

Live Session - Week 3: Discrete Response Models

Lecture 3

Professor Jeffrey Yau

Agenda

1. Quiz 2 (start promptly 5 minutes after class begins)
2. Lecture intro and review some concepts from w203
3. An extended example to practice concepts/techniques covered in this week: multiple breakout room discussions
4. Discussion of HW03_dueWeek04
5. Q&A (if time permits)

Required Readings:

BL2015: Christopher R. Bilder and Thomas M. Loughin. Analysis of Categorical Data with R. CRC Press. 2015.

- Ch. 2.2.5 – 2.2.7, 2.3

Topics covered in Week 3

- Variable transformation: interactions among explanatory variables
- Variable transformation: quadratic term
- Categorical explanatory variables
- Odds ratio in the context of categorical explanatory variables
- Convergence criteria and complete separation

Familiarity with the concepts and techniques covered in this and last lecture are critical, as they will be used frequently in the next two lectures in situations that are more general (from two categorical to $J > 2$ categories and from unordered categorical variables to ordinal variables). With multinomial logistic regression models, the notation will be heavier.

The key objectives in this live session are to learn how to incorporate various transformation of variables (or, in machine learning terminology, “feature engineering”) and interpret the results when these transformed variables are part of the model specification. Variable transformations (or feature engineering) are useful in real life statistical and machine learning modeling.

In general, the odd ratios answer the question “how much the odds of success have changed by k-unit increase?” The amazing feature of logistic regression model is that the odd ratios (of the odds of success before and after the k-unit increase in a particular explanatory variable) is simplified to the exponential of the product between k and the coefficient estimate associated with that variable. That is, “the odds of a success change by $\exp(k\beta_j)$ times for every k-unit increase in x_j ”

Review some concepts from w203

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (\beta_3 x_1 \times x_2) + \epsilon$$

$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

$$\frac{\partial y}{\partial x_1} = \beta_1 + 2\beta_2 x_1$$

Start-up code

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

# Set working directory
#setwd("~/Documents/Teach/Cal/w271/LiveSessions/week03/")
wd <- getwd()
wd

## [1] "/Users/macbook/Desktop/271/main-2019-summer/live-session-files/week03"

# Start with a clean R environment
rm(list = ls())

# Load Libraries
library(car)
library(dplyr)
library(Hmisc)
library(skimr)
library(ggplot2)
library(stargazer)
library(mcprofile)
```

In this live session, we will practice binary logistic regression modeling, with a focus on the materials covered in week 3, using an autism screening dataset (for toddlers). I've obtained this dataset from kaggle

Thanks to the generosity of Dr. Fadi Fayeze Thabtah, who kindly provides this dataset to kaggle, we have this rarely available dataset on autism screening to practice the concepts and techniques on binary logistic regression modeling that we've learned in the last few weeks.

The dataset also comes with a very detailed description, whose Word document I also share with you. Below are some brief description provided on the aforementioned kaggle webpage:

Context: The dataset was developed by Dr Fadi Fayeze Thabtah (fadifayeze.com) using a mobile app called ASDTests (ASDtests.com) to screen autism in toddlers. See the description file attached with the CSV data to know more about the variables and the class. This data can be used for descriptive and predictive analyses such as classification, clustering, regression, etc. You may use it to estimate the predictive power of machine learning techniques in detecting autistic traits.

Brief Description of the Variables This data page on kaggle also provide some very basic descriptive graphs on the variables in this dataset.

A1 - A10: Items within Q-Chat-10 in which questions possible answers : “Always, Usually, Sometimes, Rarely & Never” items’ values are mapped to “1” or “0” in the dataset.

There are two variables in the data that will not be used in our analysis: 1. **Case_No**: the individual case number; this is an identifier variable 2. **Qchat.10.Score**: the dataset document suggests that this variable not be used in a classification problem, as the score variable is used to defined the **Class.ASD.Traits**.

