ALM week 9: logistic regression

Schedule for today

- Discussion activity
- Logistic regression workshop
  - *Slides and workshop to follow along are in GitHub*
  - *pew_libraries_2016_cleaned_ch10.csv in GitHub*
    - These data are from the Pew Internet & American Life website, which has great publicly available data sources
- New R packages (all optional today)
  - *sjPlot*
  - *sjmisc*
  - *sjlabelled*

What predicts library use?

The logistic regression process:

- Exploratory data analysis
- Assumption checking
- Model development
- Model diagnostics
- Model reporting

Import the data

```
# import the libraries data file
libraries <- read.csv("pew_libraries_2016_cleaned_ch10.csv")

# check the data
summary(object = libraries)
```

```
##       age            sex              parent       disabled    uses.lib
##  Min.   :16.00   female:768   not parent:1205   no  :1340   no :809
##  1st Qu.:33.00   male  :833   parent    : 391   yes : 253   yes:792
##  Median :51.00                NA's      :   5   NA's:   8
##  Mean   :49.31
##  3rd Qu.:64.00
##  Max.   :95.00
##  NA's   :30
##      ses                     raceth                             educ
##  high  : 158   Hispanic          : 194   < HS                       :171
##  low   : 246   Non-Hispanic Black: 170   Four-year degree or more:658
##  medium:1197   Non-Hispanic White:1097   HS to 2-year degree     :772
##                NA's              : 140
##
##
##
##     rurality
##  rural   :879
##  suburban:355
##  urban   :353
##  NA's    : 14
##
##
##
```

Codebook

The variables are:

- **age:** age in years
- **sex:** sex (female/male)
- **rurality:** lives in rural/suburban/urban area
- **disabled:** has a disability (yes/no)
- **uses.lib:** has visited a public library in the last year (yes/no)
- **ses:** socioeconomic status (high/medium/low)
- **raceth:** race and ethnicity (Hispanic, Non-Hispanic Black, Non-Hispanic White)
- **educ:** highest education completed (< HS, HS to 2-year degree, Four-year degree or more)
- **parent:** parent status (parent/not parent)

Analysis plan

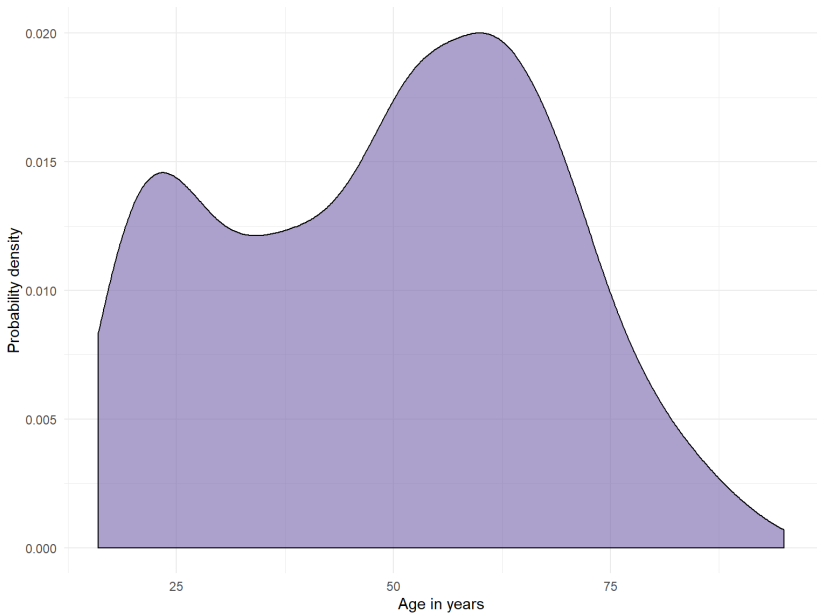
- There is published evidence that library use varies by age, sex, race-ethnicity, education, SES, and rurality
- There is not much evidence related to disability or parent status and library use
- Use the Pew Internet & American Life data to explain or predict library use based on age, sex, race-ethnicity, education, SES, rurality, disability, and parent status

Exploratory data analysis: examine distributions

- Check the distribution of any continuous predictors

```
# open tidyverse
library(package = "tidyverse")

# examine the distribution of age
libraries %>%
  ggplot(aes(x = age)) +
  geom_density(fill = "#7463AC", alpha = .6) +
  theme_minimal() +
  labs(y = "Probability density", x = "Age in years")
```



