Binary Logistic Regression in Ascertaining Important Factors for Autistic Spectrum Disorder

Chi Nguyen, Huong Nguyen, Phuong Dao

STA 9797 - Advanced Data Analysis, Baruch College

Abstract

Autistic Spectrum Disorder (ASD) is a disorder of neurodevelopment that negatively affects how people communicate, learn, behave, and socially interact. It is mostly influenced by a combination of genetic and environmental factors. Many studies have shown that early interventions might help improve these conditions. However, diagnosing ASD is a challenging and costly process that may take years. The screening procedure is the initial stage in diagnosing ASD, and the results of this phase will assist the doctor to determine whether individuals should seek further examinations. This phase has recently been shortened while being highly accurate, thanks to the assistance of intelligent technologies based on machine learning. In the scope of this project, we applied logistic regression model to the Autism Spectrum Disorder Screening data set to assess the associations of various aspects to the recognition of autism. The application study indicated some interesting factors that have effect on ASD.

Keywords: Logistic Regression, autistic spectrum disorder, likelihood ratio test, Wald test.

Background and Data Description

The data set is made up of survey responses from 704 participants who completed an app form. The survey includes 20 questions, in which there are 10 questions relating to demographics of respondents and 10 questions asking about individual reactions to specific circumstances. The 10 behavioral answers will be graded based on AQ-10 (Autism Spectrum Quotient), which combine various criteria of social interactions, communication as well as restricted, repetitive patterns of behaviors. If the participant scores 6 or above, a specialist diagnostic assessment can be considered later.

The detailed description of the survey question can be found as below:

- Social communication questions:
  - Question 5: I find it easy to read between the lines when someone is talking to me.
  - Question 7: When I'm reading a story, I find it is difficult to work out the characters' intention.
  - Question 6. I know how to tell if someone listening to me is getting bored.
  - Question 9: I find it easy to work out what someone is thinking or feeling just by looking at their face.
  - Question 10: I find it difficult to work out people's intention.

The answer "agree" for question 7 and 10, and 'disagree' for question 5,6,9 deliver the similar messages. People having ASD often have deficits in nonverbal communicative behaviors used for social interaction. That leads to difficult to initiate, response, and have a poorly integrated verbal and nonverbal communication in various social contexts.

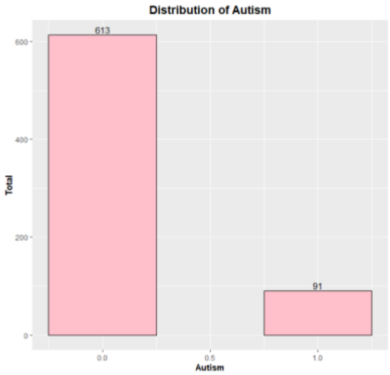
- Repetitive patterns and characteristic behaviors questions:
  - Question 1: I often notice small sounds when other do not.
  - Question 2: I usually concentrate more on the whole picture, rather than small details.
  - Question 3: I find that it easy to do more than one thing at one.
  - Question 4: If there is an interruption, I can switch back to to what I was doing very quickly.
  - Question 8: I like to collect information about categories of things (e.g. types of car, types of bird, type of train, types of plant etc).

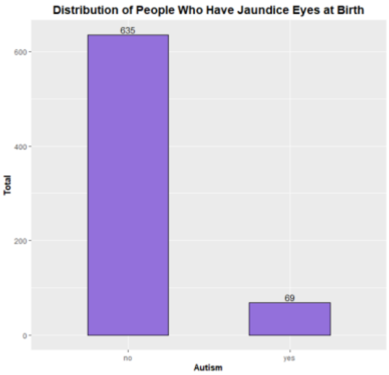
Similarly, with social impairment and communication difficulties, the agree answers for question 1 and 8 indicate the fixed interests of individuals and their hyperactivity to sensory input. The disagreement for question 2,3,4 show that there are signs of insistence on sameness, inflexible adherence to routines, and sensitivity to small changes. The points are given to the answers which indicate the symptoms commonly found in people having ASD.