```
# Load data
autism <- read.csv("autism.csv", header = TRUE, sep = ",")

# Attach the dataframe autism
attach(autism)

# View(autism)

# Examine the structure of the data
str(autism)
```

```
## 'data.frame': 1054 obs. of 19 variables:
## $ Case_No : int 1 2 3 4 5 6 7 8 9 10 ...
## $ A1 : int 0 1 1 1 1 1 1 0 0 1 ...
## $ A2 : int 0 1 0 1 1 1 0 1 0 1 ...
## $ A3 : int 0 0 0 1 0 0 0 0 0 1 ...
## $ A4 : int 0 0 0 1 1 0 1 0 0 0 ...
## $ A5 : int 0 0 0 1 1 1 1 1 0 1 ...
## $ A6 : int 0 1 0 1 1 1 1 0 0 1 ...
## $ A7 : int 1 1 1 1 1 1 0 1 1 0 ...
## $ A8 : int 1 0 1 1 1 1 0 1 0 1 ...
## $ A9 : int 0 0 0 1 1 1 1 1 0 1 ...
## $ A10 : int 1 0 1 1 1 1 0 1 1 1 ...
## $ Age_Mons : int 28 36 36 24 20 21 33 33 36 22 ...
## $ Qchat.10.Score : int 3 4 4 10 9 8 5 6 2 8 ...
## $ Sex : Factor w/ 2 levels "f","m": 1 2 2 2 1 2 2 2 2 2 ...
## $ Ethnicity : Factor w/ 11 levels "asian","black",...: 5 11 5 3 11 2 1 1 1 10 ...
## $ Jaundice : Factor w/ 2 levels "no","yes": 2 2 2 1 1 1 2 2 1 1 ...
## $ Family_mem_with_ASD : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
## $ Who.completed.the.test: Factor w/ 5 levels "family member",...: 1 1 1 1 1 1 1 1 1 3 ...
## $ Class.ASD.Traits. : Factor w/ 2 levels "No","Yes": 1 2 2 2 2 2 2 2 1 2 ...

# Conduct some very basic EDA
describe(autism)
```

```
## autism
##
## 19 Variables 1054 Observations
## -----
## Case_No
##      n missing distinct      Info      Mean      Gmd      .05      .10
##    1054      0     1054        1     527.5     351.7     53.65     106.30
##      .25      .50      .75      .90      .95
##    264.25    527.50    790.75    948.70   1001.35
##
## lowest :      1      2      3      4      5, highest: 1050 1051 1052 1053 1054
## -----
## A1
##      n missing distinct      Info      Sum      Mean      Gmd
```

```

##      1054      0      2      0.738      594      0.5636      0.4924
##
## -----
## A2
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.742      473      0.4488      0.4952
##
## -----
## A3
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.721      423      0.4013      0.481
##
## -----
## A4
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.75      540      0.5123      0.5002
##
## -----
## A5
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.748      553      0.5247      0.4993
##
## -----
## A6
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.732      608      0.5769      0.4887
##
## -----
## A7
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.683      685      0.6499      0.4555
##
## -----
## A8
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.745      484      0.4592      0.4971
##
## -----
## A9
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.75      516      0.4896      0.5003
##
## -----
## A10
##      n missing distinct      Info      Sum      Mean      Gmd
##      1054      0      2      0.728      618      0.5863      0.4856
##
## -----
## Age_Mons
##      n missing distinct      Info      Mean      Gmd      .05      .10
##      1054      0      25      0.971      27.87      8.859      12      15
##      .25      .50      .75      .90      .95
##      23      30      36      36      36
##

```

```

## lowest : 12 13 14 15 16, highest: 32 33 34 35 36
## -----
## Qchat.10.Score
##      n missing distinct      Info      Mean      Gmd      .05      .10
##    1054      0      11    0.991    5.213    3.338      0      1
##      .25      .50      .75      .90      .95
##        3        5        8        9      10
##
## Value      0      1      2      3      4      5      6      7      8      9
## Frequency    54    88    88    96   110   120    96   135    97   95
## Proportion 0.051 0.083 0.083 0.091 0.104 0.114 0.091 0.128 0.092 0.090
##
## Value      10
## Frequency    75
## Proportion 0.071
## -----
## Sex
##      n missing distinct
##    1054      0      2
##
## Value      f      m
## Frequency   319   735
## Proportion 0.303 0.697
## -----
## Ethnicity
##      n missing distinct
##    1054      0      11
##
## asian (299, 0.284), black (53, 0.050), Hispanic (40, 0.038), Latino (26,
## 0.025), middle eastern (188, 0.178), mixed (8, 0.008), Native Indian (3,
## 0.003), Others (35, 0.033), Pacifica (8, 0.008), south asian (60, 0.057),
## White European (334, 0.317)
## -----
## Jaundice
##      n missing distinct
##    1054      0      2
##
## Value      no      yes
## Frequency   766   288
## Proportion 0.727 0.273
## -----
## Family_mem_with_ASD
##      n missing distinct
##    1054      0      2
##
## Value      no      yes
## Frequency   884   170
## Proportion 0.839 0.161
## -----
## Who.completed.the.test
##      n missing distinct
##    1054      0      5
##
## family member (1018, 0.966), Health care professional (5, 0.005), Health

```

