Optimization applied 1.5: Maximum likelihood estimation continued

Marcin Lewandowski

November 28, 2024

Last meeting:

1. We have assumed that the data is generated from Bernoulli:

Last meeting:

1. We have assumed that the data is generated from Bernoulli:
   - $Y \sim Bern(p)$

$$f\left(y;p\right) = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases}$$

Last meeting:

1. We have assumed that the data is generated from Bernoulli:
    - $Y \sim Bern(p)$

$$f\left(y;p\right)=\begin{cases} p & \text{if } y=1 \\ 1-p & \text{if } y=0 \end{cases}$$

    - The same can be also written as:

$$f\left(y;p\right)=p^{y}\left(1-p\right)^{1-y}$$

Last meeting:

1. We have assumed that the data is generated from Bernoulli:
   - $Y \sim Bern(p)$

   $$f\left(y;p\right)=\begin{cases} p & \text{if } y=1 \\ 1-p & \text{if } y=0 \end{cases}$$

   - The same can be also written as:

   $$f\left(y;p\right)=p^{y}\left(1-p\right)^{1-y}$$

   - Examples:

Last meeting:

1. We have assumed that the data is generated from Bernoulli:
   - $Y \sim Bern(p)$

$$f(y;p) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases}$$

   - The same can be also written as:

$$f(y;p) = p^y (1-p)^{1-y}$$

   - Examples:
     - Whether a person has a disease or not

Last meeting:

1. We have assumed that the data is generated from Bernoulli:
    - $Y \sim Bern(p)$

$$f(y; p) = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases}$$

    - The same can be also written as:

$$f(y; p) = p^y (1 - p)^{1-y}$$

    - Examples:
        - Whether a person has a disease or not
        - Whether a person will default on a loan or not

Last meeting:

1. We have assumed that the data is generated from Bernoulli:
   - $Y \sim Bern(p)$

   $$f(y;p) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases}$$

   - The same can be also written as:

   $$f(y;p) = p^y (1-p)^{1-y}$$

   - Examples:
     - Whether a person has a disease or not
     - Whether a person will default on a loan or not
     - Whether a person will get to college or not

Last meeting:

1. We have assumed that the data is generated from Bernoulli:
   - $Y \sim Bern(p)$

$$f(y;p) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases}$$

   - The same can be also written as:

$$f(y;p) = p^{y}(1-p)^{1-y}$$

   - Examples:
     - Whether a person has a disease or not
     - Whether a person will default on a loan or not
     - Whether a person will get to college or not
2. Next: we have assumed that we **observe** a vector of data.

Last meeting:

1. We have assumed that the data is generated from Bernoulli:
    - $Y \sim Bern(p)$

$$f(y;p) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases}$$

    - The same can be also written as:

$$f(y;p) = p^y (1-p)^{1-y}$$

    - Examples:
        - Whether a person has a disease or not
        - Whether a person will default on a loan or not
        - Whether a person will get to college or not
2. Next: we have assumed that we **observe** a vector of data.
3. And want to estimate the underlying parameter p.

Last meeting:

We have defined the likelihood function:

- y_i are independent
- The likelihood of observing $\mathbf{y} = \{1, 1, 1, 0, 1, 0\}$ is:

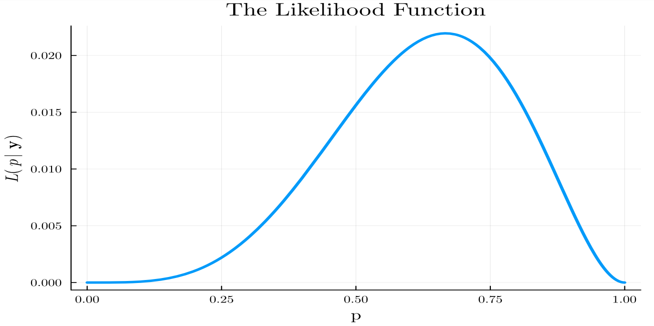
$$L(p \mid \mathbf{y}) = \prod_{i=1}^{6} p^{y_i} (1-p)^{1-y_i}$$