Exploratory data analysis: make a table

```
# open tableone
library(package = "tableone")
# create the table
table.desc <- CreateTableOne(data = libraries,
                             strata = 'uses.lib',
                             vars = c("age", "sex", "parent", "disabled",
                                      "ses", "raceth", "educ", "rurality"))
print(table.desc,
      nonnormal = 'age',
      showAllLevels = TRUE)
```

Exploratory data analysis: make a table

Stratified by uses.lib

	level	no	yes	p	test
n		809	792		
age (median [IQR])		53.00 [35.00, 65.00]	49.00 [31.00, 62.00]	0.001	nonnorm
sex (%)	female	330 (40.8)	438 (55.3)	<0.001	
	male	479 (59.2)	354 (44.7)		
parent (%)	not parent	639 (79.1)	566 (71.8)	0.001	
	parent	169 (20.9)	222 (28.2)		
disabled (%)	no	661 (82.0)	679 (86.3)	0.024	
	yes	145 (18.0)	108 (13.7)		
ses (%)	high	67 (8.3)	91 (11.5)	0.088	
	low	130 (16.1)	116 (14.6)		
	medium	612 (75.6)	585 (73.9)		
raceth (%)	Hispanic	111 (14.9)	83 (11.6)	0.110	
	Non-Hispanic Black	79 (10.6)	91 (12.7)		
	Non-Hispanic White	557 (74.6)	540 (75.6)		
educ (%)	< HS	102 (12.6)	69 (8.7)	<0.001	
	Four-year degree or more	276 (34.1)	382 (48.2)		
	HS to 2-year degree	431 (53.3)	341 (43.1)		
rurality (%)	rural	478 (59.7)	401 (51.0)	0.002	
	suburban	159 (19.9)	196 (24.9)		
	urban	164 (20.5)	189 (24.0)		

Assumption checking

- Independent observations (not tested, based on sampling)
- No multicollinearity
- Linearity of independent variables with the log-odds of the outcome

Assumption 1: independent observations

- Even if you did not collect the data yourself, you can still check this assumption by examining available information from the data collectors:
  - *These data are from Pew Internet & American Life*
  - *The website for the Pew Internet & American Life data describes the data collection process (https://www.pewinternet.org/2016/09/09/libraries-2016/)*

18
PEW RESEARCH CENTER

Methodology

The analysis in this report is based on a Pew Research Center survey conducted March 7-April 4, 2016, among a national sample of 1,601 adults, 16 years of age or older, living in all 50 U.S. states and the District of Columbia. Fully 401 respondents were interviewed on landline telephones, and 1,200 were interviewed on cellphones, including 667 who had no landline telephone. The survey was conducted by interviewers at Princeton Data Source under the direction of Princeton Survey Research Associates International. A combination of landline and cellphone random-digit-dial samples were used; both samples were provided by Survey Sampling International. Interviews were conducted in English and Spanish. Respondents in the landline sample were selected by randomly asking for the youngest adult male or female who was at home. Interviews in the cellphone sample were conducted with the person who answered the phone, if that person was 16 years of age or older. For detailed information about our survey methodology, visit: http://www.pewresearch.org/methodology/u-s-survey-research/

- The assumption is **met**

Assumption 2: No multicollinearity

- Multicollinearity is when variables are highly correlated with one another
- It is identified by the VIF in linear regression and the GVIF in logistic
  - *The GVIF, or generalized variance inflation factor, re-runs the model for each predictor as the outcome with the other predictors as the independent variables*
  - *If model fit is very high for these models, that means the predictors are strongly related to each other and do not all need to be in the model*
  - *Bottom line, if the GVIF score is too high, a variable is too strongly related to another variable and one of them should be removed*
  - *A variable should be removed if* $GVIF^{1/(2*df)}$ *> 4*

The logistic regression model is needed in order to check this assumption!

Clean the outcome for the regression model

```
# recode so yes = 1 and no = 0
libraries.cleaned <- libraries %>%
  mutate(uses.lib.num = recode(uses.lib,
                               `yes` = 1,
                               `no` = 0))

# check recoding
table(libraries.cleaned$uses.lib.num, libraries.cleaned$uses.lib)
```

```
##
##      no yes
##   0 809   0
##   1   0 792
```


Estimate the model and get GVIF

```
# predict library use
libUseModel <- glm(uses.lib.num ~ age + sex + parent +
                     disabled + ses + raceth + educ + rurality,
                   data = libraries.cleaned,
                   family = "binomial")

# compute GVIF for libUseModel
library(car)
vif(mod = libUseModel)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.254322  1        1.119965
## sex      1.051221  1        1.025291
## parent   1.101618  1        1.049580
## disabled 1.153173  1        1.073859
## ses      1.249162  2        1.057194
## raceth   1.212126  2        1.049269
## educ     1.309506  2        1.069737
## rurality 1.118617  2        1.028420
```

