# Test 2

Brea Koenes

November 05, 2021

```
letsMove <- readr::read_csv("https://sldr.netlify.app/data/lets-move.csv")
```

## Research Question

I am looking to explore if the rate at which kids picked candy as their treat changes as age increases.

## Rationale

My response variable is the rate at which kids picked candy as their treat because that is what was asked of me to predict. My predictors are year, age, male, and obama. I chose Obama because whether the kid selected their treat from the side of the porch with a photo of Michelle Obama (1), or the "no Obama photo" side of the porch (0) may be correlated to whether they chose fruit or candy. I chose age as a predictor because the age of the child may impact what they selected. Specifically, I wondered if kids would choose fruit over candy as they got older. Also, it is a primary part of my research question. Gender may also predict whether a child chooses candy or fruit, so I am including gender as a predictor. Maybe boys gravitate towards candy more than girls, or vise versa. Lastly, the kid's age may vary across the years that the study was collected and whether children chose candy or not may vary over time as well. Becuase of its impact on the response varible and other factors, I also found it to be an important predictor to include.

In addition, the size of the data set is relatively small compared to most data sets (at least that I have worked with). Model accuracy increases as the number of observations increase, so I want to include all of the variables that are relevant to my research question in order to get the most information for my model.

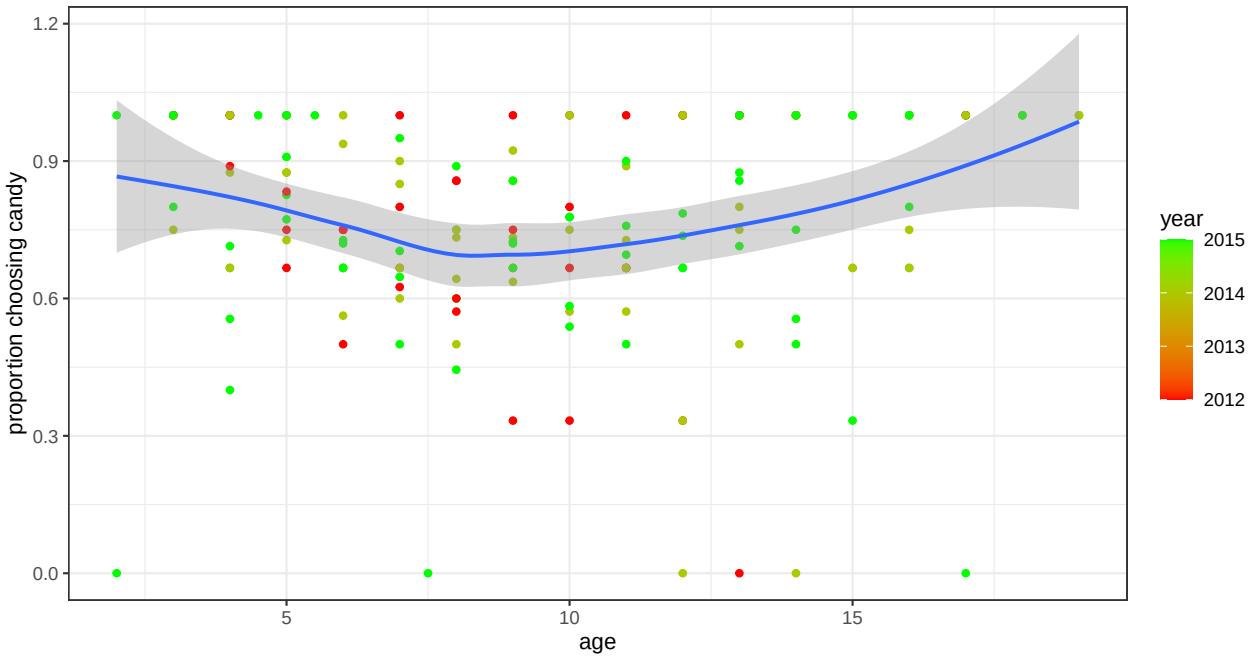
## Data exploration/graphics

```
letsMove <- letsMove %>%
  mutate(obama = ifelse(obama == 1, 'Yes', 'No'))

letsMove <- letsMove %>%
  mutate(male = ifelse(male == 1, 'Male', 'Female'))

letsMove <- letsMove %>%
  na.omit()
```

```
gf_point((chose_candy / (chose_fruit + chose_candy)~ age) , color =~ year, data = letsMove) %>%
  gf_refine(scale_colour_gradientn(colors = c('red', 'green'))) %>%
  gf_labs(y='proportion choosing candy') +
  geom_smooth()
```




