Logistic Regression

READING:	9.1-9.4 7.3-7.6
EXERCISES:	CH 9. 21,33,35,37
ASSIGNED:	NONE
PRODUCER:	DR. MARIO

IMG CREDIT: ALEX RIEGERT-WATERS

Motivation

Binary Logistic Regression Model

- Response Variable is Binary (Coded as Indicator)

$$Y = \begin{cases} 1 & if\ Yes\ (Success) \\ 0 & if\ No\ (Failure) \end{cases}$$

- Predictor Variable Could Be **Numeric** or **Categorical**

- Bad Idea -> **Linear Regression** or **ANOVA**

$$Y \neq \beta_0 + \beta_1 X + \epsilon \qquad Y \neq \mu + \alpha_i + \epsilon$$

Binary Logistic Regression Model

- Requirements

$Y = Binary\ Response \qquad X = Predictor\ Variable$

$\pi = P(Y = 1|X = x) = Proportion\ of\ 1's\ \ if\ X = x$

- Logistic Regression Model

$$\log\left(\frac{\pi}{1-\pi}\right) = \underbrace{\boxed{\beta_0 + \beta_1 X}}_{\text{Linear Model}} \quad \text{or} \quad \pi = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Binary Logistic Regression Model

- Parameter π Versus Statistic p

- Example
  - π = Probability of Having Green Eyes
  - Sample of 10 Random People

    $0,1,1,0,1,0,0,0,0,0$
  - p = Sample Average or Proportion of Blue-Eyed People in Sample

$$p = \frac{0+1+1+0+1+0+0+0+0+0}{10} = \frac{3}{10} = 0.3 = 30\%$$

- Another Reason Why Linear Regression or ANOVA Wouldn't Work

Binary Logistic Regression Model

- Odds that $Y = 1$

  Parameter: $\frac{\pi}{1-\pi}$ Estimate: $\frac{p}{1-p}$

- Odds are a Ratio of $\mathrm{P}(Y = 1)$ to $\mathrm{P}(Y \neq 1) = \mathrm{P}(\mathrm{Y} = 0)$ (Binary Case)

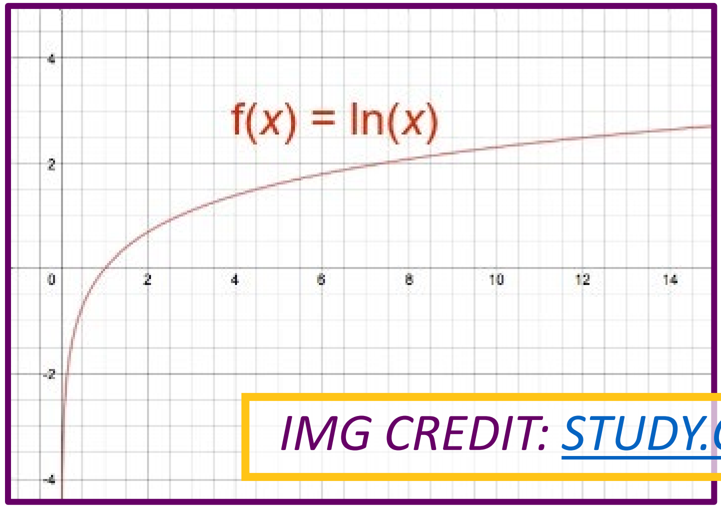
- Example: The Odds of a Horse Winning a Race is 4 to 1 or 4:1 or 4
  - Interpretation: "4 Wins for Every 1 Loss"
  - Probability: $P(Win) = 4/5$ and $P(Loss) = 1/5$
  - Calculation of Odds: $P(Win)/\, P(Loss) = \frac{4}{5} * \frac{5}{1} = 4$

Binary Logistic Regression Model

- Reason for Logistic Regression Model

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

- Notice
  - Probability: $0 < \pi < 1$
  - Odds: $0 < \frac{\pi}{1-\pi} < \infty$
  - Log Odds: $-\infty < \log\left(\frac{\pi}{1-\pi}\right) < \infty$