Check assumption 2: No multicollinearity

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.254322  1        1.119965
## sex      1.051221  1        1.025291
## parent   1.101618  1        1.049580
## disabled 1.153173  1        1.073859
## ses      1.249162  2        1.057194
## raceth   1.212126  2        1.049269
## educ     1.309506  2        1.069737
## rurality 1.118617  2        1.028420
```

- The GVIF values are near 1 and none of those in the $GVIF^{1/(2*df)}$ column are over 4, this assumption is **met**
  - *You may find different thresholds suggested by different people which is due to differing opinions about how much correlation among variables is too much*
  - *Thresholds recommended may range from 3 to 10, 4 and 5 are common*
  - *Choosing a lower threshold means that you are being more strict about how much correlation there can be among variables*

Check assumption 3: Linearity

- Examines whether there is a linear relationship between any **continuous** predictors and the log odds of the predicted values
  - *This shows whether the predicted values are equally accurate along the range of values of the predictor*
  - *In this case, is library use predicted equally accurately along the range of* ***ages****?*

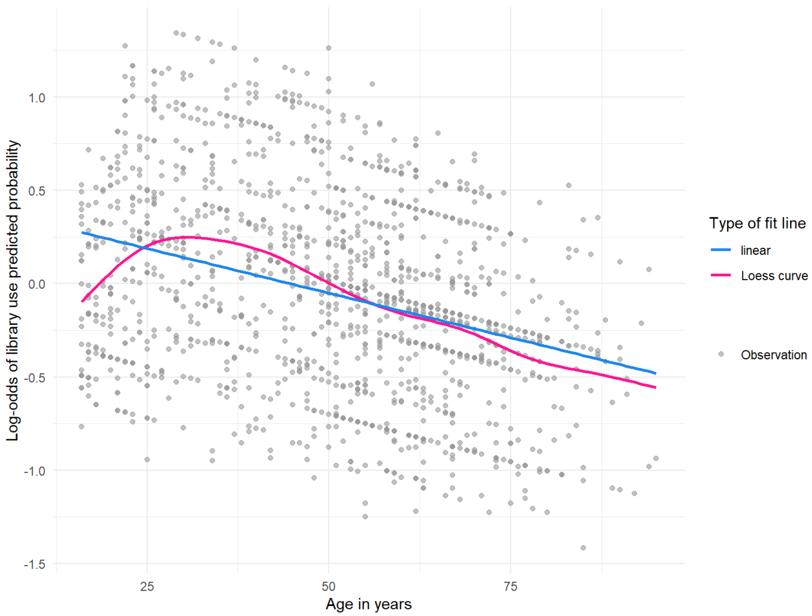
```
# make a variable of the logit of the predicted probabilities
logodds.use <- log(x = libUseModel$fitted.values/(1-libUseModel$fitted.values))

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

# create a plot with linear and actual relationships shown
linearity.data %>%
  ggplot(aes(x = age, y = logodds.use))+
  geom_point(aes(size = "Observation"), color = "gray60", alpha = .6) +
  geom_smooth(se = FALSE, aes(color = "Loess curve")) +
  geom_smooth(method = lm, se = FALSE, aes(color = "linear")) +
  theme_minimal() +
  labs(x = "Age in years", y = "Log-odds of library use predicted probability") +
  scale_color_manual(name="Type of fit line", values=c("dodgerblue2", "deeppink")) +
  scale_size_manual(values = 1.5, name = "")
```

Check assumption 3: Linearity

- Examines whether there is a linear relationship between any **continuous** predictors and the predicted probabilities
  - *This shows whether the predicted probabilities are equally accurate along the range of values of the predictor*
  - *In this case, is library use predicted accurately along the range of ages*



Check assumption 3: Linearity

- Some deviation at the very youngest ages, which go down to 16
- Might be better to limit age range to adults 18 and over?
- For now we will say close enough and **met**

Model development

- We already developed the model `libUseModel`

Model diagnostics

- Same diagnostics as for linear regression, just adjust the cutoff for leverage