Last meeting:

We have defined the likelihood function:

- y_i are independent
- The likelihood of observing $\mathbf{y} = \{1, 1, 1, 0, 1, 0\}$ is:

$$L\left(p \mid \mathbf{y}\right) = \prod_{i=1}^{6} p^{y_i}(1-p)^{1-y_i}$$



MLE: Log likelihood function (Bernoulli)

$$L\left(p \mid \mathbf{y}\right) = \prod_{i=1}^{n} p^{y_i} \cdot \left(1 - p\right)^{1 - y_i}$$

Necessary condition for maxima:

MLE: Log likelihood function (Bernoulli)

$$L\left(p \mid \mathbf{y}\right) = \prod_{i=1}^{n} p^{y_i} \cdot \left(1 - p\right)^{1 - y_i}$$

$$\log\left(L\left(p \mid \mathbf{y}\right)\right) = \sum_{i=1}^{n} \log\left(p^{y_i}\right) + \sum_{i=1}^{n} \log\left(\left(1 - p\right)^{1 - y_i}\right)$$

Necessary condition for maxima:

MLE: Log likelihood function (Bernoulli)

$$
\begin{aligned}
L\left(p \mid \mathbf{y}\right) &= \prod_{i=1}^{n} p^{y_i} \cdot \left(1-p\right)^{1-y_i} \\
\log\left(L\left(p \mid \mathbf{y}\right)\right) &= \sum_{i=1}^{n} \log\left(p^{y_i}\right) + \sum_{i=1}^{n} \log\left(\left(1-p\right)^{1-y_i}\right) \\
&= \sum_{i=1}^{n} y_i \log\left(p\right) + \sum_{i=1}^{n} \left(1-y_i\right) \log\left(1-p\right)
\end{aligned}
$$

Necessary condition for maxima:

MLE: Log likelihood function (Bernoulli)

$$
\begin{aligned}
L(p \mid \mathbf{y}) &= \prod_{i=1}^{n} p^{y_i} \cdot (1-p)^{1-y_i} \\
\log(L(p \mid \mathbf{y})) &= \sum_{i=1}^{n} \log(p^{y_i}) + \sum_{i=1}^{n} \log\left((1-p)^{1-y_i}\right) \\
&= \sum_{i=1}^{n} y_i \log(p) + \sum_{i=1}^{n} (1-y_i) \log(1-p) \\
&= \log(p) \sum_{i=1}^{n} y_i + \log(1-p) \sum_{i=1}^{n} (1-y_i)
\end{aligned}
$$

Necessary condition for maxima:

MLE: Log likelihood function (Bernoulli)

$$
\begin{aligned}
L(p \mid \mathbf{y}) &= \prod_{i=1}^{n} p^{y_i} \cdot (1-p)^{1-y_i} \\
\log(L(p \mid \mathbf{y})) &= \sum_{i=1}^{n} \log(p^{y_i}) + \sum_{i=1}^{n} \log\left((1-p)^{1-y_i}\right) \\
&= \sum_{i=1}^{n} y_i \log(p) + \sum_{i=1}^{n} (1-y_i) \log(1-p) \\
&= \log(p) \sum_{i=1}^{n} y_i + \log(1-p) \sum_{i=1}^{n} (1-y_i)
\end{aligned}
$$

Necessary condition for maxima:

$$\frac{\partial \log(L(p \mid y))}{\partial p} = 0$$

MLE: Log likelihood function (Bernoulli)

$$
\begin{aligned}
L(p \mid \mathbf{y}) &= \prod_{i=1}^{n} p^{y_i} \cdot (1-p)^{1-y_i} \\
\log(L(p \mid \mathbf{y})) &= \sum_{i=1}^{n} \log\left(p^{y_i}\right) + \sum_{i=1}^{n} \log\left((1-p)^{1-y_i}\right) \\
&= \sum_{i=1}^{n} y_i \log(p) + \sum_{i=1}^{n} (1-y_i) \log(1-p) \\
&= \log(p) \sum_{i=1}^{n} y_i + \log(1-p) \sum_{i=1}^{n} (1-y_i)
\end{aligned}
$$