The survey result is retrieved from:
https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults

Except for questions 7 and 8, all other behavior questions are used in this project to assess chance of having autism. Meanwhile, only 3 demographic variables are kept. Thus, there are total of 12 variables, including response variable `austim` used in the project. The details of other variables, besides behavior answers, can be found in the table below:

Variables	Types	Description
austim	binary	Have autism or not (1: yes, 0: no
age	numeric	Age of a person
gender	categorical, nominal	Gender (f: female, m: male)
jundice	binary	Have yellow eyes or not





Quickly exploring we see that the data is quite unbalance between the proportion of participants who were labeled with ASD and participants who had jaundice at birth. The variation of age variable is greater for people who were labeled with autism than people who were not. Besides, there are two missing values in age, and an outlier in age (380) which is clearly a typo. We removed those three observations from the data set.

Methods

a. Logistic Regression Model

Logistic regression is a statistical analysis method to predict binary outcome (yes or no) based on predictor variables (independent variables). The predictor can categorical or continuous. In details, the logistic model models the probability of an event taking place by having the log-odds of the event be a linear combination of one or more dependent variables.

The logistic model for probability of success can be written as:

$$Pr(Y = 1|X = 1) = p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

where β_0 and β_1 are parameters, $x = 0$ or 1, and Y is the binary outcome of 0 (failure, no response) or 1 (success, response) . This model can be also written in terms of logit:

$$logit(p_x) = \beta_0 + \beta_1 x$$

The odds for $x(x = 0$ or 1) is equal to

$$\frac{p_x}{1 - p_x} = \frac{\left(e^{\beta_0+\beta_1 x}\right)/\left(1 + e^{\beta_0+\beta_1 x}\right)}{1/(1 + e^{\beta_0+\beta_1 x})}$$

The odds can take on any value between 0 and ∞. Values of the odds close to 0 and ∞ indicate very low and high probabilities of default, respectively.

By taking the logarithm of both sides of odds formula, we have

$$logit(p_x) = \log\left(\frac{p_x}{1 - p_x}\right) = \beta_0 + \beta_1 X$$

This is called the log odds or logit. From this formula, we can see that increasing X by one unit changes the log odds by β_1, and it multiplies the odds by e^{β_1} (e^{β_1} is called the odds ratio or OR).

To generalize the relationship between parameters β_1, X, and $p(X)$, if β_1 is positive then increasing X will be associated with increasing $p(X)$, and vice versa.

b. Test statistics for H_0: $\beta_1 = 0$ in logistic model

Wald Test is commonly used to test the significant of individual coefficient in Logistic Regression. The hypothesis test of Wald test is:

$$\boldsymbol{H_0}: \boldsymbol{\beta_1} = \mathbf{0}$$

$$\boldsymbol{H_a}: \boldsymbol{\beta_1} \neq \mathbf{0}$$

The Wald test statistic Z is approximately distributed as $N(0,1)$ under the null if the sample size is large. In this study, we use Wald Test to examine the importance of individual covariates.

c. Likelihood Ratio Test

The Likelihood-Ratio Test (or likelihood-ratio chi-squared test) is a statistical test used to compare the goodness of fit of two models based on the ratio of their likelihoods. In this study we use the likelihood ratio test to test the "goodness of fit" between the full model (f) and the reduced model (r), where null hypothesis is that the reduced model is better. The better model maximizes the likelihood function. The test statistics calculation is as followed:

$$G = -2\log_e\left(\frac{L_r(\hat{\theta})}{L_f(\hat{\theta})}\right)$$

G follows a Chi-squared distribution with the degrees of freedom equals the difference of the number of parameters between the two models.

