bernoulliBinomialMLE

October 21, 2019

0.1 Goal

Our goal for this class is to understand how to compute the optimal beta coefficients for logistic regression. We'll need a bit of background. Logistic regression attempts to predict a binary variable and so we'll need to understand distributions related to binary variables. To find the optimal coefficients, we'll maximize a function called the likelihood. The below summarized distributions of binary variables, the maximum likelihood approach in general, and as it relates to logistic regression (linear regression is included as an aside).

0.2 Bernoulli distributed random variables

A random variable X is Bernoulli distributed is can take two values: zero and one. The variable X equals 1 with probability p and X equals 0 with probability $1 - p$.

The expected value—the average over all values the random variable can take weighted by their corresponding probabilities—equals p.

$$E(X) = \mu = p \times 1 + (1 - p) \times 0$$

The variance—or average squared difference around the expected value—equals $p(1 - p)$.

$$\begin{aligned} \text{Var}(X) &= E(X - \mu)^2 = E(X - p)^2 = p \times (1 - p)^2 + (1 - p) \times (0 - p)^2 && (1) \\ &= p(1 - p)\left[(1 - p) + p\right] = p(1 - p) && (2) \end{aligned}$$

The data we study with logistic regression can be viewed as a stream of 1s and 0s together with covariate data. If we assume our data is generated by a Bernoulli-distributed variable, then the goal is to estimate p (the only parameter) and how covariate information changes p.

0.3 Maximum likelihood

A model assigns a probability to possible events we may see in our data and can depend on a set of parameters. For example, a model can assign a probability to seeing a stream of data $y_1, y_2, \cdots y_n$.

$$p(y_1, y_2, \cdots y_n | \theta)$$

The vector θ represents all the parameters we need to estimate in our model.

For example, a linear regression *model* assumes individual y values are normally distributed with constant variance

$$p(y_i|\theta) = p(y_i|\beta, \sigma^2) \sim \mathcal{N}(X_i\beta, \sigma^2)$$

If we assume that the observations in our dataset are independent, a theorem from probability tells us we can compute the above probability by multiplying the probabilities of observing each individual y_i value.

$$p(y_1, y_2, \cdots y_n) = p(y_1) \times p(y_2) \times \cdots \times p(y_n) = \prod_{i=1}^{n} p(y_i)$$

In the above linear regression, this product would equal

$$p(y_1, y_2, \cdots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\beta, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}(X_i\beta, \sigma^2)$$

We can assume the data is fixed in the above probability model and treat this model as a function of the parameters β. This function is called a **likelihood function**. Intuitively, this function ask "How likely is it that the parameter values θ generated this data?"

0.4 Maximum likelihood for simple linear regression

A linear regression model assumes N data points (x_i, y_i) come from the following distribution

$$y_i \sim N(\beta' x_i, \sigma^2)$$

The probability model above assumes β and σ are constants. Assuming the (x,y) observations are independent, the likelihood function takes the following form

$$\begin{aligned}
p(y_1, y_2, \cdots, y_n|\beta, \sigma) &= \prod_{i=1}^{N} \mathcal{N}(X_i\beta, \sigma^2) && (3)\\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - [\beta' x_i])^2}{2\sigma^2}\right\} && (4)\\
&= \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^N \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - [\beta' x_i])^2\right\} && (5)
\end{aligned}$$

Assume σ is known. The parameter β only appears in the exponential above, and is maximized if we maximize the exponential's argument

$$f(\beta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - [\beta' x_i])^2$$

or minimize

$$g(\beta) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - [\beta' x_i])^2$$

The function g is the sum squares error and minimized when

$$\beta = (X'X)^{-1} X'Y$$

Minimizing the sum squares error is equivalent to maximizing the above likelihood.

0.5 Maximum likelihood for logistic regression

A logistic regression model assumes the N data points follow a Bernoulli distribution

$$y_i \sim \text{Bernoulli}\left(\frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} \right)$$

The probability assumes β is a constant. If the probability of a "1" is

$$p(y_i = 1) = \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}}$$

then the probability of "0" is $1 - p(1)$ or

$$p(y_i = 0) = 1 - p(y_i = 1) = 1 - \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} \tag{6}$$

$$p(y_i = 0) = \frac{1}{1 + e^{\beta' x_i}} \tag{7}$$

The likelihood function equals

$$p(y_1, y_2, \cdots, y_n | \beta) = \left[\frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} \right]^{z_i} \left[\frac{1}{1 + e^{\beta' x_i}} \right]^{1 - z_i}$$

where

$$z_i = \begin{cases} 1 & \text{when } y_i = 1 \\ 0 & \text{when } y_i = 0 \end{cases}$$