6:14 - Class Starts

Logistic Regression

Prof. Maria Tackett

03.25.20

STA 210

[Click for PDF of slides]

Part I: Categorical Response Variables

12:30

Quantitative vs. Categorical Response Variables

Quantitative response variable:

- Sales price of a house in Levittown, NY
- **Model**: variation in the mean sales price given values of the predictor variables (`bedrooms`, `lot_size`, `year_built`, etc.)

Categorical response variable:

- Patient at risk of coronary heart disease (Yes/No)
- **Model**: variation in the probability a patient is at risk of coronary heart disease given values of the predictor variables (`age`, `currentSmoker`, `totChol`, etc.)

STA 210

13:38

Models for categorical response variables

Logistic Regression

2 Outcomes

Agree/Disagree

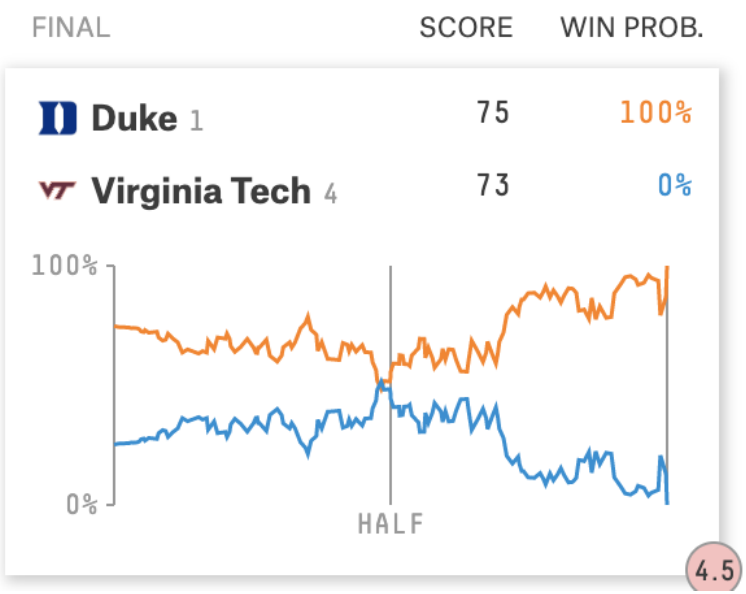
Multinomial Logistic Regression

3+ Outcomes

Strongly Agree, Agree, Disagree, Strongly Disagree

Let's focus on logistic regression models for now.

FiveThirtyEight Live Win Probabilities



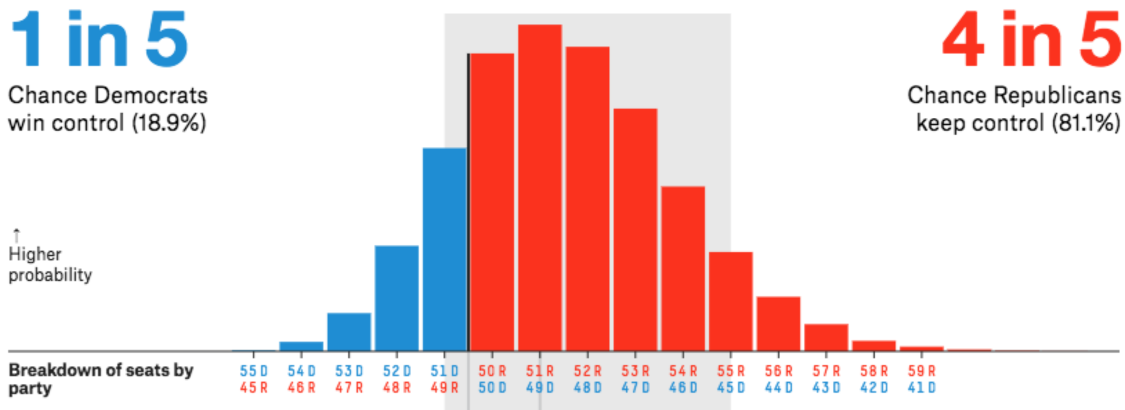
[FiveThirtyEight: 2019 March MadnessLive Win Probabilities]

"These probabilities are derived using ***logistic regression analysis****, which lets us plug the current state of a game into a model to produce the probability that either team will win the game.*