```
## Care Professional (24, 0.023), Others (3, 0.003), Self (4, 0.004)
## -----
## Class.ASD.Traits.
##      n missing distinct
##    1054      0        2
##
## Value      No   Yes
## Frequency  326  728
## Proportion 0.309 0.691
## -----
```

```
summary(autism)
```

```
##      Case_No      A1      A2      A3
## Min.   : 1.0   Min.   :0.0000 Min.   :0.0000 Min.   :0.0000
## 1st Qu.: 264.2 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000
## Median : 527.5 Median :1.0000 Median :0.0000 Median :0.0000
## Mean   : 527.5 Mean   :0.5636 Mean   :0.4488 Mean   :0.4013
## 3rd Qu.: 790.8 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max.   :1054.0 Max.   :1.0000 Max.   :1.0000 Max.   :1.0000
##
##      A4      A5      A6      A7
## Min.   :0.0000 Min.   :0.0000 Min.   :0.0000 Min.   :0.0000
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000
## Median :1.0000 Median :1.0000 Median :1.0000 Median :1.0000
## Mean   :0.5123 Mean   :0.5247 Mean   :0.5769 Mean   :0.6499
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max.   :1.0000 Max.   :1.0000 Max.   :1.0000 Max.   :1.0000
##
##      A8      A9      A10      Age_Mons
## Min.   :0.0000 Min.   :0.0000 Min.   :0.0000 Min.   :12.00
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:23.00
## Median :0.0000 Median :0.0000 Median :1.0000 Median :30.00
## Mean   :0.4592 Mean   :0.4896 Mean   :0.5863 Mean   :27.87
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:36.00
## Max.   :1.0000 Max.   :1.0000 Max.   :1.0000 Max.   :36.00
##
## Qchat.10.Score Sex      Ethnicity      Jaundice
## Min.   : 0.000 f:319   White European:334 no :766
## 1st Qu.: 3.000 m:735   asian          :299 yes:288
## Median : 5.000      middle eastern:188
## Mean   : 5.213      south asian   : 60
## 3rd Qu.: 8.000      black         : 53
## Max.   :10.000      Hispanic      : 40
##              (Other) : 80
## Family_mem_with_ASD      Who.completed.the.test Class.ASD.Traits.
## no :884      family member      :1018      No :326
## yes:170      Health care professional: 5      Yes:728
##              Health Care Professional: 24
##              Others                : 3
##              Self                   : 4
##
##
```

```
skim(autism)
```

```
## Skim summary statistics
## n obs: 1054
## n variables: 19
##
## -- Variable type:factor -----
##      variable missing complete    n n_unique
##      Class.ASD.Traits.      0    1054 1054      2
##      Ethnicity              0    1054 1054     11
##      Family_mem_with_ASD     0    1054 1054      2
##      Jaundice                0    1054 1054      2
##      Sex                    0    1054 1054      2
##      Who.completed.the.test  0    1054 1054      5
##
##      top_counts ordered
##      Yes: 728, No: 326, NA: 0  FALSE
##      Whi: 334, asi: 299, mid: 188, sou: 60  FALSE
##      no: 884, yes: 170, NA: 0  FALSE
##      no: 766, yes: 288, NA: 0  FALSE
##      m: 735, f: 319, NA: 0  FALSE
##      fam: 1018, Hea: 24, Hea: 5, Sel: 4  FALSE
##
## -- Variable type:integer -----
##      variable missing complete    n  mean    sd p0    p25    p50    p75
##      A1          0    1054 1054    0.56    0.5  0    0      1      1
##      A10         0    1054 1054    0.59    0.49 0    0      1      1
##      A2          0    1054 1054    0.45    0.5  0    0      0      1
##      A3          0    1054 1054    0.4     0.49 0    0      0      1
##      A4          0    1054 1054    0.51    0.5  0    0      1      1
##      A5          0    1054 1054    0.52    0.5  0    0      1      1
##      A6          0    1054 1054    0.58    0.49 0    0      1      1
##      A7          0    1054 1054    0.65    0.48 0    0      1      1
##      A8          0    1054 1054    0.46    0.5  0    0      0      1
##      A9          0    1054 1054    0.49    0.5  0    0      0      1
##      Age_Mons     0    1054 1054   27.87    7.98 12   23     30     36
##      Case_No      0    1054 1054   527.5   304.41 1  264.25 527.5 790.75
##      Qchat.10.Score 0    1054 1054    5.21    2.91 0    3      5      8
##      p100      hist
##      1
##      1
##      1
##      1
##      1
##      1
##      1
##      1
##      1
##      1
##      1
##      36
##      1054
##      10
```

