# Logistic Regression

## Comparing two logistic models

**Jenine Harris**
**Brown School**

# Importing and cleaning the data

```r
# import the libraries cleaned file
libraries <- read.csv("/Users/harrisj/Box/teaching/Teaching/Fall2020/dat

# change data types
library(package = "tidyverse")
libraries.cleaned <- libraries %>%
  mutate(age = as.numeric(age))
```

# The models

```
# large model
lib.model <- glm(formula = uses.lib ~ age + sex + educ + parent + disabl
                 data = libraries.cleaned,
                 na.action = na.exclude,
                 family = binomial("logit"))
```

```
# large model with interaction
lib.model.int <- glm(formula = uses.lib ~ age + sex + educ + parent + di
                 data = libraries.cleaned,
                 family = binomial("logit"))
```

# Using the likelihood ratio (LR) test to compare two nested models

- The LR test compares two **nested** models where one model includes a subset of the variables in the other model.
- For example, the simple logistic regression model with age as the only predictor could be compared statistically to any of the larger models because they all have the age variable in them.
- The small model is nested in each of the larger models.
- In addition, the larger model without the interaction could be compared to the model with the interaction term.
- Models where the variables of one are completely different from the variables in the other cannot be compared with this test.

# Theory behind the test

- The idea behind the LR test is to determine if the additional variables in a model make the model *better enough* to warrant the complexity of adding more variables to the model.

- The lmtest package has the `lrtest()` function, which can be used to compare two nested models.

- The LR test computes the difference between the log-likelihoods for the two models and multiplies this by two; the result has a chi-squared distribution.

# NHST Step 1: Write the null and alternate hypotheses

H0: The larger model with the interaction term is the same at explaining library use compared to the model without the interaction term.

HA: The interaction term model is better than the smaller model at explaining library use.

# NHST Step 2: Compute the test statistic

- It does not matter which model is listed as the first argument and which is listed second.

```
# compare simple logistic with age to
# full library use model
lmtest::lrtest(object = lib.model, lib.model.int)
```

```
## Likelihood ratio test
##
## Model 1: uses.lib ~ age + sex + educ + parent + disabled + rurality +
##     raceth + ses
## Model 2: uses.lib ~ age + sex + educ + parent + disabled + rurality +
##     ses + raceth + sex * parent
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  13 -941.32
## 2  14 -940.54  1 1.5599     0.2117
```

The output from `lrtest()` shows the test statistic is $\chi^2 = 1.56$.

# NHST Step 3: Compute the probability of the test statistic

The probability of the test statistic was included in the output from the `lrtest()` command. The test statistic of $\chi^2 = 1.56$ had a p-value of .21.

# NHST Steps 4 & 5: Make a decision and write a conclusion

The null hypothesis was retained; the model with the interaction term was no different in explaining library use from the model without the interaction term ( $\chi^2 = 1.56$; p = .21).

- When the larger model is not statistically significantly better, it is often preferred to use the smaller model to aid in interpretation.
- The more complex a model becomes, the more difficult it is to interpret.
- Generally speaking, parsimony is preferable.
- However, there are exceptions to this when the larger model has variables in it that have been consistently related to the outcome in other research or are important to understanding the outcome for some other reason.

# Complete interpretation of final model

A logistic regression model with age, sex, education, parent status, socioeconomic status, race-ethnicity, and disability status was statistically significantly better than a baseline model at explaining library use [ $\chi^2$ (12) = 94.74; p < .001]. A likelihood ratio test comparing this model to a model that also included an interaction between sex and parent status showed that the larger model was not statistically significantly better than the smaller model [ $\chi^2$ (1) = 1.56; p = .21], so the smaller model was retained. The odds of library use were 51% lower for males compared to females (OR = .49; 95% CI: .39 - .61). The odds of library use were 1.90 times higher for those with a four-year degree compared to those with less than a high school education (OR = 1.90; 95% CI: 1.26 - 2.90). The odds of library use are 1.55 times higher for non-Hispanic Black participants compared to Hispanic participants (OR = 1.55; 95% CI: 1.002 - 2.42). The odds of library use are 1% lower for every one year increase in a person's age (OR = .99; 95% CI: .985 - .997). The odds of library use are not statistically significantly different for urban or suburban residents compared to rural residents, for parents compared to non-parents, for non-Hispanic Whites compared to Hispanics, or for people with low or medium SES compared to high SES. Assumption checking revealed a possible problem with the linearity of the age predictor, especially at the youngest ages. The other assumptions were met. Diagnostics found two problematic outlying or influential observations, but the observations did not appear to be data entry mistakes or much different from the rest of the sample in any way.