Logistic regression, 1

Frank Edwards

2/19/2023

Read in the data for today

```
admissions <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
head(admissions)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

```
nrow(admissions)
```

```
## [1] 400
```

Evaluate distribution of binary admission variable

```
table(admissions$admit)
```

```
##
##   0   1
## 273 127
```

```
mean(admissions$admit)
```

```
## [1] 0.3175
```

Evaluate bivariate relationships between admission and focal predictors

```
admissions %>%
    group_by(admit) %>%
    summarize(gpa_mean = mean(gpa), gre_mean = mean(gre))
```

```
## # A tibble: 2 x 3
##   admit gpa_mean gre_mean
##   <int>    <dbl>    <dbl>
## 1     0     3.34     573.
## 2     1     3.49     619.
```

Lab: Let's fit a model

What else might predict admission?

```
head(admissions)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

1. Theorize

Think carefully about what we want to accomplish. How do admissions work?

2. Write a model

$$\text{admit} \sim \text{Bernoulli}(p)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{GPA}_i + \beta_2 \text{GRE}_i$$

2. Fit the model

```
m1 <- glm(admit ~ gpa + gre, data = admissions, family = "binomial"
```

3. Compare model fits

```r
## baseline gpa model
m0 <- glm(admit ~ gpa, data = admissions, family = "binomial")

broom::tidy(m0)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    -4.36     1.04      -4.21 0.0000257
## 2 gpa             1.05     0.299      3.52 0.000437
```

```r
broom::tidy(m1)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic    p.value
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept) -4.95      1.08        -4.60 0.00000415
## 2 gpa          0.755     0.320        2.36 0.0182
```

3. Comparing model fits

```
AIC(m1, m0)
```

```
##    df      AIC
## m1  3 486.3440
## m0  2 490.9676
```

Iterate

Repeat steps 1-3 until satisfied that we've built a theoretically justified well-fitting model.

4. Interpet

1. Identify what we'd like to make inferences about. Which groups or cases would be theoretically interesting for our topic.
2. Create these cases in a new data frame.
3. Predict
4. Visualize

Homework 5

Who was most (and least) likely to die on the Titanic? Use `~/hw/data/titanic.csv` for this one.

1. Develop a theory for survival
2. Write a model
3. Fit the model
4. Think about revising the model
5. Compare model fits
6. Interpret the model