IMG CREDIT: STUDY.COM

Binary Logistic Regression Model

- Interpretation of Logistic Regression Model
  - Model:

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \beta_0 + \beta_1 X$$

  - Default in Statistics is the Natural Logarithm
  - Intercept Represents the Log Odds When *X=0*
  - Slope Represents the Change in Log Odds When *X* Increases by 1

Binary Logistic Regression Model

- Suppose We Have Model

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 3 - 2X$$

Log Odds Decreasing by 2 for Every 1 Unit Increase in X is not Equivalent to Saying that Odds Decreases by e^2 for Every 1 Unit Increase in X

- Then for Odds

$$\frac{\hat{\pi}}{1-\hat{\pi}} = e^{3-2X} \neq e^3 - e^2 X$$

- Slope $\beta_1 < 0$ Indicates Odds Decreases as X Increases

Binary Logistic Regression Model

- Suppose We Have Model

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 3 - 2X$$

- Then for Probability

$$\hat{\pi} = \frac{e^{3-2X}}{1+e^{3-2X}} = \frac{Odds}{1+Odds}$$

- Notice What Happens When $X = \frac{-\beta_0}{\beta_1} = \frac{3}{2}$

$$\hat{\pi} = \frac{e^{3-2X}}{1+e^{3-2X}} = \frac{e^{3-2\left(\frac{3}{2}\right)}}{1+e^{3-2\left(\frac{3}{2}\right)}} = \frac{e^{0}}{1+e^{0}} = \frac{1}{1+1} = 1/2$$

Binary Logistic Regression Model

- Estimating Parameters β_0 and β_1

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

- Recall in Linear Regression We Chose Estimates Based Off Minimization of a Bad Thing (SSE)

- In Logistic Regression We Choose Estimates that Maximize the Likelihood (Good Thing)

- The Likelihood the Probability of Our Data

Binary Logistic Regression Model

- Function in R that Estimates Logistic Regression Models

  glm(*formula*, family=binomial, *data*)

- GLM Stands for Generalized Linear Model

- The "family=binomial" Argument Uses a Logit Link Function to Connect the Mean π to a Linear Predictor

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Example: Putts

- Data from 587 Different Putts

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
Total	101	119	108	125	134

- Question: *What is the relationship between the length of a putt and the probability of making that putt?*

Supplement for Lecture 27

- Plot Raw Data and Fit Linear Regression -> Problematic
- Plot Summarized Data and Fit Linear Regression -> Problematic
- Fit Logistic Regression and Redo Plots -> Reasonable

Supplement for Lecture 27

- Comparing Sample Proportions to Estimates From Model

$$\frac{\#\ Made}{\#\ Attempts}$$

Length	3	4	5	6	7
$\hat{p}$	0.832	0.739	0.565	0.488	0.328
$\hat{\pi}$	0.826	0.730	0.605	0.465	0.330

Why the Difference?

$$\frac{e^{3.26-0.57X}}{1+e^{3.26-0.57X}}$$

- Estimate Odds
  - When are the Odds Greater Than 1?
  - When are the Odds Less Than 1?

Odds Ratios

- Way to Compare Two Groups
  - Example: Compare a Putt at 3ft Versus a Putt at 4ft

- Formula

$$OR = \frac{Odds_1}{Odds_2} = e^{\beta_0+\beta_1 X_1} / e^{\beta_0+\beta_1 X_2}$$

- Interpretation: *Odds of success in Group 1 is _____ times the odds of success in Group 2*

Supplement for Lecture 27

- Calculate Odds Ratios
- Compare Odds Ratios to Slope of Logistic Regression Model

$$Slope = \frac{Rise}{Run} = \frac{Change\ in\ Log\ Odds}{1-0} = \frac{\log(Odds_{a+1}) - \log(Odds_a)}{1-0} = \log(OR)$$