The proportion of kids that pick candy decreases from 2 until almost 8 years old, then the proportion that chose candy increases with age. When it comes to year, there are no noticeable trends with 2014 and 2015. With 2012, it appears as if the proportion of kids that chose candy were younger on average than 2014 and 2015's ages.

In reference to my research question, it appears as if kids choose more candy as they get older past the age of 8.

## Model Rationale

I chose binary regression because of the binary nature of the response variable; the child either chooses fruit or candy. I used the logit link function because it allows me to evaluate the results using logarithms, as seen in the equation below. In addition, logit enables me to do exact logistical regression with my model, which is a quality unique to the logit function.

## Binary Regression Model

```
letsMove_logit <- glm(cbind(chose_candy, chose_fruit) ~ obama + male + year + age,
data = letsMove,
family = binomial(link = 'logit'))

msummary(letsMove_logit)
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 206.41822  141.44910   1.459   0.1445
## obamaYes     -0.24008    0.13838  -1.735   0.0827 .
## maleMale      0.14847    0.13281   1.118   0.2636
## year         -0.10188    0.07022  -1.451   0.1468
## age          -0.01086    0.02143  -0.507   0.6124
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 186.67  on 166  degrees of freedom
## Residual deviance: 180.38  on 162  degrees of freedom
## AIC: 436.5
##
## Number of Fisher Scoring iterations: 4
```

model equation:

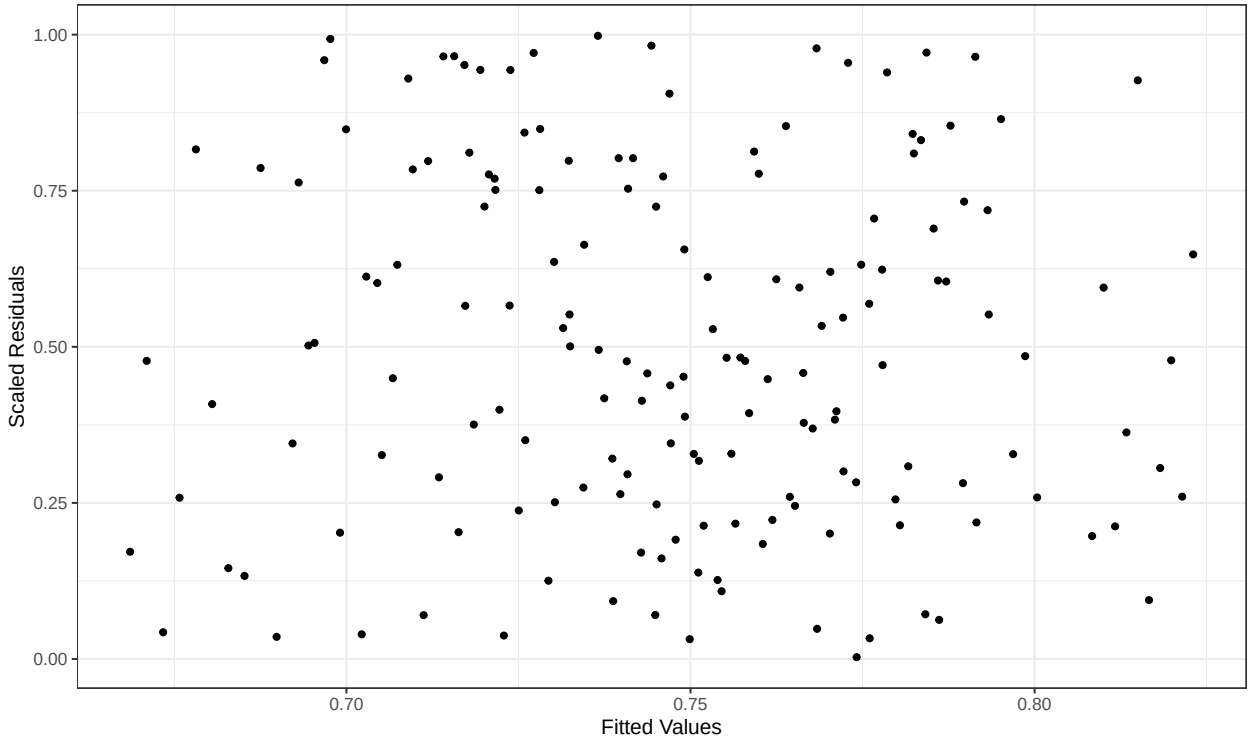
$$log(p_i/1 - p_i) = 206.41822 - 0.24008I_{obama} + 0.14847I_{male} - 0.10188x_{year} - 0.01086x_{age} + E$$

