# Logistic Regression - Understanding Model Selection Criteria

Ryan Miller

# Introduction

- Logistic regression is a generalized linear model for binary outcomes
  - As is the case for all GLMs, logistic regression involves a probability distribution and therefore has a *likelihood*
- An implication is that most of the model selection approaches we covered while studying multiple regression can be applied to logistic regression

# Model Selection Criteria

Both the AIC and BIC model selection criteria can be directly applied to the logistic regression setting:

- $AIC = -\text{Log-Likelihood} + 2k$
- $BIC = -\text{Log-Likelihood} + log(n) * k$

Recall that both criteria aim to balance a model's *goodness of fit* (measured by the log-likelihood) and its complexity (measured by $k$, the number of model parameters)

# Maximum Likelihood Estimation

- In *maximum likelihood estimation*, the goal is to solve for a set of parameters that maximize the conditional probability of observing the sample data given a specified probability model

# Maximum Likelihood Estimation

- In *maximum likelihood estimation*, the goal is to solve for a set of parameters that maximize the conditional probability of observing the sample data given a specified probability model
- In logistic regression, each observed outcome follows a *Bernoulli distribution*, which is just a *binomial distribution* with $n = 1$ and a success probability of $\pi$
  - Maximum likelihood estimation is the basis for the log-likelihood in criteria like AIC and BIC
  - Today we'll work through a brief example akin to an intercept only model (which implies all subjects having the same success probability)

# Bernoulli Variables

- Suppose we observe a single outcome, $y_1 = 1$, how *likely* was this outcome?

# Bernoulli Variables

- Suppose we observe a single outcome, $y_1 = 1$, how *likely* was this outcome?
  - The probability of observing $y_1 = 1$ is $\pi$
- Similarly, if we observe a second outcome, $y_2 = 0$, the probability of seeing this outcome is $1 - \pi$

# Bernoulli Variables

- Suppose we observe a single outcome, $y_1 = 1$, how *likely* was this outcome?
  - The probability of observing $y_1 = 1$ is $\pi$
- Similarly, if we observe a second outcome, $y_2 = 0$, the probability of seeing this outcome is $1 - \pi$
- The likelihood function describes the *joint probability* of *all of the observed data*
  - If the data-points are independent, it can be expressed as a product of individual likelihoods:

$$P(\mathbf{y}) = P(y_1) * P(y_2) * \ldots * P(y_n)$$

# Likelihood

We're now ready to define the likelihood function:

$$
\begin{aligned}
L(\mathbf{y}|\pi) &= P(y_1) * P(y_2) * \ldots * P(y_n) \\
&= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}
\end{aligned}
$$

# Likelihood

We're now ready to define the likelihood function:

$$\begin{aligned} L(\mathbf{y}|\pi) &= P(y_1) * P(y_2) * \ldots * P(y_n) \\ &= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \end{aligned}$$

Logarithms are one-to-one, monotone transformations, so there's no difference in maximizing the likelihood or the log-likelihood:

$$\begin{aligned} l(\mathbf{y}|\pi)) &= log(L(\mathbf{y}|\pi)) \\ &= \sum_{i=1}^{n} log(\pi^{y_i}(1-\pi)^{1-y_i}) \\ &= log(\pi)\sum_{i=1}^{n} y_i + log(1-\pi)\sum_{i=1}^{n}(1-y_i) \end{aligned}$$

# Maximizing the Likelihood

- The goal of maximum likelihood estimation is to find a value of the parameter $\pi$ that maximizes $l(\mathbf{y}|\pi))$
  - Not surprisingly, this can be down by differentiating with respect to $\pi$, setting the resulting expression equal to zero, then solving for the maximizer

# Maximizing the Likelihood

- The goal of maximum likelihood estimation is to find a value of the parameter $\pi$ that maximizes $l(\mathbf{y}|\pi))$
  - Not surprisingly, this can be down by differentiating with respect to $\pi$, setting the resulting expression equal to zero, then solving for the maximizer

$$l(\mathbf{y}|\pi)) = log(\pi)\sum_{i=1}^{n} y_i + log(1-\pi)\sum_{i=1}^{n}(1-y_i)$$

$$\frac{\partial l}{\partial \pi} = \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{sum_{i=1}^{n}(1-y_i)}{1-\pi} \stackrel{\text{set}}{=} 0$$

