Logistic Regression

Adding an interaction

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))
```

Larger model

```
# run the odds.n.ends code again
lib.model <- glm(formula = uses.lib ~ age + sex + educ + parent + disabl
                 data = libraries.cleaned,
                 na.action = na.exclude,
                 family = binomial("logit"))
```

Adding and interpreting interaction terms in logistic regression

- Perhaps sex and parent status might work together to influence the odds of library use.
- An interaction term examines how two (or more) variables might work together to influence an outcome.

```
# the relationship between parent status and library use
libraries.cleaned %>%
  drop_na(parent) %>%
  ggplot(aes(x = parent, fill = factor(uses.lib))) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(x = "Parent status", y = "Number of participants",
       title = "Parent status and library use from 2016\nPew Research Ce
  scale_fill_manual(values=c("#7463AC", "gray"),
                    name="Library use")
```

Adding and interpreting interaction terms in logistic regression

- Perhaps sex and parent status might work together to influence the odds of library use.
- An interaction term examines how two (or more) variables might work together to influence an outcome.

Plotting the second variable

- It looks like there are more parents who are library users than there are parents who are non-users, while fewer non-parents are users than non-users of the library.

```
# library use by sex
libraries.cleaned %>%
  drop_na(parent) %>%
  ggplot(aes(x = sex, fill = factor(uses.lib))) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(x = "Sex", y = "Number of participants",
       title = "Library use by sex from 2016 Pew\nResearch Center survey
  scale_fill_manual(values=c("#7463AC", "gray"),
                    name="Library use")
```

Plotting the variables together

```
# the relationship between parent status and library use
libraries.cleaned %>%
  drop_na(parent) %>%
  ggplot(aes(x = parent, fill = factor(uses.lib))) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(x = "Parent status", y = "Number of participants") +
  scale_fill_manual(values=c("#7463AC", "gray"),
                    name="Library use") +
  facet_grid("sex")
```

Interpreting the interaction

- It does look like there is a difference between males and females who are parents and non-parents.
- For females, both parents and non-parents are more often library users than they are non-users.
- For males, non-parents are not library users and parents are almost equally library users and non-users.
- This suggests that sex and parental status **interact** to influence library use.
- That is, the two characteristics work together to influence the outcome.

Including the interaction in the model

- This possible interaction between sex and parental status can be included in the logistic regression model by adding a term, `+ sex*parent` to the formula.
- When an interaction is included in a model, it is customary to also include the interacting variables separately.
- In a model with interaction terms, the terms that are not part of the interaction are called *main effects*.
- Start by writing out the model:
  - $p(uses.lib) = \frac{1}{1+e^{-(b_0+b_1 \cdot age+b_2 \cdot sex+b_3 \cdot educ+b_4 \cdot parent+b_5 \cdot disabled+b_6 \cdot rurality+b_7 \cdot ses+b_8 \cdot raceth+b_9 \cdot sex \cdot parent)}}$

Estimating the model in R

Now that she had the model in mind, use `glm()` to estimate it.

```
# estimate the library use model and print results
lib.model.int <- glm(formula = uses.lib ~ age + sex + educ + parent + di
                     data = libraries.cleaned,
                     family = binomial("logit"))
odds.n.ends::odds.n.ends(x = lib.model.int)
```

```
## $`Logistic regression model significance`
## Chi-squared        d.f.           p
##      96.296      13.000       0.000
##
## $`Contingency tables (model fit): percent predicted`
##                    Percent observed
## Percent predicted         1         0       Sum
##               1   0.2683952 0.1850035 0.4533987
##               0   0.2193413 0.3272600 0.5466013
##               Sum 0.4877365 0.5122635 1.0000000
##
## $`Contingency tables (model fit): frequency predicted`
##                     Number observed
## Number predicted    1    0  Sum
##               1     383  264  647
##               0     313  467  780
##               Sum   696  731 1427
##
```

NHST Step 1: Write the null and alternate hypotheses

H0: A model including age, sex, education, parent status, disability status, rurality, SES, race-ethnicity, and an interaction between sex and parent status is not useful in explaining library use.

HA: A model including age, sex, education, parent status, disability status, rurality, SES, race-ethnicity, and an interaction between sex and parent status is useful in explaining library use.

NHST Step 2: Compute the test statistic

The test statistic is chi-squared = 96.30 with 13 degrees of freedom.

NHST Step 3: Compute the probability for the test statistic (p-value)

The p-value is less than .001.

NHST Steps 4 & 5: Interpret the probability and write a conclusion

The model including age, sex, education, parent status, disability status, rurality, SES, race-ethnicity, and an interaction between sex and parent status is useful in explaining library use [χ^2 (13) = 96.30; p < .001].

Compute and interpret odds ratios

- Age, sex, having a four-year degree or more, and non-Hispanic Black race-ethnicity were statistically significant predictors of library use.
- For every one year increase in age, the odds of library use decreased by 1% (OR = .99; 95% CI: .983 - .996).
- Males have 55% lower odds of library use compared to females (OR = .45; 95% CI: .35 - .58) and those with a four-year degree have 1.93 times higher odds of library use compared to those with less than a high school education (OR = 1.93; 95% CI: 1.28 - 2.94).
- Finally, non-Hispanic Black participants had 1.56 times higher odds of library use compared to Hispanic participants.
- Disability status, SES, and rurality were not significantly associated with library use and those with between high school and a 2-year degree had odds no higher nor lower odds of library use than those with less education.
- Likewise, non-Hispanic White participants had no higher or lower odds of library use compared to Hispanic participants.
- There was no statistically significant interaction between sex and parent status on the odds of library use.

Compute and interpret model fit

- The model correctly predicted 383 of 696 of those who use the library and 467 of 731 of those who do not use the library.
- Overall the model correctly predicted 850 of 1427 observations (59.6%); the model was more specific (63.9%) than sensitive (55.0%), indicating that it is better at classifying non-library users than library users.

Check assumptions: Independence of observations

The data source is the same as for the previous analyses; this assumption is met.

No multicollinearity