Necessary condition for maxima:

$$
\begin{aligned}
\frac{\partial \log(L(p \mid y))}{\partial p} &= 0 \\
\frac{\sum_{i=1}^{n} y_i}{p} - \frac{\sum_{i=1}^{n} (1-y_i)}{1-p} &= 0
\end{aligned}
$$

But what if we want to model the probability:

How to model *probability* of a binary outcome, $Y \in \{0, 1\}$.
Recall that Y can be:

- Whether a person has a disease or not
- Whether a person will default on a loan or not
- Whether a person will get to college or not

But what if we want to model the probability:

How to model *probability* of a binary outcome, $Y \in \{0, 1\}$.
Recall that Y can be:

- Whether a person has a disease or not
- Whether a person will default on a loan or not
- Whether a person will get to college or not

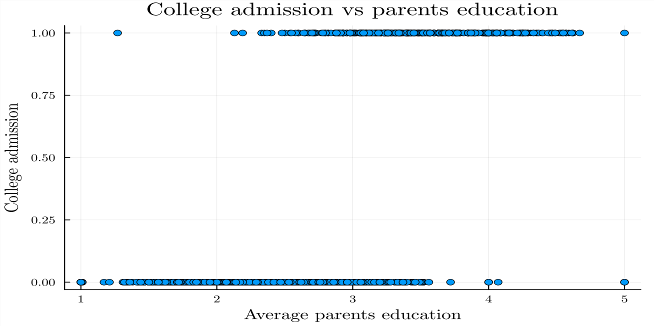
Let's assume that $P(Y = y_i \mid x)$ depends on a single predictor, x.

Making it more concrete:

Suppose you were asked to model the relationship between the probability of getting to college and the (average) parents education.

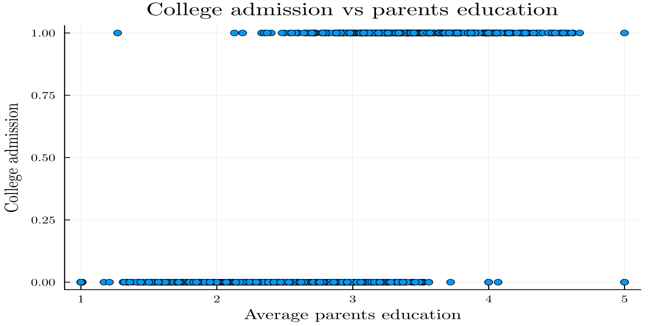
Making it more concrete:

Suppose you were asked to model the relationship between the probability of getting to college and the (average) parents education.



Making it more concrete:

Suppose you were asked to model the relationship between the probability of getting to college and the (average) parents education.



In this case:

- Y is whether a person gets to college or not
- x is the average parents education

The Logit Model:

- Let's assume that $P\left(Y = y_i \mid x\right)$ depends on a single predictor, x.

The Logit Model:

- Let's assume that $P\left(Y = y_i \mid x\right)$ depends on a single predictor, x.
- Thus p is some function of x!

The Logit Model:

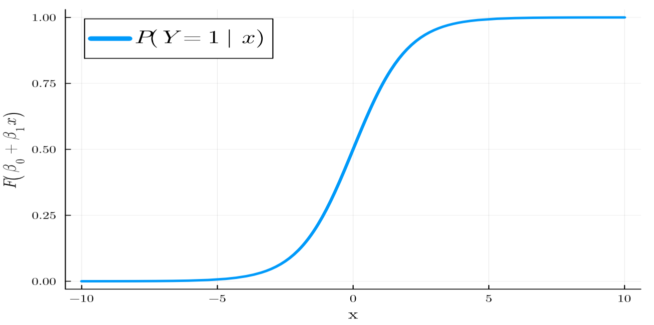
- Let's assume that $P(Y = y_i \mid x)$ depends on a single predictor, x.
- Thus p is some function of x!
- $P(Y = y_i \mid x_i) = F(\beta_0 + \beta_1 x)$