Rule of Logarithms:
$\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

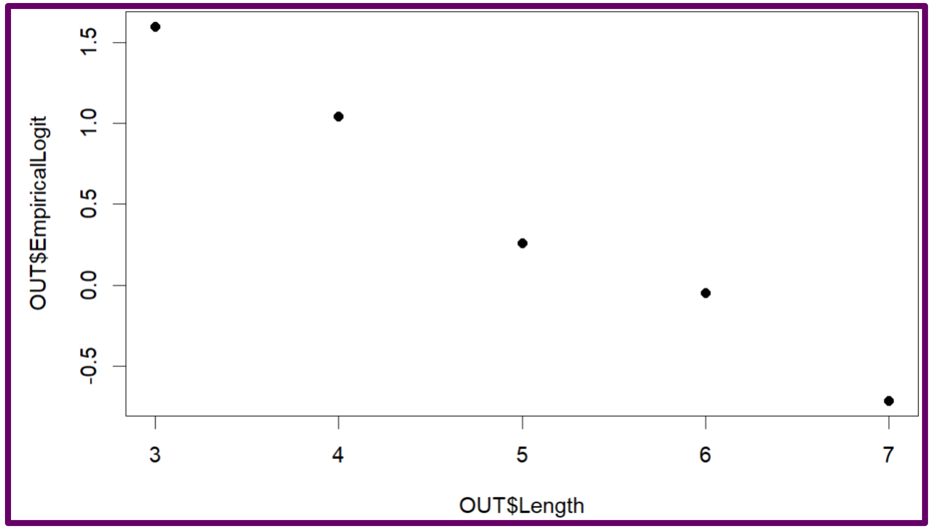
- What Happens When We Increase X by 1?

$$\underbrace{e^{\beta_0+\beta_1(X+1)}}_{\text{Odds}} = e^{\beta_0+\beta_1 X+\beta_1} = (e^{\beta_1})\underbrace{e^{\beta_0+\beta_1 X}}_{\text{Odds}}$$

Rule of Exponents:
$x^{a+b} = x^a x^b$

Assumptions for Logistic Regression

- Linearity
  - Assume the Linear Model for the Log Odds is Reasonable
  - Assess by Plotting the Log Odds from Proportions in Sample Against Your X Variable
  - Check for Linearity



- Large Sample Size
  - Recall: $np>10$ and $n(1-p)>10$
  - Require 10 Successes and Failures Per Predictor Variable

Assumptions for Logistic Regression

- Randomness
  - Is Flipping a Weighted Coin Reasonable for Deciding Whether or Not the Outcome is 0 or 1?
  - Tied to the Bernouilli Distribution (Binary Variables)

- Independence
  - Observed Successes/Failures Independent of Each Other
  - Tied to the Binomial Distribution

- No Multicollinearity (Applies if Doing Multiple Logistic Regression)

Hypothesis Test and CI for Slope

- Hypotheses
  - $H_0: \beta_1 = 0$
  - $H_a: \beta_1 \neq 0$

- Test Statistic
  - $z^* = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$
- P-Value
  - Use Standard Normal Distribution
  - Use **2*(1-pnorm(abs(zstar),mean=0,sd=1))** Function in R

Hypothesis Test and CI for Slope

- Decision – Same As Always
- Interpret – Similar to Interpretation of Test from SLR
- Alternative: Confidence Interval
  - $\hat{\beta}_1 \pm 1.96 * SE_{\hat{\beta}_1}$
  - Does it Contain 0? Yes or No?
- CI for Odds Ratio – Exponentiate Both Bounds of CI

Likelihood Ratio Test

- Tests Overall Effectiveness of the Model
- Hypothesis Test for Comparing Empty Model to Full Model
- Similar to F-test in Linear Regression
- Almost Equivalent to Previous Hypothesis Test (P-values Similar)
- Let *L* Represent the Likelihood of our Model – We Want to Maximize

Likelihood Ratio Test

- The **glm()** Function Minimizes $-2 * \log(L)$ (Same as Maximizing L)
- The **glm()** Function also Estimates L_0 Which is the Likelihood of the Constant Model or Empty Model (Only an Intercept)
- Effectiveness of Model Can Be Measured by the Test Statistic
  - $G^* = -2 * \log(L_0) - (-2 * \log(L))$
  - Notice: $G^* = -2(\log(L_0) - \log(L)) = -2 * \log\left(\frac{L_0}{L}\right)$