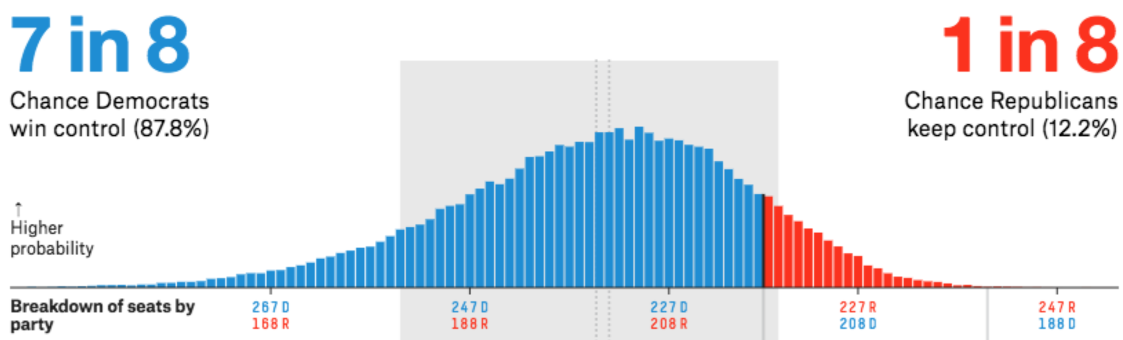
- ["How Our March Madness Predictions Work"]

STA 210

2018 Election Forecasts



FiveThirtyEight.com Senate forecast



FiveThirtyEight.com House forecast

*Our models are probabilistic in nature; we do a lot of thinking about these probabilities, and the goal is to develop **probabilistic estimates** that hold up well under real-world conditions.*

-"How FiveThirtyEight's House, Senate, and Governor Models Work"

STA 210

14:20

Response Variable, Y

- Y is a binary response variable
  - 1: yes (success)
  - 0: no (failure)
- $\text{Mean}(Y) = \pi$
  - π is the proportion of "yes" responses in the population
  - $\hat{\pi}$ is the proportion of "yes" responses in the sample
- $\text{Variance}(Y) = \pi(1 - \pi)$
  - Sample variance: $\hat{\pi}(1 - \hat{\pi})$
- $\text{Odds(Y=1)} = \frac{\pi}{1-\pi}$
  - Sample odds: $\frac{\hat{\pi}}{1-\hat{\pi}}$

STA 210

Odds

- Given π, the population proportion of "yes" responses (i.e. "success"), the corresponding **odds** of a "yes" response is

$$\omega = \frac{\pi}{1-\pi}$$

- The *sample odds* are $\hat{\omega} = \frac{\hat{\pi}}{1-\hat{\pi}}$

- Ex: Suppose the sample proportion $\hat{\pi} = 0.3$. Then, the sample odds are

$$\hat{\omega} = \frac{0.3}{1-0.03} = 0.4286 \approx \text{ 2 in 5}$$

15:34

Properties of the odds

- $\text{odds} \geq 0$
- If $\pi = 0.5$, then odds $= 1$
- If odds of "yes" $= \omega$, then the odds of "no" $= \frac{1}{\omega}$
- If odds of "yes" $= \omega$, then $\pi = \frac{\omega}{(1+\omega)}$

STA 210

16:20

Risk of coronary heart disease

This dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. We want to predict if a patient has a high risk of getting coronary heart disease in the next 10 years.

Response:

`TenYearCHD`:

- 0 = Patient is not high risk of having coronary heart disease in the next 10 years
- 1 = Patient is high risk of having coronary heart disease in the next 10 years

Predictors:

- **`age`**: Age at exam time.
- **`currentSmoker`**: 0 = nonsmoker; 1 = smoker
- **`totChol`**: total cholesterol (mg/dL)

STA 210

P(democrats win) = 0.189

$$\text{odds} = \frac{0.189}{1-0.189} = 0.233$$

≈ 1 in 5

referring to slide 7.

24:39

Response Variable, `TenYearCHD`

```
## # A tibble: 2 x 3
##   TenYearCHD      n proportion
##   <fct>       <int>      <dbl>
## 1 0            3101      0.848
## 2 1             557      0.152
```

- $\hat{\pi}$ = 0.152
- Sample variance = 0.152 * (1- 0.152) = 0.128896
- Odds(Y = 1) = 0.152/(1 - 0.152) = 0.1792453
- Odds(Y = 0) = 1 / 0.1792453 = 5.5789474

STA 210

Let's incorporate more variables

- We want to use information about a patient's age, cholesterol, and whether or they are a smoker to understand the probability they're high risk of having coronary heart disease.