```
# use same code as from linear regression
# change the cutoff for leverage to reflect the 13 parameters in the model
libraries.cleaned.diag <- libraries.cleaned %>%
  drop_na() %>%
  mutate(standardres = rstandard(model = libUseModel)) %>%
  mutate(cooks.dist = cooks.distance(model = libUseModel)) %>%
  mutate(lever = hatvalues(model = libUseModel)) %>%
  mutate(outlier.infl = as.numeric(x = lever > 2*13/n()) +
           as.numeric(x = cooks.dist > 4/n()) +
           as.numeric(x = abs(x = standardres) > 1.96))

# examine the outliers & influential
libraries.cleaned.diag %>%
  filter(outlier.infl >= 2)
```

```
##   age    sex     parent disabled uses.lib    ses             raceth educ
## 1  91 female not parent      yes      yes medium Non-Hispanic Black < HS
## 2  76   male not parent       no      yes    low Non-Hispanic Black < HS
##   rurality uses.lib.num standardres  cooks.dist      lever outlier.infl
## 1    rural            1    1.452884 0.002976640 0.02049212            2
## 2    rural            1    1.562863 0.003526649 0.01906898            2
```

Model reporting (finally!)

- For a logistic regression model reporting includes
  - *Odds ratios and confidence intervals*
  - *Model significance*
  - *Model fit*

Model results

```
# use odds.n.ends to get model results
library(package = "odds.n.ends")
odds.n.ends(libUseModel)
```

```
## $`Logistic regression model significance`
## Chi-squared        d.f.           p
##      94.736      12.000       0.000
##
## $`Contingency tables (model fit): percent predicted`
##                  Percent observed
## Percent predicted         1         0       Sum
##               1   0.2648914 0.1744919 0.4393833
##               0   0.2228451 0.3377715 0.5606167
##               Sum 0.4877365 0.5122635 1.0000000
##
## $`Contingency tables (model fit): frequency predicted`
##                 Number observed
## Number predicted    1    0  Sum
##              1    378  249  627
##              0    318  482  800
##              Sum  696  731 1427
##
## $`Predictor odds ratios and 95% CI`
##                                    OR     2.5 %    97.5 %
## (Intercept)                 1.3180091 0.6778733 2.5644803
## age                         0.9899123 0.9835415 0.9962814
## sexmale                     0.4891734 0.3921079 0.6091430
## parentparent                1.2652862 0.9710624 1.6500243
## disabledyes                 0.8003756 0.5836481 1.0949054
## seslow                      0.9323567 0.5720558 1.5162449
## sesmedium                   0.8471423 0.5747503 1.2441896
## racethNon-Hispanic Black    1.5539262 1.0018330 2.4167032
## racethNon-Hispanic White    1.3152888 0.9312650 1.8632329
## educFour-year degree or more 1.9040694 1.2584331 2.8953329
## educHS to 2-year degree     1.1475517 0.7808490 1.6947789
## ruralitysuburban            1.1899804 0.9019925 1.5704210
## ruralityurban               1.2300055 0.9281956 1.6307183
##
## $`Model sensitivity`
## [1] 0.5431034
##
## $`Model specificity`
## [1] 0.6593707
```

Statistically significant odds ratios

- Odds ratios with confidence intervals that do not include 1 are statistically significant
  - *age*
  - *sex*
  - *race-ethnicity*
  - *education*

Interpreting the continuous predictor odds ratios

- For continuous predictors, odds ratios show increase in odds of the outcome for a one-unit increase in the predictor
  - *There is a statistically significant relationship between age and library use status. For every one year increase in age, the odds of library use decrease by .01 or 1% (OR = .99; 95% CI: .98 - .996).*

Interpreting the categorical predictor odds ratios

- For significant categorical predictors, odds ratios show the increase or decrease in odds of the outcome for the category shown compared to the *reference* group (category not shown)
  - *Compared to females, males have 51% lower odds of library use (OR = .49; 95% CI: .39 -.61)*
    - 51% was computed by subtracting 1 - .49 = .51 and then multiply by 100 to get a percent
  - *Compared to Hispanic participants, Non-Hispanic Black participants have 1.55 times the odds of library use (OR = 1.55; 95% CI: 1.00 - 2.42)*
  - *Compared to those with less than a high-school education, those with a four-year degree or more have 1.90 times the odds of library use (OR = 1.90; 95% CI: 1.26 - 2.90)*

Model significance

- Is the model statistically significantly better than the baseline?
- The baseline is the percentage of library use
  - *Check the `odds.n.ends()` output to see the baseline*

