Logistic regression

The Basics

Dr. Maria Tackett

10.23.19

Click for PDF of slides

Announcements

- Reading 08 for Monday
- Project Proposal **due Wed, Oct 30 at 11:59p**

Packages

```
library(tidyverse)
library(knitr)
library(broom)
library(fivethirtyeight)
library(pROC) #ROC curves
library(questionr) #odds ratio function
```

Review

- Y: binary response
  - 1: yes
  - 0: no
- $Mean(Y) = p$
- $Var(Y) = p(1-p)$
- Odds of "yes": $\omega = \frac{p}{1-p}$

Comparing Odds

Suppose we have two independent groups with odds ω_1 and ω_2

- Odds Ratio: $\phi = \frac{\omega_1}{\omega_2}$
- Use inference to assess if groups have equal odds, i.e. $\phi = 1$
  - Hypothesis Test:

$$H_0 : \log(\phi) = 0$$

  - Confidence Interval:

$$\exp\left\{ \log(\phi) \pm z^* SE(\log(\phi)) \right\}$$

Diff. in proportions vs. odds ratio

Suppose that the probability of a disease is 0.00369 in a population of unvaccinated subjects and that the probability of the disease is 0.001 in a population of vaccinated subjects.

a. What is the difference in the proportion of subjects expected to get the disease in the unvaccinated group versus the vaccinated group?

b. What are the odds of disease without vaccine relative to the odds of disease with vaccine?

Now suppose the probability of the disease is 0.48052 in the population of unvaccinated subjects and 0.2 in the population of vaccinated subjects.

a. What is the difference in the proportion of subjects expected to get the disease in the unvaccinated group versus the vaccinated group?

b. What are the odds of disease without vaccine relative to the odds of disease with vaccine?

Compare your responses from the two scenarios:

1. How do the difference in proportions compare?
2. How do the odds ratios compare?

Is it rude to recline your seat on a plane?

```
flying <- fivethirtyeight::flying %>%
  drop_na(recline_rude, height, age) %>%
  mutate(
    rude = if_else(recline_rude %in%
                     c("Somewhat", "Very"), 1, 0),
         rude = factor(rude),
         age = factor(age, order = FALSE)) # to display in model
```

- **height**: self-reported height in feet and inches
- **age**: 18-29, 30-44, 45-60, > 60
- **rude**: 1: yes, 0: no (Is it rude to recline your seat on a plane?)

Source: *41 Percent of Fliers Think You're Rude If You Recline Your Seat*

Opinions about flying

age	0	1
18-29	78	94
30-44	143	79

Is there a significant difference in the proportion of 18-29 year olds versus 30-44 year olds who think reclining a seat on a plane is rude?

`odds.ratio` function

- We will use the **`odds.ratio`** function in the **questionr** package to compute odds ratios and the corresponding confidence interval

```
#calculate odds ratio and 95% confidence interval
flying %>%
  filter(age %in% c("18-29", "30-44")) %>%
  glm(rude ~ age, data = ., family = binomial) %>%
  odds.ratio(level=0.95) %>%
  kable(format="markdown", digits = 3)
```

	OR	2.5 %	97.5 %	p
(Intercept)	1.205	0.893	1.630	0.223
age30-44	0.458	0.304	0.687	0.000

We are 95% confident that the interval 0.304 to 0.687 contains the true odds ratio of 30-44 year olds versus 18-29 year olds who think reclining a seat.

Hypothesis Test for Odds Ratio

- We want to test whether two groups have equal odds, i.e. $\phi = \frac{\omega_1}{\omega_2} = 1$
- Null Hypothesis: $H_0 : \log(\phi) = \log\left(\frac{\omega_1}{\omega_2}\right) = 0$
- Test Statistic:

$$z = \frac{\log(\hat{\phi}) - 0}{SE_0[\log(\hat{\phi})]} = \frac{\log(\hat{\phi}) - 0}{\sqrt{\frac{1}{n_1\hat{p}_c(1-\hat{p}_c)} + \frac{1}{n_2\hat{p}_c(1-\hat{p}_c)}}}$$