- To do this, we need to fit a model!

STA 210

25:50

Consider possible models

- y: Whether a patient in the sample is high risk of having coronary heart disease.
- $\pi_i = P(y_i = 1|\text{age}_i, \text{currentSmoker}_i, \text{totChol}_i)$: probability a patient i is high risk for coronary heart disease given their age, smoking status, and total cholesterol

Let's consider fitting a multiple linear regression model. Below are 3 possible response variables. For each response variable, briefly explain why a multiple linear regression model is **not** appropriate.

Model 1: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\text{age} + \hat{\beta}_2\text{currentSmoker} + \hat{\beta}_3\text{totChol}$

Model 2: $\hat{\pi}_i = \hat{\beta}_0 + \hat{\beta}_1\text{age} + \hat{\beta}_2\text{currentSmoker} + \hat{\beta}_3\text{totChol}$

Model 3: $\widehat{\log(\pi)}_i = \hat{\beta}_0 + \hat{\beta}_1\text{age} + \hat{\beta}_2\text{currentSmoker} + \hat{\beta}_3\text{totChol}$

STA 210

Model 1 : y is a categorical variable that only takes values 0 or 1.

in multiple linear regression, we assume y is Normally distributed

Model 2 : $0 \leq \pi \leq 1$ but a linear model could produce predictions outside of this range

Model 3 : $-\infty < \log(\pi) \leq 0$ but a linear model could produce predictions outside this range.

Part 2: Basics of logistic regression

43:35

Logistic Regression Model

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

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

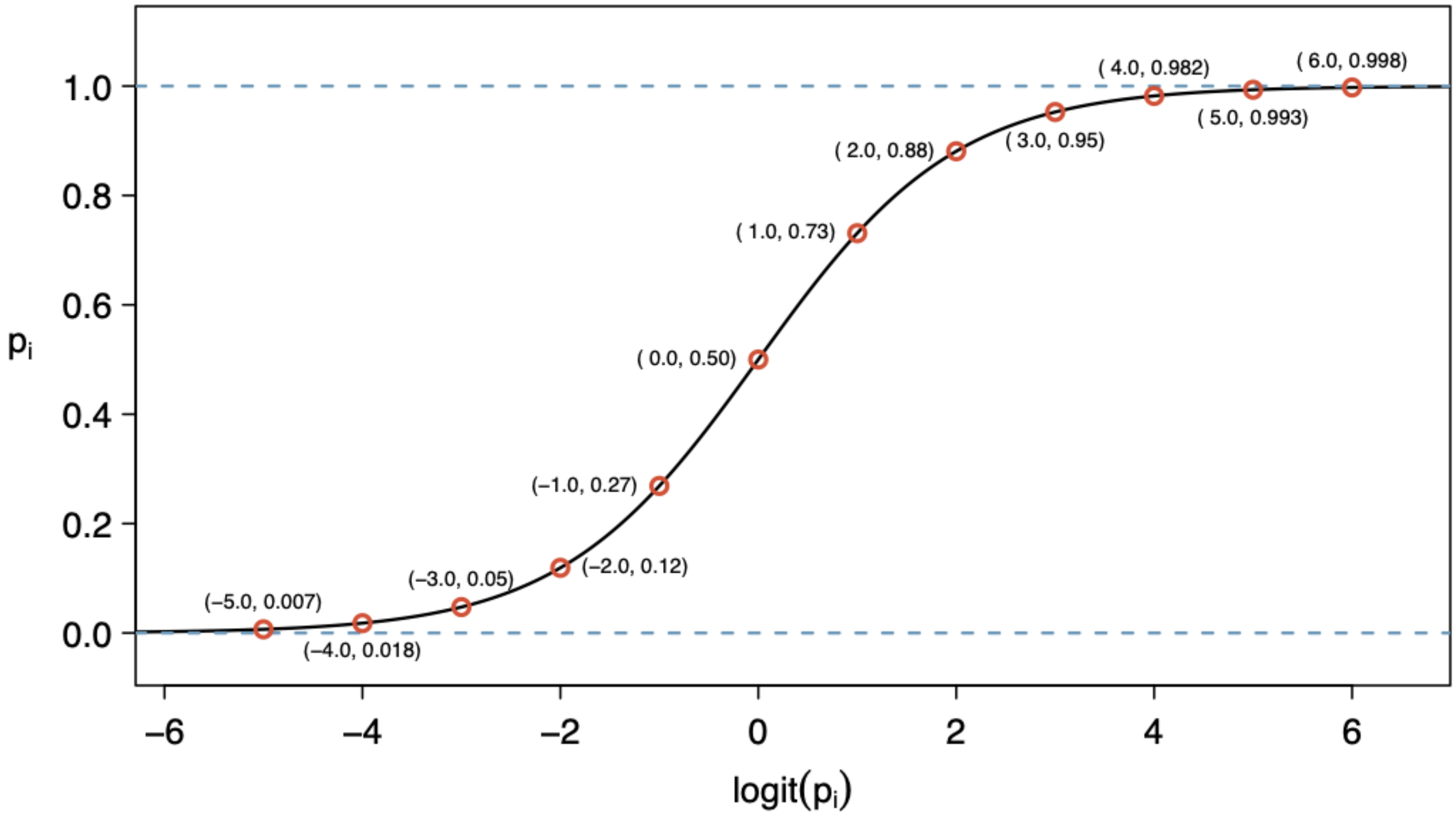
- $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ is called the **logit** function