Detailed Output

1. Fitting logistics regression using all 11 variables

Estimating a logistic regression model using the generalized linear model function

```
data.fit <- glm(autism~., data = df, family = "binomial")
summary(data.fit)

##
## Call:
## glm(formula = autism ~ ., family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4464  -0.5660  -0.3926  -0.2796   2.6232
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.489831   0.422085  -8.268  < 2e-16 ***
## A1_Score     0.479902   0.313500   1.531 0.125822
## A2_Score     0.040844   0.251710   0.162 0.871097
## A3_Score     0.056016   0.275432   0.203 0.838841
## A4_Score     0.944442   0.300099   3.147 0.001649 **
## A5_Score    -0.195790   0.291944  -0.671 0.502449
## A6_Score    -0.109499   0.297843  -0.368 0.713143
## A9_Score     0.586105   0.297140   1.972 0.048554 *
## A10_Score    0.410112   0.277406   1.478 0.139306
## age          0.009460   0.005324   1.777 0.075585 .
## genderm     -0.451427   0.240394  -1.878 0.060399 .
## jundiceyes   1.065297   0.315046   3.381 0.000721 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 541.50  on 701  degrees of freedom
## Residual deviance: 482.81  on 690  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 506.81
##
## Number of Fisher Scoring iterations: 5
```

Three variables: `A4_Score`, `A9_Score`, `jundiceyes` are significant at 0.05 level of significance. The logistic regression coefficients state the change in the log odds of the response for one unit increase in the predictor variables. In details:

- For every one unit change in `A4_Score`, the log odds of having autism (versus non-autistic) increases by 0.94.
- For every one unit change in `A9_Score`, the log odds of having autism (versus non-autistic) increases by 0.59.
- For every one unit change in `jundiceyes`, the log odds of having autism (versus non-autistic) increases by 1.07.

Interpreting the odd-ratios and 95% Confidence Interval

Interpreting the odd-ratios:

- For a one unit increase in `A4_Score`, the odds of being autistic (versus not being autistic) increase by a factor of 2.57.
- For a one unit increase in `A9_Score`, the odds of being autistic (versus not being autistic) increase by a factor of 1.80.
- For a one unit increase in `jundiceyes`, the odds of being autistic (versus not being autistic) increase by a factor of 2.9.

Interpreting 95% Confidence Interval

- The estimated odds ratio of having ASD among people who were born with jaundiced eyes is greater than the odds of having ASD among people who did not have jaundiced eyes at birth from as low as 1.54 to as large as 5.33 times.
- The estimated odds ratio of having ASD among people who have the positive answers for the restricted and repetitive behavior question is greater than the odds of having ASD among people who do not from 1.44 to 4.70 times.

- The estimated odds ratio of having ASD among people who have the negative answers for the nonverbal communication understanding question is greater than the odds of having ASD among people who do not from 1 to 3.2 times.

```
exp(cbind(OR = coef(data.fit), confint(data.fit)))

##                     OR      2.5 %     97.5 %
## (Intercept) 0.03050602 0.01279777 0.06746407
## A1_Score    1.61591666 0.89478187 3.08229015
## A2_Score    1.04168927 0.63449694 1.70642099
## A3_Score    1.05761485 0.61619330 1.81878521
## A4_Score    2.57137694 1.44218045 4.69595984
## A5_Score    0.82218523 0.46191856 1.45536195
## A6_Score    0.89628342 0.49707502 1.60136004
## A9_Score    1.79697497 1.00172410 3.21812038
## A10_Score   1.50698669 0.88212603 2.62789768
## age         1.00950463 0.99852067 1.02322617
## genderm     0.63671862 0.39528991 1.01691561
## jundiceyes  2.90169956 1.54192574 5.32961473
```