- p-value: proportion of $N(0, 1)$ distribution as extreme or more extreme than the test statistic

Standard error $SE_0[\log(\hat{\phi})]$

- The null hypothesis is that odds ratio is 1, i.e. the proportions are equal
- To calculate standard error, we estimate $\hat{p}_c$, the sample proportion from the combined data

$$SE_0[\log(\hat{\phi})] = SE_0\left[\log\left(\frac{\hat{\omega}_1}{\hat{\omega}_2}\right)\right] = \sqrt{\frac{1}{n_1\hat{p}_c(1-\hat{p}_c)} + \frac{1}{n_2\hat{p}_c(1-\hat{p}_c)}}$$

Opinions about reclining seat

Do the odds of thinking it's rude to recline a seat on a plane differ between 18-29 and 30-44 year olds?

$$H_0 : \log(\phi) = 0$$
$$H_a : \log(\phi) \neq 0$$

- $\hat{p}_c$ = 0.439
- **18 - 29**: $n = 172$, $\hat{\omega} = 1.208$
- **30 - 44**: $n = 222$, $\hat{\omega} = 0.553$

1. Calculate the test statistic.
2. Calculate p-value and make a conclusion.

Looking at the odds ratio is useful...

...but we want to build a model to incorporate more variables that could potentially explain the odds of a flier having the opinion that reclining a seat is rude.

Linear model?

- We want to use a model to predict a binary response y
- Suppose we use a linear regression model to predict y using some explanatory variable x

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

- This model assumes that y could be any continuous value; however, it can only be 0 or 1
- So linear regression is **not** appropriate

Other model choices

Let $P(y_i = 1|x_i) = p_i$ and $P(y_i = 0|x_i) = 1 - p_i$

Potential models for p_i:

- Linear: $p_i = \beta_0 + \beta_1 x_i$
  - could predict that p_i is outside of $(0, 1)$
- Log-linear: $\log(p_i) = \beta_0 + \beta_1 x_i$
  - could predict that p_i is greater than 1

Logistic Regression Model

- Suppose $P(y_i = 1|x_i) = p_i$ and $P(y_i = 0|x_i) = 1 - p_i$
- The **logistic regression model** is

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

- $\log\left(\frac{p_i}{1-p_i}\right)$ is called the **logit** function

Logistic Regression Model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

- We can calculate p_i by solving the logit equation:

$$p_i = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}$$

Solving Logit Equation

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

$$\Rightarrow \quad \exp\left\{\log\left(\frac{p_i}{1-p_i}\right)\right\} = \exp\{\beta_0 + \beta_1 x_i\}$$

$$\Rightarrow \quad \frac{p_i}{1-p_i} = \exp\{\beta_0 + \beta_1 x_i\}$$

$$\Rightarrow \quad p_i = \frac{\exp\{\beta_0 + \beta_1 X_i\}}{1 + \exp\{\beta_0 + \beta_1 x_i\}}$$

Interpreting the intercept: β_0

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

- When $x = 0$, log-odds of y are β_0
  - Won't use this interpretation in practice
- **When $x = 0$, odds of y are $\exp\{\beta_0\}$**

Interpreting slope coefficient β_1

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

If x is a quantitative predictor

- As x_i increases by 1 unit, we expect the log-odds of y to increase by β_1
- **As x_i increases by 1 unit, we expect the odds of y to multiply by a factor of $\exp\{\beta_1\}$**

If x is a categorical predictor

- The difference in the log-odds between group x and the baseline is β_1
- **The odds of y for group k are expected to be $\exp\{\beta_1\}$ times the odds of y for the baseline group.**

Inference for coefficients

- The standard error is the estimated standard deviation of the sampling distribution of $\hat{\beta}_1$
- We can calculate the C confidence interval based on the large-sample Normal approximations
- CI for β_1:

$$\hat{\beta}_1 \pm z^* SE(\hat{\beta}_1)$$

CI for $\exp\{\beta_1\}$:

$$\exp\{\hat{\beta}_1 \pm z^* SE(\hat{\beta}_1)\}$$

Estimating the coefficients

- Estimate coefficients using **maximum likelihood estimation**
  - covered in STA 250 and STA 360
- **Basic Idea:**
  - Find values of β_0 and β_1 that make observed values of y the most likely to have occurred
  - Use multivariable calculus and numerical methods to estimate coefficients
- In this class, we will use R to estimate the coefficients