```
# compute GVIF
car::vif(lib.model.int)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## age         1.272824  1         1.128195
## sex         1.394525  1         1.180900
## educ        1.317568  2         1.071379
## parent      2.373418  1         1.540590
## disabled    1.155785  1         1.075074
## rurality    1.122362  2         1.029279
## ses         1.255048  2         1.058437
## raceth      1.214412  2         1.049764
## sex:parent 2.570189  1         1.603181
```

None of the values in the right hand column were greater than 2; this assumption is met.

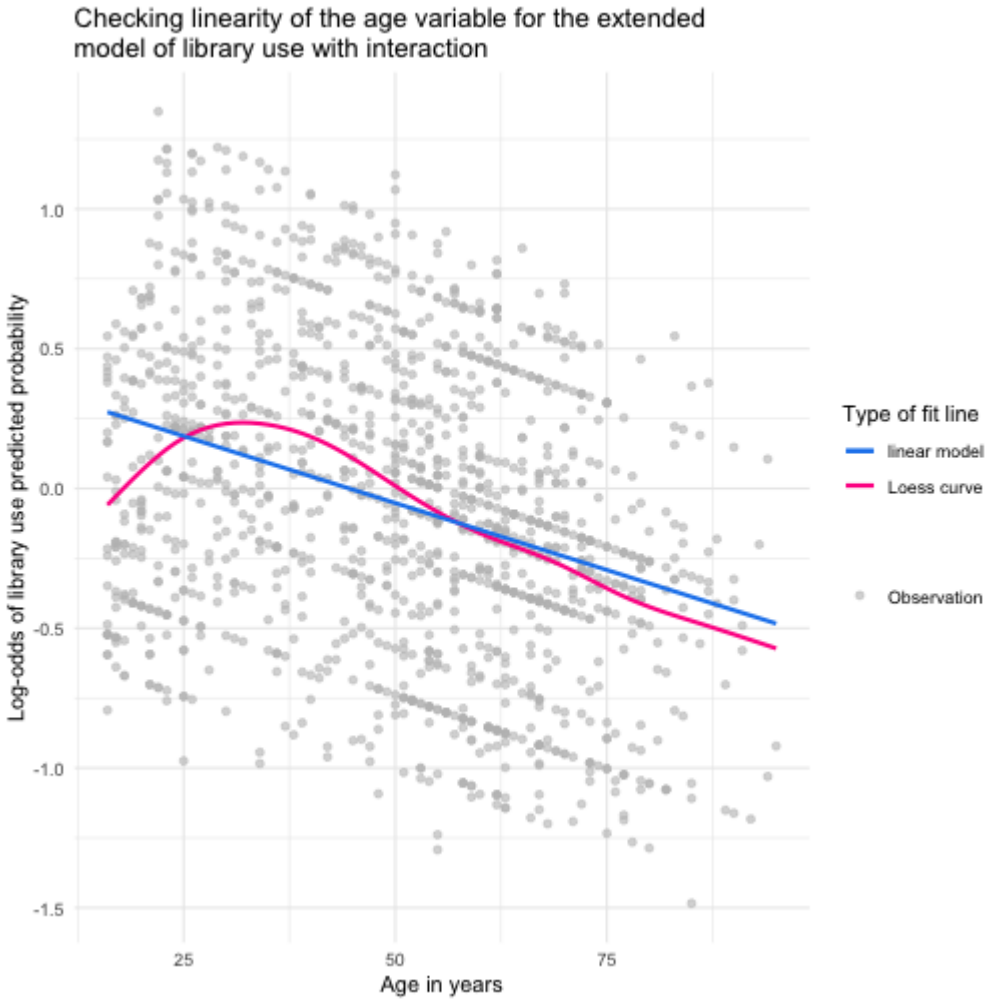
Linearity

```
# make a variable of the logit of the outcome
logit.use.int <- log(lib.model.int$fitted.values/(1-lib.model.int$fitted

# make a small data frame with the logit variable and the age predictor
linearity.data.int <- data.frame(logit.use.int, age.int = lib.model.int$

# create a plot with linear and actual relationships shown
linearity.data.int %>%
  ggplot(aes(x = age.int, y = logit.use.int))+
  geom_point(aes(size = "Observation"), color = "gray", alpha = .6) +
  geom_smooth(se = FALSE, aes(color = "Loess curve")) +
  geom_smooth(method = lm, se = FALSE, aes(color = "linear model")) +
  scale_color_manual(name="Type of fit line", values=c("dodgerblue2", "d
  scale_size_manual(values = 1.5, name = "") +
  theme_minimal() +
  labs(x = "Age in years", y = "Log-odds of library use predicted probab
       title = "Checking linearity of the age variable for the extended\
```

Linearity


Checking linearity of the age variable for the extended model of library use with interaction
Log-odds of library use predicted probability
1.0
0.5
0.0
-0.5
-1.0
-1.5
25
50
75
Age in years
Type of fit line
linear model
Loess curve
Observation

Interpretation

- The linearity concerns are similar to the prior model; there is a large deviation from linearity at the younger end of the age range, and some more minor deviations throughout.
- Given that the interaction term was not statistically significant and the model violated the linearity assumption, it seems preferable to report the previous model without the interaction term as the final model.
- However, there is a statistical test that can be used to determine whether a larger model is statistically significantly better than a smaller model. The test is called the Likelihood Ratio (LR) test.