STA 210

45:10

Logit function

$$0 \le \pi \le 1 \quad \Rightarrow \quad -\infty < \log\left(\frac{\pi}{1-\pi}\right) < \infty$$



OpenIntro Statistics, 4th ed (pg. 373)

50:25

Estimating the coefficients

- Estimate coefficients using **maximum likelihood estimation**
- Basic Idea:
  - Find values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that give observed data the maximum probability of occuring
  - More details pg. 156 - 157 of the textbook
- We will fit logistic regression models using R

STA 210

Y = Bernoulli $\quad P(Y=y) = \pi^{y}(1-\pi)^{1-y}$

$$P(Y=1) = \pi^{1}(1-\pi)^{1-1}$$

$$= \pi$$

$$P(Y_1=y_1, Y_2=y_2, \ldots, Y_n=y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \left[\frac{\exp\{\beta_0+\beta_1 X_i\}}{1+\exp\{\beta_0+\beta_1 X_i\}}\right]^{y_i}\left[1-\frac{\exp\{\beta_0+\beta_1 X_i\}}{1+\exp\{\beta_0+\beta_1 X_i\}}\right]^{1-y_i}$$

Interpreting the intercept: β_0

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

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

58:00

Interpreting slope coefficient β_1

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

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, the odds of y multiply by a factor of $\exp\{\beta_1\}$**

If x is a categorical predictor. Suppose $x_i = k$

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

STA 210

58:50

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 $\exp\{\beta_1\}$:

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

STA 210

59:20

Modeling risk of coronary heart disease

Let's use the mean-centered variables for age and `totChol`.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.111	0.077	-27.519	0.000	-2.264	-1.963
ageCent	0.081	0.006	13.477	0.000	0.070	0.093
currentSmoker1	0.447	0.099	4.537	0.000	0.255	0.641
totCholCent	0.003	0.001	2.339	0.019	0.000	0.005

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.111 + 0.081\text{ageCent} + 0.447\text{currentSmoker} + 0.003\text{totChol}$$

59:50

Modeling risk of coronary heart disease

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.111 + 0.081\text{ageCent} + 0.447\text{currentSmoker} + 0.003\text{totChol}$$

Use the model to interpret the following. Write all interpretations in terms of the odds of a patient being high risk for coronary heart disease.

1. Interpret the intercept.
2. Interpret `ageCent` and its 95% confidence interval.
3. Interpret `currentSmoker1` and its 95% confidence interval.

STA 210

1:05:00

Intercept: The odds that a patient who is average age, does not currently smoke, and who has average cholesterol is high risk of coronary heart disease are exp(-2.111).

ageCent: Holding all else constant, the odds that a patient is high risk for coronary heart disease multiply by a factor of exp(0.081) for each year increase in age.

totCholCent: Holding all else constant, the odds that a patient is high risk for coronary heart disease multiply by a factor of exp(0.003) for each additional mg/dL in total cholesterol.

currentSmoker: Holding all else constant, the odds a patient who currently smokes is high risk for coronary heart disease are exp(0.447) times the odds a patient who doesn’t smoke is high risk of coronary heart disease.