Logistic regression in R

- Fit a logistic model using the Use the **`glm()`** function
  - Set **`family=binomial`** for a binary response variable

```
my.model <- glm(y ~ x1 + ... + xp, data = my.data,
                family = binomial)
```

- Display model with log-odds as the response

```
tidy(my.model, exponentiate = FALSE)
```

- Display model with odds as response

```
tidy(my.model, exponentiate = TRUE)
```

Recoding height

We want to use height to predict whether a flier will think reclining a seat on an airplane is rude. To do so, we will recode height so it's quantitative.

```
flying <- flying %>%
  separate(height, c("feet", "inches"), remove = FALSE) %>%
  mutate(height_in = case_when(
    height == "Under 5 ft." ~ 60,
    TRUE ~ as.numeric(feet)*12 + as.numeric(inches)))
```

Recoding height

```
flying %>%
  select(height, height_in) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   height   height_in
##   <ord>        <dbl>
## 1 "6'3\""         75
## 2 "5'8\""         68
## 3 "5'11\""        71
## 4 "5'7\""         67
## 5 "5'9\""         69
```

Reclining vs. height

Use the mean-centered `height` in the model, so the intercept will have a meaningful interpretation

```
flying <- flying %>%
  mutate(heightCent = height_in - mean(height_in))
```

```
ht_model <- glm(rude ~ heightCent, data = flying, family = binomia
kable(tidy(ht_model, exponentiate = FALSE, conf.int = TRUE),
      format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.348	0.070	-4.970	0.000	-0.485	-0.211
heightCent	0.012	0.018	0.703	0.482	-0.022	0.047

Reclining vs. height

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.348	0.070	-4.970	0.000	-0.485	-0.211
heightCent	0.012	0.018	0.703	0.482	-0.022	0.047

- For each additional inch taller a flier is, the odds they think reclining the seat on a plane is rude are expected to multiply by a factor of 1.013), with 95% confidence interval 0.978 to 1.049.
- The odds a flier of average height thinks reclining the seat on a plane is rude are 0.706 to 1, with 95% confidence interval 0.615 to 0.81.

Is `height` a significant predictor of whether a flier thinks reclining the seat is rude?

Reclining vs. height & age

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.188	0.153	1.226	0.220	-0.112	0.490
heightCent	0.013	0.018	0.741	0.459	-0.022	0.049
age30-44	-0.782	0.208	-3.766	0.000	-1.192	-0.377
age45-60	-0.590	0.203	-2.901	0.004	-0.990	-0.193
age> 60	-0.669	0.208	-3.216	0.001	-1.078	-0.263

1. Interpret the coefficient of `age30-44` in the context of the data.
2. Describe the relationship between a flier's age and the odds they think reclining the seat on a plane is rude.

Predictions & Model Fit

Predictions

- We are often interested in predicting if a given observation will have a "yes" response
- To do so, we will use the logistic regression model to predict the probability of a "yes" response for the given observation. If we have one predictor variable, then...

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

- We will use the predicted probabilities to classify the observation as having a "yes" or "no" response

Will the passenger think I'm rude?

- Suppose you want to recline your seat on an airplane, but you first want to determine if the passenger behind you will think you're rude. The passenger is about 6ft tall and around 35-40 years old.
- Predicted log-odds that this passenger thinks reclining the seat is rude:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 0.188 + 0.013 \times (72 - 67.44) - 0.782 = -0.534$$

- The probability this passenger thinks reclining the seat is rude:

$$\hat{p}_i = \frac{\exp\{-0.534\}}{1+\exp\{-0.534\}} = 0.3696$$

Predictions in R

```
x0 <- data_frame(heightCent = (72 - 67.44), age = "30-44")
```

- Predicted log-odds

```
predict(ht_age_model, x0)
```

```
##          1
## -0.5337904
```

- Predicted probabilities

```
predict(ht_age_model, x0, type = "response")
```

```
##         1
## 0.3696333
```

Will the passenger think I'm rude?