$p_i$ is the probability that the child chose candy. $x_{year}$ is the year. $x_{age}$ is the age. $I_{obamaYes}$ is 1 if the child got candy from the side with Michelle Obama's picture on it and 0 if the child went to the side with no picture. $I_{maleMale}$ is 1 if the child is a male and 0 if the child is a female.

## Check conditions

```
require(DHARMa)
letsMove_sim <- simulateResiduals(letsMove_logit)

gf_point(letsMove_sim$scaledResiduals ~ fitted(letsMove_logit)) %>%
  gf_labs(x = 'Fitted Values',
          y = 'Scaled Residuals')
```




This fitted vs. residuals plot checks the lack of non-linearity and error variance. There isn't a clear linear trend in the data. However, it is okay if there is not a distinct trend; it still passes this test.

In addition, the width of distribution is not changing much. The points form a rectangle and not a trumpet shape, so the error variance is constant. It passes this test as well.

## Model Selection: Hypothesis testing

```
car::Anova(letsMove_logit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(chose_candy, chose_fruit)
##       LR Chisq Df Pr(>Chisq)
## obama  2.98891  1    0.08384 .
## male   1.24961  1    0.26363
## year   2.15612  1    0.14200
## age    0.25603  1    0.61286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

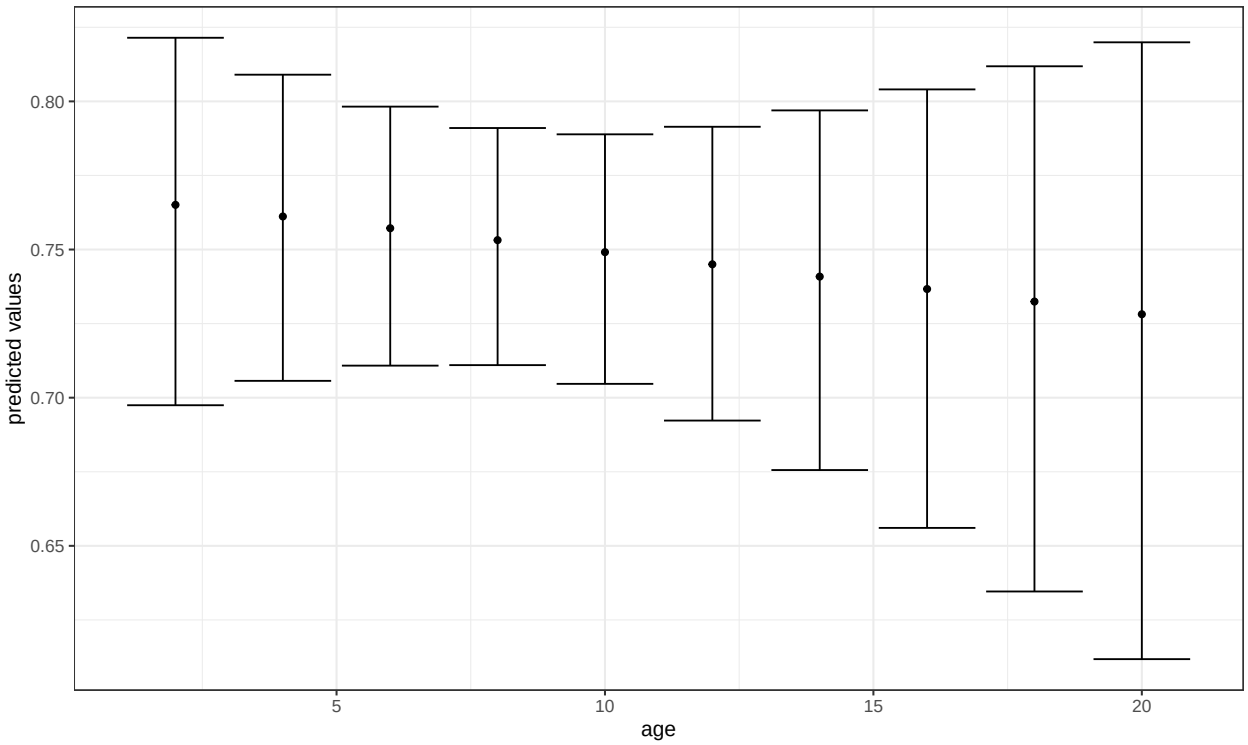
I chose hypothesis testing in order to see if my predictors are really associated with the response. Anova tests null hypothesis for each variable to see if it is not associated with the response. We look at the p-values to evaluate the tests. A general rule is that a p-value less than 0.05 is statistically significant, while a p-value higher than 0.05 indicates strong evidence for the null hypothesis.

Looking at the p-values, none of my predictors appear to have any statistical significance when it comes to the response. More analysis on these values is given in my conclusion.