Likelihood Ratio Test

- P-value
  - Use Chi-Squared Distribution
  - Degrees of Freedom for Chi-Squared is 1 When Full Model has 1 Predictor

- Hypotheses the Same as Previous Test
  - $H_0: \beta_1 = 0$
  - $H_a: \beta_1 \neq 0$

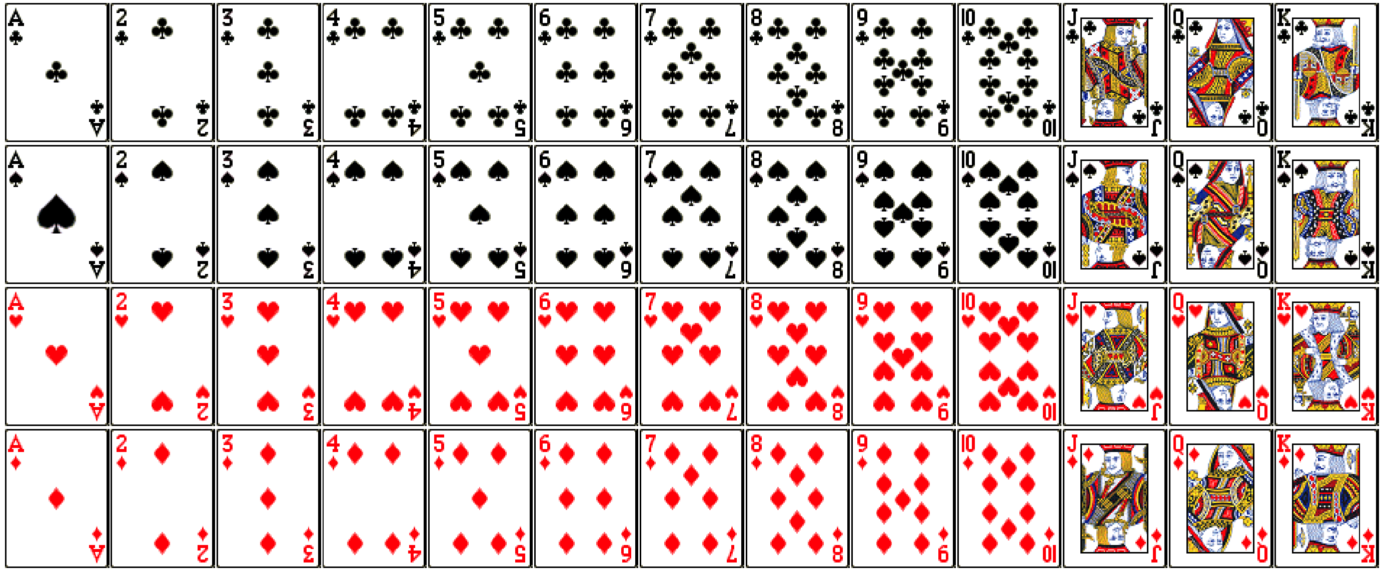
- Testing Same Hypothesis Test but Trust LRT Over Previous Test

Supplement for Lecture 27

- Examine Output from Logistic Regression
- Get Confidence Intervals for Slope
- Get Confidence Intervals for Odds Ratio
- Perform Likelihood Ratio Test

Maximizing Likelihood

- Suppose There are Three Decks of Playing Cards
  - Standard 52 Card Deck
  - Euchre Deck (9,10,J,Q,K)
  - Only Red Cards from the Deck (26 Cards)



Maximizing Likelihood

- Sample 2 Cards: Get Jack of Hearts and then the Jack of Diamonds

- Let L Represent the Likelihood of Our Sample

- Probability of the Sample Under All Three Situations
  - Full Deck: $L = \left(\frac{1}{52}\right) * \left(\frac{1}{51}\right) \approx 0.00038$
  - Euchre Deck: $L = \left(\frac{1}{24}\right) * \left(\frac{1}{23}\right) \approx 0.0018$ ← Most Likely
  - Red Deck: $L = \left(\frac{1}{26}\right) * \left(\frac{1}{25}\right) \approx 0.0015$

Thank You

Make Reasonable Decisions