```
predict(ht_age_model, x0, type = "response")
```

```
##         1
## 0.3696333
```

The probability the passenger will think you're rude is 0.3696.

Based on this probability, do you expect the passenger to think you're rude? Why or not why not?

Confusion Matrix

- We can use the estimated probabilities to predict outcomes
- *Ex.*: Establish a threshold such that $y = 1$ if predicted probability is greater than the threshold ($y=0$ otherwise)
- Determine how many observations were classified correctly and incorrectly and put the results in a 2×2 table
  - This table is the **confusion matrix**
- If the proportion of misclassifications is high, then we might conclude the model doesn't fit the data well

Confusion Matrix

Suppose we use 0.5 as the threshold to classify responses

```
threshold <- 0.5
ht_age_aug <- augment(ht_age_model, type.predict = "response")
```

```
ht_age_aug %>%
  mutate(rude_predict = if_else(.fitted > threshold, "Yes", "No")) %>%
  group_by(rude, rude_predict) %>%
  summarise(n = n()) %>%
  spread(rude, n) %>%
  kable(format="markdown")
```

rude_predict	0	1
No	416	255
Yes	78	94

Confusion matrix

rude_predict	0	1
No	416	255
Yes	78	94

What proportion of observations were misclassified?

Sensitivity & Specificity

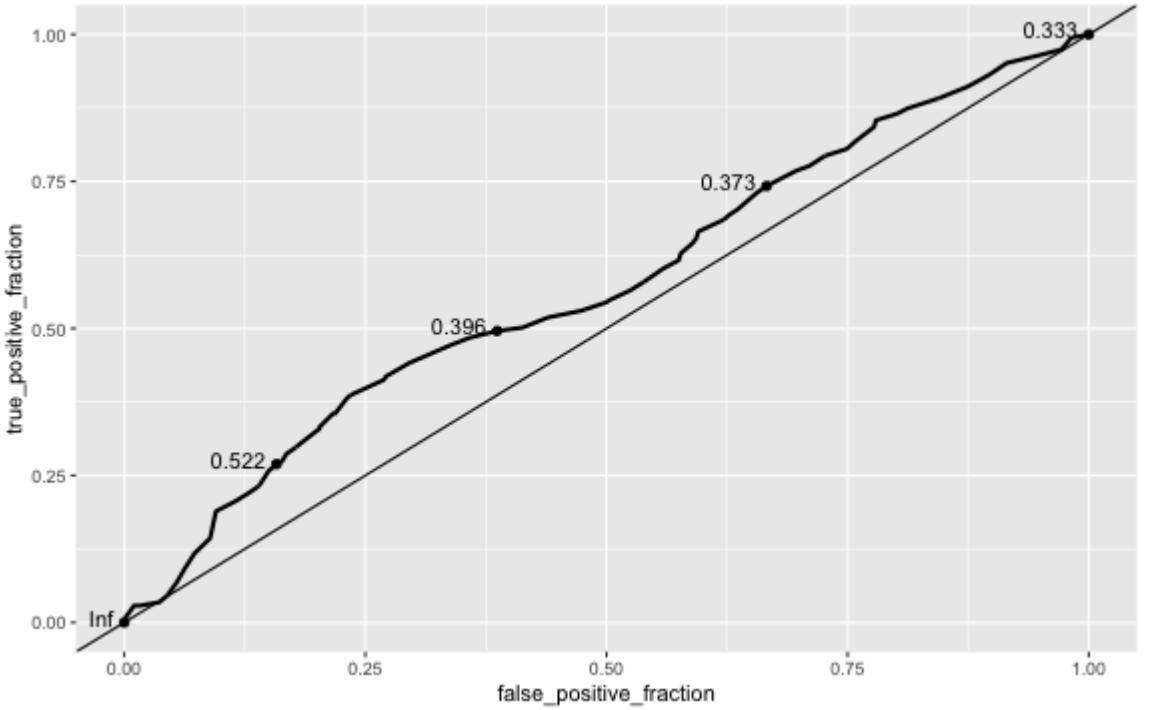
- **Sensitivity:** Proportion of observations with $y = 1$ that have predicted probability above a specified threshold
  - Called true positive rate
- **Specificity:** Proportion of observations with $y = 0$ that have predicted probability below a specified threshold
  - (1 - specificity) called false positive rate
- What we want:
  - High sensitivity
  - Low values of 1-specificity

ROC Curve

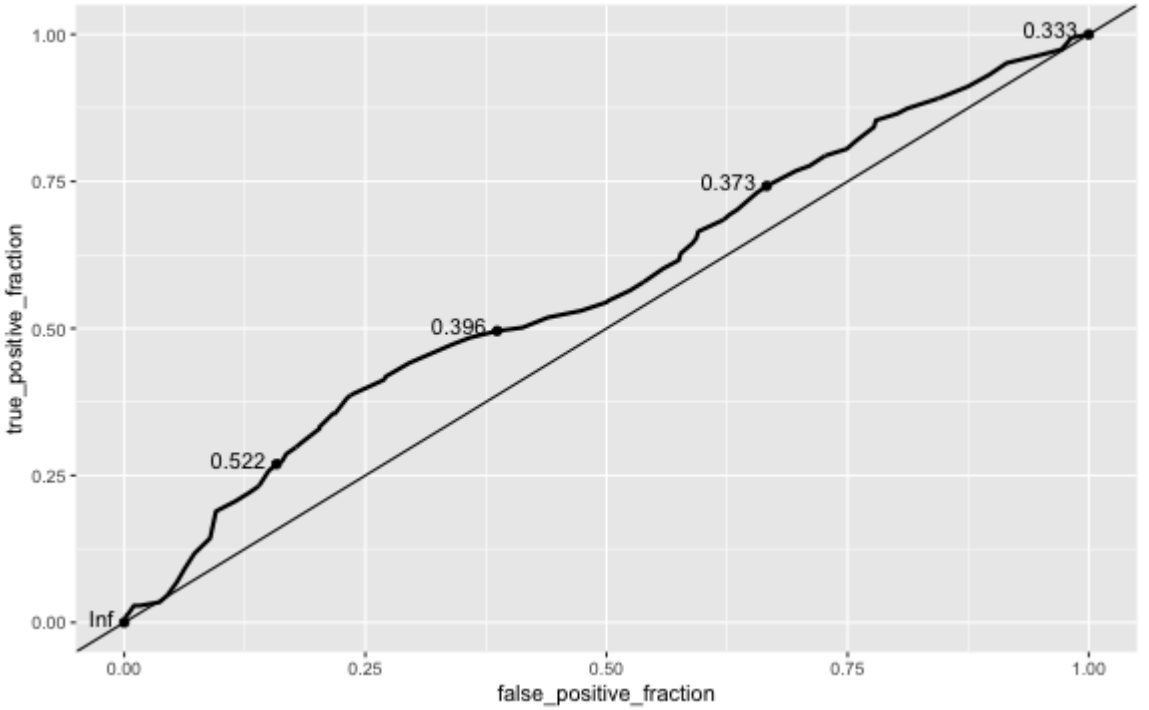
- Receive Operating Characteristic (ROC) curve :
  - *x-axis*: $1 -$ specificity
  - *y-axis*: Sensitivity
- Evaluated with a lot of different values for the threshold
- Logistic model fits well if the area under the curve (AUC) is close to 1
- ROC in R
  - Use the `roc` function in the pROC to calculate AUC
  - Use `geom_roc` layer in ggplot to plot the ROC curve

Visualize ROC curve

```
library(plotROC) #extension of ggplot2
ggplot(ht_age_aug, aes(d = as.numeric(rude), m = .fitted)) +
  geom_roc(n.cuts = 5, labelround = 3) +
  geom_abline(intercept = 0)
```



Area under curve




```
## [1] 0.5719233
```