```
# skim(Age_Mons)
```

```
# Define a function to examine factor variables:
```

```
exam_cat_var = function(var.names) {  
  table(var.names)  
  round(prop.table(table(var.names)), 2)  
}  
apply(autism[, 14:19], 2, table)
```

```
## $Sex
```

```
##
```

```
## f m
```

```
## 319 735
```

```
##
```

```
## $Ethnicity
```

```
##
```

```
## asian black Hispanic Latino middle eastern
```

```
## 299 53 40 26 188
```

```
## mixed Native Indian Others Pacifica south asian
```

```
## 8 3 35 8 60
```

```
## White European
```

```
## 334
```

```
##
```

```
## $Jaundice
```

```
##
```

```
## no yes
```

```
## 766 288
```

```
##
```

```
## $Family_mem_with_ASD
```

```
##
```

```
## no yes
```

```
## 884 170
```

```
##
```

```
## $Who.completed.the.test
```

```
##
```

```
## family member Health care professional Health Care Professional
```

```
## 1018 5 24
```

```
## Others Self
```

```
## 3 4
```

```
##
```

```
## $Class.ASD.Traits.
```

```
##
```

```
## No Yes
```

```
## 326 728
```

```
apply(autism[, 14:19], 2, exam_cat_var)
```

```
## $Sex
```

```
## var.names
```

```
## f m
```

```
## 0.3 0.7
```

```
##
```

```
## $Ethnicity
```

```
## var.names
```

```
## asian black Hispanic Latino middle eastern
```

```
## 0.28 0.05 0.04 0.02 0.18
```



```
##          mixed Native Indian      Others      Pacifica      south asian
##          0.01          0.00          0.03          0.01          0.06
## White European
##          0.32
##
## $Jaundice
## var.names
##   no  yes
## 0.73 0.27
##
## $Family_mem_with_ASD
## var.names
##   no  yes
## 0.84 0.16
##
## $Who.completed.the.test
## var.names
##          family member Health care professional Health Care Professional
##          0.97          0.00          0.02
##          Others          Self
##          0.00          0.00
##
## $Class.ASD.Traits.
## var.names
##   No  Yes
## 0.31 0.69
```

```
# Age describe(Age_Mons) summary(Age_Mons)
skim(Age_Mons)
```

```
##
## Skim summary statistics
##
## -- Variable type:integer -----
## variable missing complete    n mean  sd p0 p25 p50 p75 p100    hist
## Age_Mons          0      1054 1054 27.87 7.98 12  23  30  36  36
```

```
# Crosstab
xtabs(~Sex + Class.ASD.Traits.)
```

```
##      Class.ASD.Traits.
## Sex  No  Yes
##   f 125 194
##   m 201 534
```

```
round(prop.table(xtabs(~Sex + Class.ASD.Traits.), 1), 2)
```

```
##      Class.ASD.Traits.
## Sex  No  Yes
##   f 0.39 0.61
##   m 0.27 0.73
```

```
xtabs(~Ethnicity + Class.ASD.Traits.)
```

```
##          Class.ASD.Traits.
## Ethnicity      No  Yes
##   asian      87 212
```

```
##      black          14  39
##      Hispanic       10  30
##      Latino          6  20
##      middle eastern 92  96
##      mixed           3   5
##      Native Indian   0   3
##      Others           6  29
##      Pacifica         1   7
##      south asian     23  37
##      White European  84 250
```

```
round(prop.table(xtabs(~Ethnicity + Class.ASD.Traits.), 1), 2)
```

```
##              Class.ASD.Traits.
## Ethnicity      No  Yes
##      asian      0.29 0.71
##      black      0.26 0.74
##      Hispanic   0.25 0.75
##      Latino     0.23 0.77
##      middle eastern 0.49 0.51
##      mixed      0.38 0.62
##      Native Indian 0.00 1.00
##      Others     0.17 0.83
##      Pacifica   0.12 0.88
##      south asian 0.38 0.62
##      White European 0.25 0.75
```