2. Testing of significance for individual regression coefficients in logistic regression with Wald Test

```
## Wald test for A1_Score
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  2.343309  on  1  and  690  df: p= 0.12628

## Wald test for A2_Score
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  0.02632983  on  1  and  690  df: p= 0.87114

## Wald test for A3_Score
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  0.04136183  on  1  and  690  df: p= 0.8389

## Wald test for A4_Score
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  9.904225  on  1  and  690  df: p= 0.00172

## Wald test for A5_Score
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  0.4497602  on  1  and  690  df: p= 0.50267

## Wald test for A6_Score
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  0.1351581  on  1  and  690  df: p= 0.71326

## Wald test for A10_Score
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  2.185608  on  1  and  690  df: p= 0.13976
## Wald test for A9_Score
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  3.8907  on  1  and  690  df: p= 0.048953

## Wald test for age
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  3.157375  on  1  and  690  df: p= 0.076025

## Wald test for gender
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  3.526391  on  1  and  690  df: p= 0.060821

## Wald test for jundice
##  in glm(formula = autism ~ ., family = "binomial", data = df)
## F =  11.43387  on  1  and  690  df: p= 0.00076184
```

From the output of the Wald test for the importance of individual covariates, we can observe that **only `A4_Score`, `A9_Score`, and jaundice variables** having **p_values < 0.05**. It indicates that the positive answer for the restricted and repetitive behavior question, the negative answer for the nonverbal communication understanding question, and the appearance of jaundice are significant and contribute most to the probability of identifying an individual having ASD traits. Meanwhile, Wald test for the coefficients of `A1_Score`, `A_2 Score`, `A3_Score`, `A_5  Score`, `A6_Score`, `A_10  Score`, `gender` and `age` are not significant with p_values = 0.12, 0.87, 0.84, 0.5, 0.71, 0.14, 0.06, and 0.08 respectively.

3. Fitting logistic regression after removing insignificant variables

a. Using the most contributed variable

Estimating a logistic regression model using the generalized linear model function

We notice that `jundice` is the variable that has the highest weight contribution to the model, so to demonstrate the interpretation meaning of the estimated coefficient, and its measure of association, we start with a model which has only `jundice` variable. We hypothesize that `jundice` alone would be good enough to classify if an individual has ASD trait. We claim that if knowing about jaundice eyes, we could predict how likely that person would have ASD.

Hypothesis:

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

```
data.fit.remove <- glm(autism~jundice, data=df, family = "binomial")
summary(data.fit.remove)

##
## Call:
## glm(formula = autism ~ jundice, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8274  -0.4870  -0.4870  -0.4870   2.0933  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -2.0724     0.1259 -16.457  < 2e-16 ***
## jundiceyes    1.1763     0.2937   4.005  6.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 542.05  on 703  degrees of freedom
## Residual deviance: 527.94  on 702  degrees of freedom
## AIC: 531.94
## 
## Number of Fisher Scoring iterations: 4
```

With p-value = 6.2e-05 < 0.05, we reject the hypothesis and conclude that `jundice` is statistically significant variable.

Interpreting the odd-ratios and 95% Confidence Interval

- The estimated odds ratio for `jundice` variable is 3.24. It means that the odd of having ASD among people who were born with jaundiced eyes is 3.24 greater than the odds of having ASD among people who did not have jaundiced eyes at birth in this study sample.

- The estimated odds ratio of having ASD among people who were born with jaundiced eyes is greater than the odds of having ASD among people who did not have jaundiced eyes at birth from as low as 1.79 to as large as 5.69 times, at 95% level of confidence.

```
exp(cbind(OR = coef(data.fit.remove), confint(data.fit.remove)))
```

```
##                    OR     2.5 %    97.5 %
## (Intercept) 0.1258865 0.0975428 0.1599191
## jundiceyes  3.2423110 1.7926816 5.6999297
```

b. Using all statistically significant variables

```
data.fit.sig <- glm(autism~jundice+A9_Score+A4_Score, data=df, family = "bino
mial")
summary(data.fit.sig)
```

```
## 
## Call:
## glm(formula = autism ~ jundice + A9_Score + A4_Score, family = "binomial", 
##     data = df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.0938  -0.5219  -0.4387  -0.3211   2.4456  
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.9390     0.2322 -12.659  < 2e-16 ***
## jundiceyes    1.0786     0.3055   3.531 0.000414 ***
## A9_Score      0.6464     0.2438   2.652 0.008013 **
## A4_Score      1.0140     0.2690   3.770 0.000163 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 542.05  on 703  degrees of freedom
## Residual deviance: 495.54  on 700  degrees of freedom
## AIC: 503.54
##
## Number of Fisher Scoring iterations: 5
```

