Logistic Regression

Estimating a simple logistic regression model

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))
```

Estimating a simple logistic regression model

- Start with a simple logistic regression with age as the only predictor of library use.
  - $p(uses.lib) = \frac{1}{1+e^{-(b_0+b_1 \cdot age)}}$
- Replace y with the name of the outcome variable, `uses.lib`, and x with the name of the predictor variable, `age`.

NHST Step 1: Write the null and alternate hypotheses

- The null and alternate hypotheses are similar to those in linear regression.
- For this logistic regression analysis, the outcome is library use and 49.5% of people in the data set use the library.
- Without any other information, it is slightly more likely that a person selected from the data frame is *not* a library user.
- This is the **baseline** value.
- By using information like age, sex, and education, the logistic model may be better able to predict the probability of library use for a given person from the data set.
- This suggests a null hypothesis to test for the logistic model:

H0: The model is no better than the baseline at predicting library use.

HA: The model is better than the baseline at predicting library use.

NHST Step 2: Compute the test statistic

- The `glm()` or **generalized linear model** function can be used to estimate a binary logistic regression model.
- The model is **generalized** because it starts with the basic linear model and generalizes it to other situations.
- The `glm()` function takes several arguments.
- Before estimating the model it is important to understand how R is interpreting the order of the categories in the outcome variable.
- The `glm()` function will treat the first category as the reference group (the group *without* the outcome) and the second category as the group *with* the outcome.
- To see the order of `uses.lib`, use `levels()` to show the two levels in order:

```
# checking the order of the outcome variable categories
levels(x = libraries.cleaned$uses.lib)
```

```
## [1] "no"  "yes"
```

Estimate the model in R

- First, enter the formula into the `glm()` function with the outcome variable `uses.lib` on the left side of the `~` and the predictor `age` on the right.
- After the formula, the data source and the **family** or model type are entered.
- Since there are different sorts of generalized linear models (glm) that are appropriate for different types of outcome variables, the `family =` argument is used to specify which type of model to estimate.

```
# estimate the library use model and print results
lib.model.small <- glm(formula = uses.lib ~ age,
                       data = libraries.cleaned,
                       family = binomial("logit"))
summary(object = lib.model.small)
```

```
##
## Call:
## glm(formula = uses.lib ~ age, family = binomial("logit"), data = libraries.c
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.291  -1.150  -1.027   1.190   1.363
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.403785   0.142194   2.840  0.00452 **
```

Interpret logistic regression R output

- The output from `glm()` contains information about the significance of the predictors, but is missing several pieces of information for reporting results including the odds ratios, model significance, and model fit.
- To get this information, use the `odds.n.ends()` function from the odds.n.ends package.

```
# open odds.n.ends
library(package = "odds.n.ends")

# get model fit, model significance, odds ratios
odds.n.ends(x = lib.model.small)
```

```
## $`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
##
```

Reading the odds.n.ends output

- The chi-squared statistic in logistic regression is computed by finding the difference between how well the model fits the data when it has no predictors in it (the null model) and how well it does with all the predictors in it.
    - A model with no predictors in it is a baseline or null model and just consists of the percentage of people with the outcome.
    - In this case, it is the percentage of people who use the library. Looking at the model fit table, the percentage of people who were observed with a value of 1 indicating they use the library was 773 out of 1571 or 49.2%.
    - Without any other information, we would predict that each person has a 49.2% probability of being a library user.
    - If a person in the data set was a library user, their probability would be 100% chance, or 1 on a scale of 0 to 1, of being a library user.
    - The predicted probability would be .492, which translates to a 49.2% chance of being a library user.
    - The difference between the actual value of library use, 1, and the observed probability of library use would be 1 - .492 = .508.
    - The predicted probability is .508 away from the correct observed value.

Calculating model fit

- Finding these differences between observed values (0s and 1s) and the predicted values (percentages), squaring each difference, and adding up all the squared differences from each person in the data set results in a value called the **deviance**.
- The deviance is a measure of how well the model fits the data.
- If the differences between the observed values and the predicted values is small, the deviance is small and the model is doing well at fitting the data.
- A smaller deviance is an indicator of a better fitting model.
- After the deviance is computed for the null model, or model without predictors, it is computed for the model with predictors in it.
- In this case the deviance is predicted for the model with age only as a predictor.
- If age was useful in making the model a better predictor of library use than the null model is, the deviance would be smaller for the model with age in it than it was for the null or baseline model.
- The difference between the deviance for the null model and the deviance for a model with predictors in it has a chi-squared distribution and is used to determine whether the full model is doing a **statistically significantly** better job at predicting the observed values than the null model.
- In the `summary()` results from the model above you will find the null deviance of 2177.5 and the "residual" or model deviance of 2166.7.

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

- The chi-squared distribution shows the probability of getting a chi-squared test statistic as large (or larger) than the one computed if the null hypothesis were true.
- In this case, the sample size is 1601 for the libraries data frame but in the `odds.n.ends()` output it shows 1571 observations in the model fit contingency table.
- A review of the data finds this is due to missing values in the outcome and predictor variables; the model only used cases with complete data for all variables in the model.
- The `odds.n.ends()` output also shows the model chi-squared of 10.815 with the corresponding degrees of freedom of 1 and a p-value of .001.

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

The chi-squared test statistic for a logistic regression model with age predicting library use had a p-value of .001. This p-value indicates there is a .1% chance of a chi-squared statistic this large or larger if the null hypothesis were true. The null hypothesis is therefore rejected in favor of the alternate hypothesis that the model is better than the baseline at predicting library use. A logistic regression model including age was statistically significantly better than a null model at predicting library use [χ^2 (1) = 10.82; p = .001].

Interpreting predictor significance

- The next thing to look at was predictor significance and interpretation.
- The `summary()` and `odds.n.ends()` output both include values and significance statistics for the age predictor.
- The `odds.n.ends()` output included odds ratios, which are easier to interpret since the outcome is transformed by the logistic function and therefore the coefficients from `summary()` are not easy to interpret directly.

Computing odds ratios

- Use the `odds.n.ends()` function to review the odds ratios:

```
# run the odds.n.ends code again
odds.n.ends(x = lib.model.small)
```

```
## $`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
##               1    338  298  636
##               0    435  500  935
##               Sum  773  798 1571
##
## $`Predictor odds ratios and 95% CI`
##                    OR     2.5 %    97.5 %
## (Intercept) 1.497482 1.1339164 1.9804811
```

Odds ratio significance

- The interpretation of an odds ratio is the increase (or decrease) in odds with a one unit increase in the predictor.
- For example, the age variable has an odds ratio of .99. This could be interpreted as, "The odds of library use decrease by 1 percent for every one year increase in age."
- The 1 percent comes from subtracting the odds ratio of .99 from 1, which is one strategy for making the odds ratio easier to interpret when it is below 1.
- It would be just as correct, although a little more difficult to understand, to conclude that "The odds of library use are .99 times as high with every one year increase in age."
- The significance of an odds ratio is determined by its confidence interval.
- Just as the other confidence intervals, the confidence interval for the odds ratio shows where the population value of the odds ratio likely lies.
- A confidence interval that includes 1 indicates that the true or population value of the relationship could have an odds ratio of 1.
- The interpretation of an odds ratio of 1 is that the odds are 1 times higher or 1 times as high for a one unit increase in x.
- This is essentially the same odds.

Interpreting significant odds ratios

- The odds ratio for the age variable is less than one and the confidence interval does not include one.
- The interpretation of this odds ratio would therefore be: *The odds of library use are .9% lower for every one year increase in age (OR = .991; 95% CI: .986 - .996)*.

Using NHST to organize the significance testing of odds ratios: NHST Step 1

H0: Library use is not associated with age.

HA: Library use is associated with age.

NHST Step 2: Compute the test statistic

- For each predictor Kiara explained that there were two possible test statistics to use to determine statistical significance.
- The `summary()` command following a `glm()` command includes a z-statistic comparing the coefficient estimate to zero.

```
##
## Call:
## glm(formula = uses.lib ~ age, family = binomial("logit"), data = libraries.c
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.291  -1.150  -1.027   1.190   1.363
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.403785   0.142194   2.840  0.00452 **
## age         -0.008838   0.002697  -3.277  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2177.5  on 1570  degrees of freedom
```

Review the distribution

- Visualize how often a z-score of -3.28 or a larger negative value would happen if there were no relationship between age and library use in a sample of 1571 observations.

NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)

- The area under the curve to the left of the test statistic is the percentage of the time you would get a z-statistic of -3.28 or more extreme under the null hypothesis of no relationship between age and library use.
- Given this area is very small, the p-value of .00105 in the output from `summary(lib.model.small)` makes sense.
- There is a .105% probability that this sample came from a population where there was no relationship between age and library use.
- There is a statistically significant relationship between age and library use (z = -3.28; p = .001).
- The other way to determine statistical significance of predictors is to examine the confidence intervals around the odds ratios.

NHST Step 4 & 5: Reject or retain the null hypothesis based on the probability

- The odds ratio for age is .991 with a 95% CI of .990 - .996.
- The confidence interval shows the range where the odds ratio likely is in the population.
- Because the confidence interval does not include 1, this indicates that the odds ratio is statistically significantly different from 1.
- The interpretation would be:

The null hypothesis of no relationship between library use and age is rejected. The odds of library use are .9% lower for every one year increase in age in the sample (OR: .991; 95% CI: .986 - .996). The 95% confidence interval indicates that the odds of library use are .4 - 1.4% lower with each one year increase in age in the population that the sample came from.