# Logistic Regression

## Model fit

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

# Importing and cleaning the data

```
# 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))
```

# Computing and interpreting two measures of model fit

- For linear regression, the $R^2$ statistic measured how well the model fit the observed data by measuring how much of the variability in the outcome was explained by the model.
- The concept of variance is appropriate for continuous but not categorical variables, so a different measure of model fit is needed.
- There are several to choose from including the **percent correctly predicted** which is sometimes referred to as the **Count** $R^2$.

# Percent correctly predicted or Count $R^2$

- The percent correctly predicted by the model was computed using the predicted probabilities, or fitted values, for each of the observations and comparing these probabilities to the true value of the outcome.
- For example, if a person in the data set were predicted to have a 56% chance of library use, this would be transformed into a "yes" or "1" value of the outcome.
- This would then be compared to the person's actual library use. If the predicted value and the true value matched, this is considered a correct prediction.
- Likewise, if the predicted probability for library use is less than 50%, the person is considered to be a "no" or "0" for library use. C
- omparing this to the true value for that person would indicates whether the model was correct or incorrect.
- The total number of people the model gets correct out of the total number of people in the data set is the *percent correctly predicted* or *Count* $R^2$.

# Getting percent correctly predicted from R

- The `odds.n.ends()` command includes a table showing how many observations were correctly predicted in each category of the outcome.
- These values can be used to determine the overall percent correctly predicted.
- The model correctly predicted 338 of those who used the library and 500 of those who do not use the library.
- Compute the overall percent correctly predicted of 838 / 1571 or 53%.

```
## $`Logistic regression model significance`
## Chi-squared        d.f.           p
##      10.815       1.000       0.001
##
## $`Contingency tables (model fit): percent predicted`
##                   Percent observed
## Percent predicted         1         0       Sum
##               1   0.2151496 0.1896881 0.4048377
##               0   0.2768937 0.3182686 0.5951623
##               Sum 0.4920433 0.5079567 1.0000000
##
## $`Contingency tables (model fit): frequency predicted`
##                  Number observed
## Number predicted    1    0  Sum
```

# Adjusted Count $R^2$

- One alternative to the percent correctly predicted is the Adjusted Count $R^2$ , which adjusts the Count $R^2$ for the number of people in the largest of the two categories of the outcome.
- The argument behind this adjustment is that a null model, or a model with no predictors, could get a good percent correctly predicted just by predicting everyone was in the outcome category that had the bigger percentage of people in it.
    - $R^2_{count.adj} = \frac{n_{correct} - n_{most.common.outcome}}{n_{total} - n_{most.common.outcome}}$
- For the library use data, the most common category is library non-use (or 0) with 798 of the 1571 participants with complete data for the model.
- Without knowing anything about library use, you could predict everyone in the data set was a non-user and be right $\frac{798}{1571}$ or 50.8% of the time.
- Using the age predictor, the model is right $\frac{338+500}{1571}$ or 53% of the time.
- While this is not a huge increase, it did classify 40 additional people correctly compared to using the percentages in the outcome categories with no other information.
- Interpreting this statistic is pretty straightforward and could go something like this:
    - The model using age to predict library use was correct 53% of the time (Count $R^2$ = .53).

# Sensitivity and specificity

- Sometimes it is useful to know whether the model is better at predicting people with the outcome or people without the outcome.
- The measures used for these two concepts are sensitivity and specificity.
- Sensitivity determines the percentage of the 1s or "yes" values the model got correct, while specificity computes the percentage of 0s or "no" values the model got correct.
- In this case, the sensitivity is 43.7% while the specificity is 62.7%.
- The model was better at predicting the no values than the yes values.
- These percentages could also be computed from the frequency table in the output, the model predicted 500 of the 798 people in the 0 category correctly (62.7%) and 338 of the 773 in the 1 category correctly (43.7%).