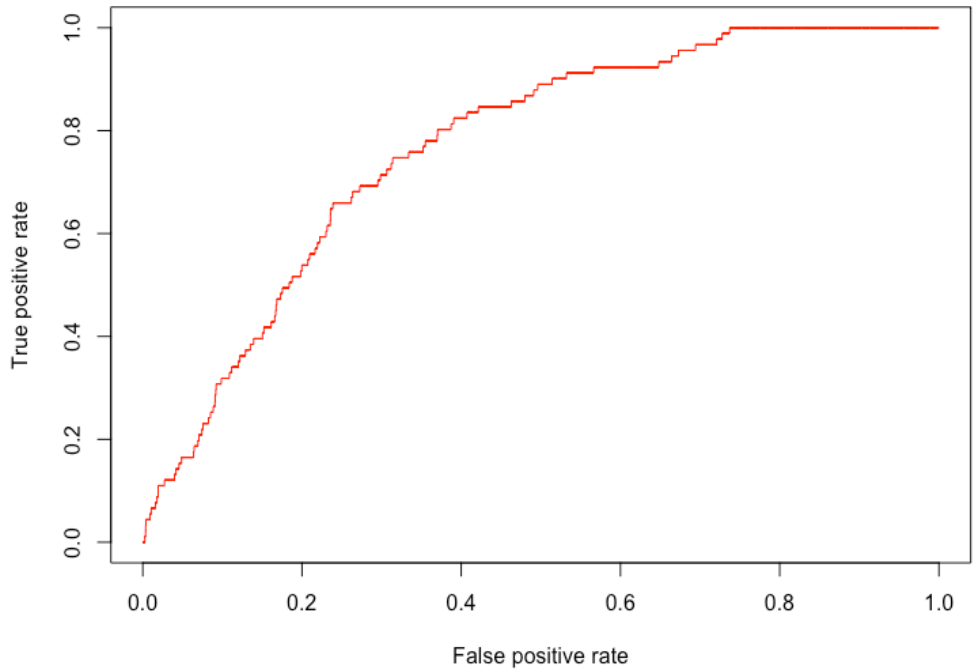

Figure 3. ROC curve

4. Discussions

The intention of this study is to build a predictive model with the best performance and to investigate the factors most related to children's chance of developing autism. One logistic regression model was built and has achieved good performance much better than randomly guessing. Also, from the logistic regression results, we are able to ascertain that the child's age, sex, physical health, birth weight, family income, and negative childhood experiences are all significant predictors of the dependent variable, which corroborates with the findings from the correlogram. Combining both results, we can see that in order to assess the child's chance of developing autism, it will be most effective to look at factors such as the child's negative childhood experience and family social-economic status.

One limitation of the study is that data entries with missing values are imputed with the mean value of the entire feature column. This is a timesaving but defective approach. Depending on the number of such data entries, it is possible that we might introduce a new bias into the dataset. For future studies, we may use more advanced techniques such as k-nearest neighbors (kNN) imputation, which replaces missing values with the mean of k (a parameter selected by the user) nearest neighbors of that sample. This technique requires more efforts but can generally achieve better performance and may help create a more accurate model.

References:

1. Centers for Disease Control and Prevention. (2020, June 29). Diagnostic criteria. Centers for Disease Control and Prevention. Retrieved February 21, 2022. From URL: https://www.cdc.gov/ncbddd/autism/hcp-dsm.html
2. Centers for Disease Control and Prevention. (2021, December 2). Autism data visualization tool. Centers for Disease Control and Prevention. Retrieved February 21, 2022. From URL: https://www.cdc.gov/ncbddd/autism/data/index.html#data
3. Google. (2020 Aug. 11). Classification: ROC curve and Auc | machine Learning crash course. Google. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.
4. National Survey of Children's Health – Data Resource Center for Child and Adolescent Health, 2019. URL: https://www.childhealthdata.org/learn-about-the-nsch/NSCH