```
xtabs(~Jaundice + Class.ASD.Traits.)
```

```
##              Class.ASD.Traits.
## Jaundice      No  Yes
##      no      253 513
##      yes      73 215
```

```
round(prop.table(xtabs(~Jaundice + Class.ASD.Traits.), 1), 2)
```

```
##              Class.ASD.Traits.
## Jaundice      No  Yes
##      no      0.33 0.67
##      yes     0.25 0.75
```

```
xtabs(~Family_mem_with_ASD + Class.ASD.Traits.)
```

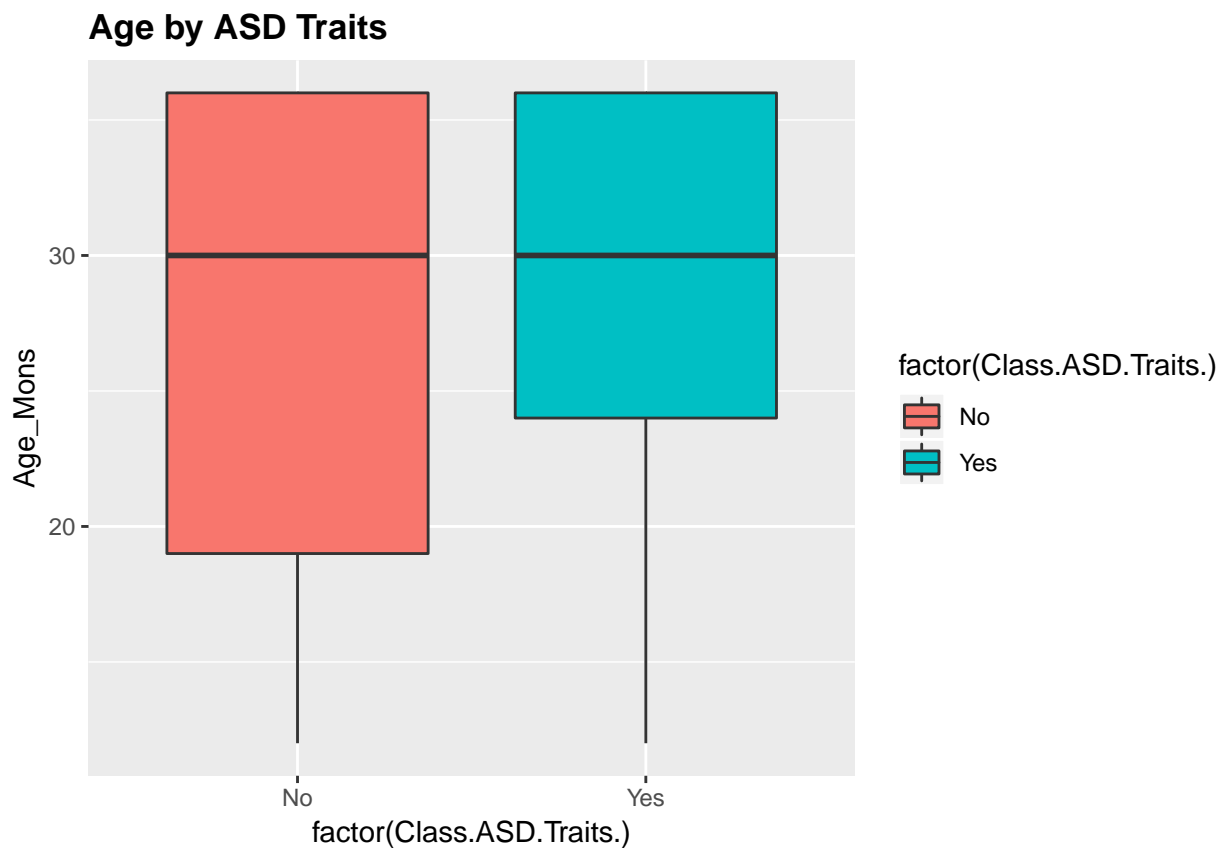
```
##              Class.ASD.Traits.
## Family_mem_with_ASD No  Yes
##      no      271 613
##      yes      55 115
```

```
round(prop.table(xtabs(~Family_mem_with_ASD + Class.ASD.Traits.),
1), 2)
```

```
##              Class.ASD.Traits.
## Family_mem_with_ASD No  Yes
##      no      0.31 0.69
##      yes     0.32 0.68
```