```
## $`Contingency tables (model fit): percent predict
##                    Percent observed
## Percent predicted         1         0       Sum
##                 1 0.2641906 0.1737912 0.4379818
##                 0 0.2235459 0.3384723 0.5620182
##               Sum 0.4877365 0.5122635 1.0000000
```

- The probability of .488 or 48.8% library users is the baseline
- Without knowing anything else, you would be more likely to predict that each person was NOT a library user (because 51.2% are NOT library users)
- Predicting everyone is not a library user would result in the prediction being right 51.2% of the time
- Can the model do (significantly) better than that?

Null and alternate hypotheses

- H0: The model is no better than the baseline percentage at explaining library use
- HA: The model is better than the baseline at explaining library use

Get the test statistic and p-value

- The `odds.n.ends()` output showed:
  - *Chi-squared = 94.74*
  - *degrees of freedom = 12*
  - *p < .05*
- The logistic regression model is statistically significantly better than the baseline at explaining library use ($\chi^2(12)$ = 94.74; p < .05).

Model Fit (ok it's significant, but HOW MUCH better than baseline is it?)

- The last thing before putting the interpretation together is model fit
- In linear regression the model fit is R^2, which is the percent of variance in the outcome accounted for by the model
- In logistic regression the model fit is the **percent correctly predicted** which is also called **Count** R^2

Interpret the model fit

- The `odds.n.ends()` output shows:
  - *860 were correctly predicted by the model out of 1427 (60.3%)*
    - 378 of 696 library users were correctly predicted to be library users (54.3% of library users)
    - 482 of 731 non-users were correctly predicted to be non-users (65.9% of library non-users)

The model correctly predicted 60.3% of the time. It was better at predicting non-users (65.9% correct) compared to users (54.3% correct).

Altogether

- Model significance
- Model fit
- Predictor odds ratios and CI
  - *Often reported in a table*
- Assumptions and diagnostics

Altogether

A logistic regression model including age, sex, race-ethnicity, parent status, rurality status, disability status, education, and ses as predictors of library use was statistically significantly better than the baseline at explaining library use ($\chi^2(12) = 94.74$; $p < .05$). The model correctly predicted 60.3% of observations including 65.9% of the non-users and 54.3% of users.

Age, race-ethnicity, and education were statistically significantly related to library use. For every one year increase in age, the odds of library use decrease by .01 or 1% (OR = .99; 95% CI: .98 - .996). Compared to females, males have 51% lower odds of library use (OR = .49; 95% CI: .39 -.61). Compared to Hispanic participants, Non-Hispanic Black participants have 1.55 times the odds of library use (OR = 1.55; 95% CI: 1.00 - 2.42). Compared to those with less than a high-school education, those with a four-year degree or more have 1.90 times the odds of library use (OR = 1.90; 95% CI: 1.26 - 2.90). Library use was not statistically significantly associated with parent status, rurality, disability, or ses.

The assumptions of independent observations and no mulitcollinearity were met; the linearity assumption checking suggested some deviation from linearity at the youngest ages. There were two observations identified as outlier or influential values; both observations were library users who were older, lived in rural areas, were Non-Hispanic Black, and had less than a high school education.

Automatically making an odds ratio table in R

```
# load package
library(package = "sjPlot")
library(package = "sjmisc")
library(package = "sjlabelled")

# make table
tab_model(libUseModel)
```

	uses.lib.num		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.32	0.68 – 2.56	0.415
age	0.99	0.98 – 1.00	**0.002**
sex: male	0.49	0.39 – 0.61	**<0.001**
parent: parent	1.27	0.97 – 1.65	0.082
disabled: yes	0.80	0.58 – 1.09	0.165
ses: low	0.93	0.57 – 1.52	0.778
ses: medium	0.85	0.57 – 1.24	0.399
raceth: Non-Hispanic Black	1.55	1.00 – 2.42	**0.050**
raceth: Non-Hispanic White	1.32	0.93 – 1.86	0.121
educ: Four-year degree or more	1.90	1.26 – 2.90	**0.002**
educ: HS to 2-year degree	1.15	0.78 – 1.69	0.486
rurality: suburban	1.19	0.90 – 1.57	0.219
rurality: urban	1.23	0.93 – 1.63	0.150
Observations	1427		
R^2 Tjur	0.065		

Alternatives for not meeting assumptions

There is no one specific test that is the alternative to logistic regression when assumptions are not met. Some of the options for dealing with failed assumptions are:

- Include additional variables in the model or drop variables from the model and check assumptions again
- Recode or transform problematic independent variable(s) and try again
- Use an alternate model for binary outcomes like negative binomial regression or tweedie regression
- Stick to visual and descriptive statistics (always a great option!)

The End

- Exercise for this week is in GitHub
- Reminder: Your TA is the guest lecturer for next week