- With p-values < 0.05, all three variables are significant.
- Likelihood Ratio (or Deviance) Test: The ratio of log-likelihood of the reduced model and full model follows a Chi-square distribution with 8 degrees of freedom under the hypothesis that the coefficients for 8 excluded variables are equal to 0. The value of Likelihood ratio statistic is $G = -2 * \left(L_r - L_f\right) = 12.73$, with $P[X^2(8) > 12.73309] = 0.12 > 0.05$. Therefore, we conclude that the full model is no better than the reduced model, meaning there is little statistical justification for including A1_Score, A2_Score, A3_Score, A5_Score, A6_Score, A10_Score, age, and gender.

c. Adding important variables to the model

However, gender variable is known as an important variable of ASD trait, we then add back gender in the model and examine it. It interestingly turns out that genderbecomes statistically significant with p_value = 0.039 in the new model.

```
data.fit.add <- glm(autism~jundice+A9_Score+A4_Score+gender, data=df, family
= "binomial")
summary(data.fit.add)

##
## Call:
## glm(formula = autism ~ jundice + A9_Score + A4_Score + gender,
##     family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1864  -0.5772  -0.3925  -0.2855   2.5379
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6915     0.2552 -10.546  < 2e-16 ***
## jundiceyes    1.0743     0.3073   3.496 0.000473 ***
## A9_Score      0.6546     0.2459   2.662 0.007758 **
## A4_Score      0.9836     0.2710   3.630 0.000284 ***
## genderm      -0.4882     0.2370  -2.060 0.039415 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 542.05  on 703  degrees of freedom
## Residual deviance: 491.24  on 699  degrees of freedom
## AIC: 501.24
##
## Number of Fisher Scoring iterations: 5
```

4. Cofounding examination

Due to data analysis exploration, we removed variables of `A7_Score` and `A8_score` at the beginning of the project. However, in this part, we decided to bring them back to discover if there is a cofounding effect of the removed variables to the existing model.

We examine the magnitude changes of coefficients associated with the remaining variables in the model after removing statistically insignificant from the model, the largest percentage change is 11% for the coefficient of `A9_Score` (point for the negative answer of I find it easy to work out what someone is thinking or feeling just by looking at their faces). By using a threshold of 20% change in coefficient, we conclude that there are no confounding variables in the excluded variables.

$$\mathit{\Delta}\beta^{T} = [0.0368,0.114,0.082,0.006]$$

Now we examine the effect of each variable that was excluded from the analysis to see if there is any confounding variable.

- First we add variable `A7_Score` back to model of four significant variables.

```
##
## Call:
## glm(formula = autism ~ jundice + A9_Score + A4_Score + A7_Score +
##     gender, family = "binomial", data = df7)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2829  -0.5799  -0.3772  -0.3008   2.6397
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6074     0.2603 -10.016  < 2e-16 ***
## jundiceyes    1.0969     0.3083   3.558 0.000373 ***
## A9_Score      0.7226     0.2502   2.888 0.003879 **
## A4_Score      1.0325     0.2732   3.780 0.000157 ***
## A7_Score     -0.3801     0.2479  -1.533 0.125298    
## genderm      -0.4654     0.2378  -1.958 0.050279 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 542.05  on 703  degrees of freedom
## Residual deviance: 488.85  on 698  degrees of freedom
## AIC: 500.85
## 
## Number of Fisher Scoring iterations: 5
```