```
# Distribution of the Toddlers' Age by ASD Traits
```

```
ggplot(autism, aes(factor(Class.ASD.Traits.), Age_Mons)) + geom_boxplot(aes(fill = factor(Class.ASD.Traits.))) +
  ggtitle("Age by ASD Traits") + theme(plot.title = element_text(lineheight = 1,
    face = "bold"))
```



Interactions between explanatory variables are needed when the effect of one explanatory variable on the probability of success depends on the value for another explanatory variable. From these graphs, interactions among some of the explanatory variables seem to be warranted.

In R, there are several ways to implement interaction terms in a logistic regression model:

- `formula = y ~ x1 + x2 + x1:x2`
- `formula = y ~ x1*x2`
- `formula = y ~ (x1 + x2)^2`

To include a quadratic term in a logistic regression model, use the following

- `formula = y ~ x1 + I(x1^2)`

For factor variables, either turn them into factor variables and enter them into a logistic regression model, which is my preferred method, or use the `factor()` function inside a logistic regression: `formula = y ~ x1 + factor(x2)`, if `x2` needs to enter the regression as a factor variable.

Binary Logistic Regression Modeling

```
# Model 1 (Base Model)
mod.glm1 <- glm(Class.ASD.Traits. ~ Age_Mons + Sex + Ethnicity +
```

```

Jaundice + Family_mem_with_ASD, family = "binomial", data = autism)
summary(mod.glm1)

##
## Call:
## glm(formula = Class.ASD.Traits. ~ Age_Mons + Sex + Ethnicity +
##     Jaundice + Family_mem_with_ASD, family = "binomial", data = autism)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0941  -1.2292   0.7224   0.8317   1.4565
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.281885   0.304062   0.927  0.35389
## Age_Mons         0.008482   0.008859   0.957  0.33834
## Sexm             0.448503   0.147869   3.033  0.00242 **
## Ethnicityblack    0.177787   0.342323   0.519  0.60351
## EthnicityHispanic 0.247918   0.392078   0.632  0.52718
## EthnicityLatino    0.423043   0.488840   0.865  0.38682
## Ethnicitymiddle eastern -0.768519  0.202754 -3.790  0.00015 ***
## Ethnicitymixed     -0.330292  0.746778  -0.442  0.65828
## EthnicityNative Indian 13.396350 508.130189  0.026  0.97897
## EthnicityOthers     0.696988  0.471729   1.478  0.13954
## EthnicityPacifica    1.080882  1.079130   1.002  0.31653
## Ethnicitysouth asian -0.425479  0.299045  -1.423  0.15480
## EthnicityWhite European 0.207318  0.186398   1.112  0.26604
## Jaundiceyes         0.353792  0.163423   2.165  0.03040 *
## Family_mem_with_ASDyes -0.276408  0.187558  -1.474  0.14056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1303.9  on 1053  degrees of freedom
## Residual deviance: 1243.4  on 1039  degrees of freedom
## AIC: 1273.4
##
## Number of Fisher Scoring iterations: 13

# Model 2 (Model with both interaction and Non-linear effect)
# interaction of age and sex a quadratic term on age

# mod.glm2 <- glm(YOUR FORMULAR HERE, family = 'binomial',
# data = autism) summary(mod.glm2)

# Display the models together stargazer(mod.glm1, mod.glm2,
# type = 'text')

```

Testing Model Differences

```
# CODE HERE (1 line)
```

Based on the test result, we will use `mod.glm2`.

Our model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \text{Age_Mons} + \beta_2 \text{Age_Mons}^2 + \beta_3 \text{Sexm} + \beta_4 \text{Ethnicityblack} + \beta_5 \text{EthnicityHispanic} + \beta_6 \text{EthnicityLatino} +$$

The odds ratio for an increase in age by c months is expressed in the following formula:

$$OR = \exp(c\beta_1 + c\beta_2(2 \times \text{age} + c))$$

which depends on the level of age.

Model Interpretation

We need some questions, such as

- What is the effect of being a 30-month old boy on the odds of having ADS traits?

```
c = 1
Age_Mons = 30
# YOUR CODE HERE (1 line)
```

- What is the effect of an one month increase in age (measured in months) of a 24 months old female toddler on the odds of having ADS traits? (Hint: use the formula above.)

```
c = 1
Age_Mons = 24
Sexm = 0
# YOUR CODE HERE (1 line)
```