Binary Regression

- Bernoulli Distribution -- distribution for a single seedling
  - SURV = 1 if seedling survives; survives with probability π
  - SURV = 0 otherwise
- Binomial Distribution -- distribution for T, total number of seedlings that survive out of n,
- Bin(n, π)
  - n = 3072
  - observe T = 755

$$\hat{\pi} = \frac{755}{3072} = 0.246$$

Estimation

How to construct an estimate of π ?

- Intuitively, the sample proportion seems like a good choice for the estimate of the population probability of seedling survival.
- But what should we do in general if we think that the probability of survival depends on other covariates?
- Use Maximum Likelihood Estimation

Maximum Likelihood Estimation

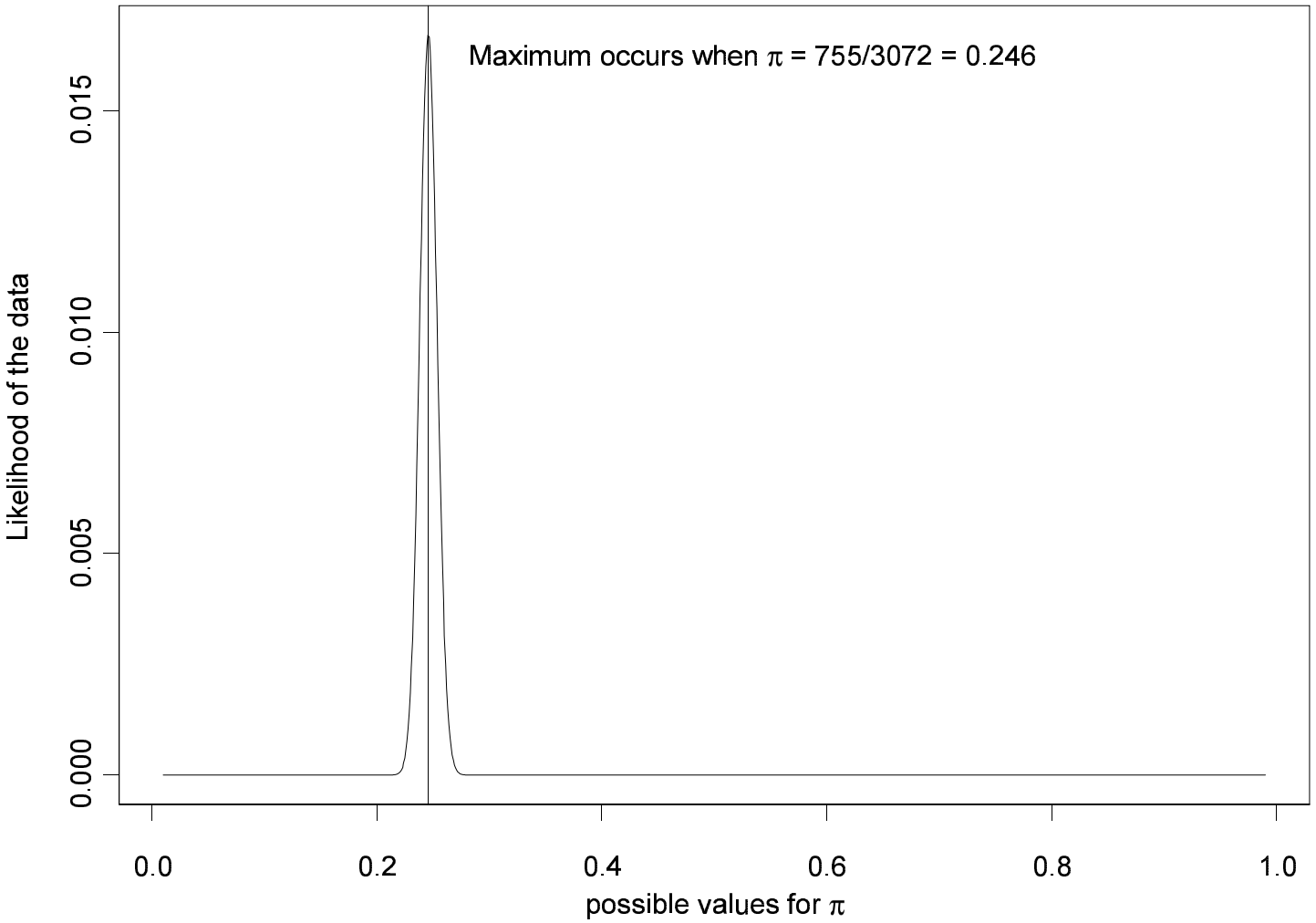
Probability that T = 755 is

$$P(T=755|\pi, n=3072)=\left(\frac{3072!}{755!(3072-755)!}\right)\pi^{755}(1-\pi)^{3072-755}$$

Goal: find the value of π that maximizes the probability or likelihood of the observed data

This approach is called maximum likelihood estimation

Plot of Binomial Distribution with T=755, n = 3072 for possible π



Sampling Distribution

$$\hat{\pi} \sim Normal\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

$$Standard\ Error\ of\ \hat{\pi} = SE(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}$$

$$95\%\ Confidence\ Interval\ \hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})}$$

Logistic Regression

How should we include covariates into π ?

For example, probability of survival may depend on whether there was a CAGE to prevent animals from eating the seedling - or - LIGHT levels, etc.

$$\pi(SURV \mid CAGE, LIGHT) = \beta_0 + \beta_1 CAGE + \beta_2 LIGHT$$

mean π,

use OLS; regress SURV on CAGE, LiGHT?

What are 2 problems with using linear regression?

Logits

To build in the necessary constraints that probabilities are between 0 and 1 convert to log-odds or logits

- Odds of survival = $\pi/(1-\pi)$

$$logit(\pi) \stackrel{\text{def}}{=} \log\left(\frac{\pi}{(1-\pi)}\right) = \beta_0 + \beta_1 CAGE + \beta_2 LIGHT = \eta$$

- η is the linear predictor
- logit is the link function that relates the mean π to the linear predictor η

Generalized **L**inear **M**odels or GLMs

Logits

- To convert from the linear predictor η to the mean π, we use the inverse transformation
- odds that (SURV = 1) = $\exp(\eta) = \omega$
- π = odds/(1 + odds)

  $= \exp(\eta)/(1 + \exp(\eta))$

  $= \omega/(1 + \omega)$
- $\omega = \pi/(1 - \pi)$

Interpretation

$$w = \exp(\beta_0 + \beta_1 CAGE + \beta_2 LIGHT)$$

- When all explanatory variables are 0 (CAGE = 0, LIGHT = 0), the odds of survivival, are $\exp(\beta_0)$
- The ratio of odds (or odds ratio) at $X_1 = A$ to $X_1 = B$, for fixed values of the other explanatory variables is

$$\frac{\omega_A}{\omega_B} = \exp(\beta_j (A - B))$$

$$\frac{\omega_A}{\omega_B} = \exp(\beta_j) \; if \; A - B = 1$$

Interpretation

- Coefficient for CAGE = 0.79
- So if CAGE increases from 0 to 1, the odds of survival will change by exp(0.79) = 2.20
- The odds of survivial for seedlings in a Cage are 2.20 times higher than the odds for surviving for seedlings left exposed.

Confidence Intervals and Testing

- Maximum Likelihood Estimates are approximately normally distributed
    - mean β
    - estimated variance $SE(\beta)^2$
- 95% Confidence Interval use normal quantiles
- t-value has an approximate normal distribution under Ho: $\beta = 0$

Example

Confidence Interval for Coefficient for CAGE