- The Log likelihood of model of 5 variables (adding `A7_Score`) is -244.3.
- Log likelihood of model of 4 variables (`jundice`, `A9_Score`, `A4_Score`, `gender`) is 245.51
- Ratio likelihood of two model is 2.37 with $P[X^2(1) > 2.37] = 0.12 > 0.05$. This result agrees with the p-value of `A7_Score` by Wald test indicating that model with `A7_Score` is no better than the one without it.
- The magnitude of coefficients in the model does not change significantly when adding `A7_Score` variable. The largest change of `A9_Score` is from 0.6449 to 0.72, equivalently 10%.
- Similarly, adding `A8_Score` to current model does not have statistically significant effect.

```
data.fit.a8 <- glm(autism~jundice+A9_Score+A4_Score+A8_Score+gender, data=df8
, family = "binomial")
summary(data.fit.a8)

## 
## Call:
## glm(formula = autism ~ jundice + A9_Score + A4_Score + A8_Score + 
##     gender, family = "binomial", data = df8)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.2050  -0.5895  -0.3737  -0.2932   2.5758  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -2.7820     0.2973  -9.358  < 2e-16 ***
## jundiceyes    1.0728     0.3075   3.489 0.000485 ***
## A9_Score      0.6323     0.2489   2.541 0.011052 *  
## A4_Score      0.9862     0.2714   3.633 0.000280 ***
## A8_Score      0.1552     0.2551   0.608 0.542901    
## genderm      -0.4984     0.2377  -2.097 0.036013 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 542.05  on 703  degrees of freedom
## Residual deviance: 490.87  on 698  degrees of freedom
## AIC: 502.87
##
## Number of Fisher Scoring iterations: 5
```

We conclude that there are no excluded variables are confounders of the relationship between any of the remaining covariates and ASD trait.

5. Explore possible interactions among the main effects

Exploring the significant interaction terms

There only significant interaction term is between `gender` and `jundice` with p-value = 0.01 < 0.05. Therefore, we will choose to include this interaction term into the model.

```
data.interact.only <- glm(autism~jundice*gender+A4_Score*A9_Score+gender*A4_S
core+A4_Score*jundice+A9_Score*gender+A9_Score*jundice, data=df, family = "bi
nomial")
summary(data.interact.only)
```

```
##
## Call:
## glm(formula = autism ~ jundice * gender + A4_Score * A9_Score +
##     gender * A4_Score + A4_Score * jundice + A9_Score * gender +
##     A9_Score * jundice, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3146  -0.5278  -0.3870  -0.2772   2.5605
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.5540     0.3211  -7.953 1.83e-15 ***
## jundiceyes            0.1782     0.7545   0.236   0.8133
## genderm              -0.6856     0.4797  -1.429   0.1529
## A4_Score              0.8042     0.4195   1.917   0.0552 .
## A9_Score              0.5580     0.6104   0.914   0.3606
## jundiceyes:genderm    1.6764     0.6585   2.546   0.0109 *
## A4_Score:A9_Score     0.3566     0.6263   0.569   0.5691
## genderm:A4_Score      0.3749     0.5736   0.654   0.5134
## jundiceyes:A4_Score  -0.6137     0.7493  -0.819   0.4127
## genderm:A9_Score     -0.7550     0.5258  -1.436   0.1510
## jundiceyes:A9_Score   0.9769     0.7033   1.389   0.1648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 542.05  on 703  degrees of freedom
## Residual deviance: 480.93  on 693  degrees of freedom
## AIC: 502.93
##
## Number of Fisher Scoring iterations: 5
```

Fitting logistic regression model with the interaction term

With p-value = 0.013 < 0.05, the interaction term between gender and jundice is statistically significant.

```
data.fit.interact <- glm(autism~jundice+A9_Score+A4_Score+gender+jundice*gend
er, data=df, family = "binomial")
summary(data.fit.interact)
```