# Maximizing the Likelihood

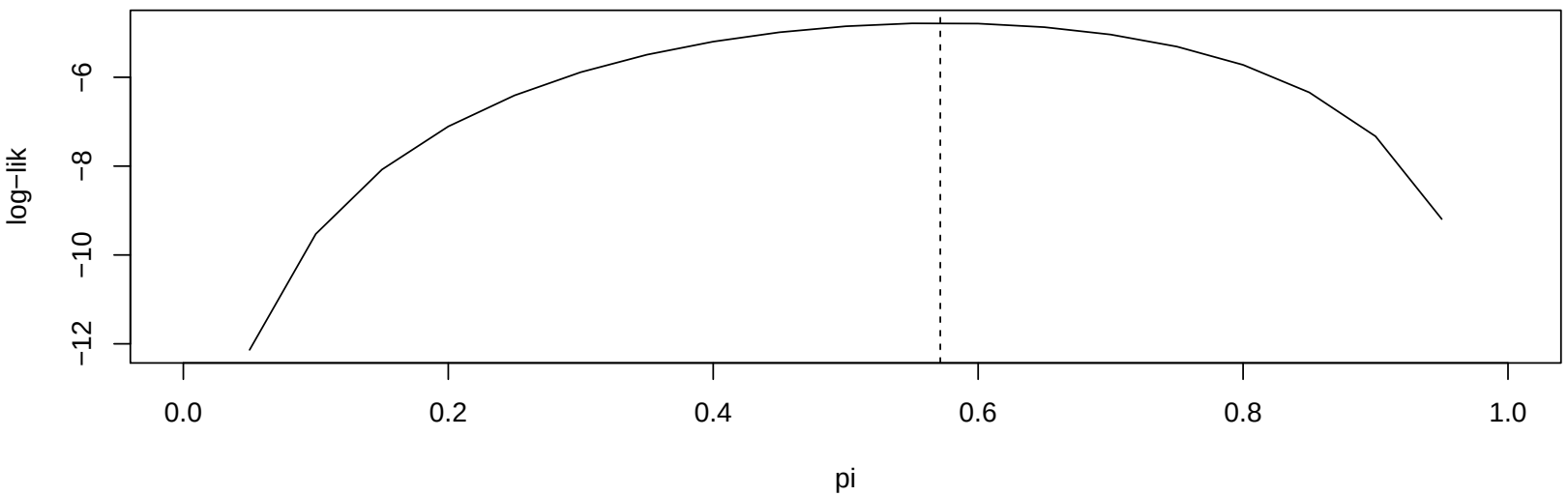
- It's easy to use algebra to solve for a closed form expression of $\hat{\pi}$, the value of $\pi$ that maximizes the likelihood of the observed data
  - I'll skip this, but the result should be unsurprising, the sample proportion $\sum_{i=1}^{n} y_i/n$ is maximizer
- Instead, let's look at a graph of the likelihood

# Maximizing the Likelihood

```r
## Sample data
y <- c(1,1,1,0,0,1,0)

## Define log-likelihood function
log_lik <- function(pi, y){
  log(pi)*sum(y) + log(1 - pi)*sum(1 - y)
}

## Plot the log-likelihood over all possible values of pi
pi_seq <- seq(0,1, by = 0.05)
plot(pi_seq, log_lik(pi = pi_seq, y = y), type = "l", ylab = "log-lik", xlab = "pi")
abline(v = sum(y)/length(y), lty = 2) ## sample proportion
```



# Comparing Different Models

- The likelihood of the observed data is maximized when $\pi = \sum_{i=1}^{n} y_i / n$
  - Notice many other values of $\pi$ near the sample proportion will also fit the data almost as well
  - These might be considered reasonable models for the observed phenomenon

# Comparing Different Models

- The likelihood of the observed data is maximized when $\pi = \sum_{i=1}^{n} y_i / n$
  - Notice many other values of $\pi$ near the sample proportion will also fit the data almost as well
  - These might be considered reasonable models for the observed phenomenon
- In contrast, values closer to zero have a substantially lower likelihood, and therefore represent models that do not fit the sample data very well

# Remarks

- The specific numeric value of the log-likelihood doesn't much in an absolute sense, but it means a lot in a relative one
  - So long as the data and the underlying probability distribution remain the same, the log-likelihood can be used to compare the relative fit of different proposed model to the sample data

# Remarks

- The specific numeric value of the log-likelihood doesn't much in an absolute sense, but it means a lot in a relative one
  - So long as the data and the underlying probability distribution remain the same, the log-likelihood can be used to compare the relative fit of different proposed model to the sample data
- Logistic regression involves an added layer of complexity beyond the example we looked at, as $\mathit{logit}(\pi) = \beta_0 + \beta_1 X_1 \ldots$
  - We'd now need to solve for a combination of parameter values that maximize the likelihood
  - There's no closed-form solution to this problem, but it's pretty easy for optimization algorithms to find it using numerical approaches

# Connection to AIC/BIC

- Hopefully this brief example provides some perspective on the log-likelihood component of the AIC and BIC model selection criteria
  - The main takeaway is that a model's log-likelihood is a relative measure describing how well it fits the sample data
- Model selection criteria, such as AIC or BIC, aim to balance fit with parsimony
  - They are suitable for comparing non-nested models with different levels of complexity
  - They also can form the basis of stepwise selection algorithms