## Prediction plot

```
library(ggeffects)
pred_data <- ggpredict(letsMove_logit, 'age')

gf_point(predicted ~ x,
         data  = pred_data) %>%
  gf_errorbar(conf.low + conf.high ~ x) %>%
  gf_labs(x = 'age', y = 'predicted values')
```




I did not include year, obama, or male as predictors. Instead, they are instead included in the my letsMove_logit model, which is used in the formula above to create the predicted values.

As apparent in the graph above, the proportion of kids that chose candy in the predicted values decreases as age increases (on average). However, the range of predicted values also generally increases as age increases.

## Interpretation and conclusion

In conclusion, it appears as if the rate at which kids picked candy as their treat decreases as age increases. However, I would not say that the claim that they are associated is reliable. Here is why:

1) My prediction plot supports my conclusion that as age increases, the rate at which kids choose candy decreases. As seen in the prediction plot above, there is a steady decrease from 0.77 to 0.726. This decrease is constant and noticeable.

2) In my model assessment, Obama has the most significance, however the p-value is not low enough to claim that it has an association with the response. When it comes to my main variable of interest, age, it most definitely indicates evidence for the null hypothesis. This is what causes me to doubt my model's predictor's association with the response. With p-values over 0.05 in the hypothesis test, the predictors are not associated with the response. Therefore, it is hard to claim that age and choosing candy are associated.

3) Despite my model not being able to claim an association between age and choosing candy, I did make the correct model selection. Due to my binary response variable with multiple kids per row, the natural choice was a binary model.