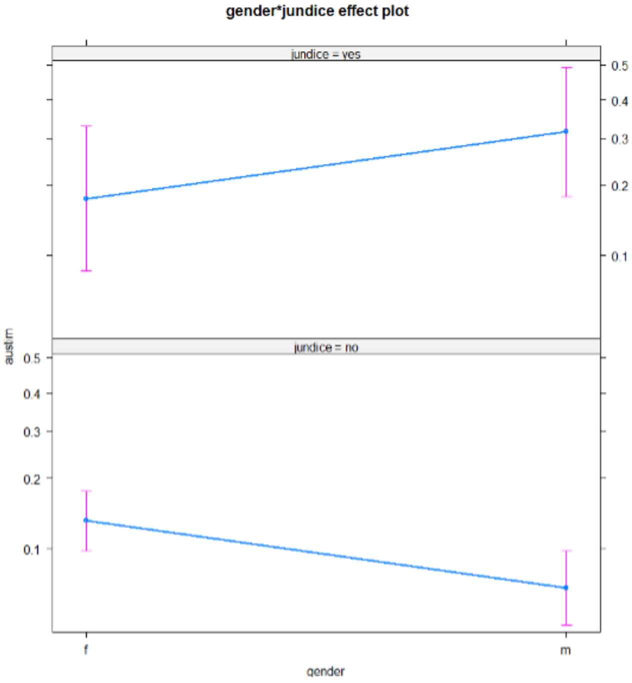
```
##
## Call:
## glm(formula = autism ~ jundice + A9_Score + A4_Score + gender +
##     jundice * gender, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2546  -0.5795  -0.3802  -0.2610   2.6065
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.5908     0.2556 -10.137  < 2e-16 ***
## jundiceyes           0.3359     0.4490   0.748 0.454382
## A9_Score             0.6735     0.2477   2.719 0.006550 **
## A4_Score             0.9902     0.2728   3.630 0.000284 ***
## genderm             -0.7721     0.2682  -2.879 0.003986 **
## jundiceyes:genderm   1.5429     0.6239   2.473 0.013392 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 542.05  on 703  degrees of freedom
## Residual deviance: 484.99  on 698  degrees of freedom
## AIC: 496.99
##
## Number of Fisher Scoring iterations: 5
```

Likelihood Ratio (or Deviance) Test

* The Log likelihood of the model with interaction term is: -242.49
* The Log likelihood of the main effect model only is: -245.62
* The ratio log likelihood of two function: 6.25

* With 1 degree of freedom, the p_value associated with the G statistic is: 0.012 < 0.05. Therefore, We conclude that the interaction between `gender` and `jundice` does contribute to the model.

Interaction between `gender` and `jundice`



The `gender*jundice` effect plot shows how the probability of having ASD changes for different combinations of gender (male/female) and jundice (yes/no).

The expected probability of having ASD in female lower than in male who had jundice at birth (given the condition that score in A4 and A9 questions are the same). Additionally, the expected probability of having ASD trait of a male who had jundice at birth is the highest at 0.31 versus 0.06 for the same gender who did not have jundice. For female, the difference is less (0.17 vs 0.31) for female who had jundice and female who did not have jundice.

Conclusion

The aim of this study is to analyze the independent variables that contribute significantly and enhance the risk of Autistic Spectrum Disorder by using Logistic Regression. Based on the results of data analysis and discussion, `A4_Score`, `A_9 Score` and `jundice` are important factors, that have been proved through our Wald test result, and `jundice` has the highest weight contribution to our model. Additionally, by adding interaction terms into the current model, we found that the interaction between `gender` and `jundice` contribute significantly to our model. It worth noting that the expected probability of having ASD trait of a male who had jaundice at birth is highest, compared to the other combinations.

By doing this project, we can understand more about how to apply logistic regression to interpret a classification problem. Although the methods applied with different test statistics are still limited and intuitive, they have given us the foundation to approach a new problem and go into deep-dive analysis in the future.