The Logit Model:

- Let's assume that $P(Y = y_i \mid x)$ depends on a single predictor, x.
- Thus p is some function of x!
- $P(Y = y_i \mid x_i) = F(\beta_0 + \beta_1 x)$

The Logit Model:

- Let's assume that $P(Y = y_i \mid x)$ depends on a single predictor, x.
- Thus p is some function of x!
- $P(Y = y_i \mid x_i) = F(\beta_0 + \beta_1 x)$



The Logit Model:

$$
\begin{aligned}
P\left(y=1 \mid x\right) &= F\left(\beta_0+\beta_1 x\right) \\
&= \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \\
&= \frac{1}{\frac{1}{e^{\beta_0+\beta_1 x}}+1}
\end{aligned}
$$

The Logit Model:

$$
\begin{aligned}
P\left(y=1 \mid x\right) &= F\left(\beta_0+\beta_1 x\right) \\
&= \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \\
&= \frac{1}{\frac{1}{e^{\beta_0+\beta_1 x}}+1}
\end{aligned}
$$

$$
\begin{aligned}
P\left(y=1 \mid x\right) &= 1-F\left(\beta_0+\beta_1 x\right) \\
&= 1-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \\
&= \frac{1}{1+e^{\beta_0+\beta_1 x}}
\end{aligned}
$$

The Logit Model:

Suppose we want to know the probability of $Y = y_i$, where $y_i \in \{0, 1\}$.

$$P\left(Y = y_i \mid x_i\right) = P\left(Y = y_i \mid x\right)^{y_i} P\left(Y = 1 - y_i \mid x\right)^{1-y_i}$$

The Logit Model:

Suppose we want to know the probability of $Y = y_i$, where $y_i \in \{0, 1\}$.

$$P\left(Y = y_i \mid x_i\right) = P\left(Y = y_i \mid x\right)^{y_i} P\left(Y = 1 - y_i \mid x\right)^{1-y_i}$$

Recall that in the Bernoulli case this was simply:

$$P\left(Y = y_i\right) = p^{y_i}\left(1 - p\right)^{y_i}$$

The Logit Model:

Suppose we want to know the probability of $Y = y_i$, where $y_i \in \{0, 1\}$.

$$P\left(Y = y_i \mid x_i\right) = P\left(Y = y_i \mid x\right)^{y_i} P\left(Y = 1 - y_i \mid x\right)^{1-y_i}$$

Recall that in the Bernoulli case this was simply:

$$P\left(Y = y_i\right) = p^{y_i}\left(1 - p\right)^{y_i}$$

In logit model:

$$P\left(Y = y_i \mid x_i\right) = \left(\frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta_0+\beta_1 x}}\right)^{1-y_i}$$

The Logit Model:

Suppose we want to know the probability of $Y = y_i$, where $y_i \in \{0, 1\}$.

$$P\left(Y = y_i \mid x_i\right) = P\left(Y = y_i \mid x\right)^{y_i} P\left(Y = 1 - y_i \mid x\right)^{1-y_i}$$

Recall that in the Bernoulli case this was simply:

$$P\left(Y = y_i\right) = p^{y_i}\left(1 - p\right)^{y_i}$$

In logit model:

$$P\left(Y = y_i \mid x_i\right) = \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x}}\right)^{1-y_i}$$

The likelihood function is:

$$L\left(\beta_0, \beta_1 \mid \mathbf{y}, \mathbf{x}\right) = \prod_{i=1}^{N} \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x}}\right)^{1-y_i}$$

The likelihood function:

$$L\left(\beta_0, \beta_1 \mid \mathbf{y}, \mathbf{x}\right) = \prod_{i=1}^{N} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta_0+\beta_1 x}}\right)^{1-y_i}$$

The likelihood function:

$$L\left(\beta_0, \beta_1 \mid \mathbf{y}, \mathbf{x}\right) = \prod_{i=1}^{N} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 x}}\right)^{1-y_i}$$

For example, suppose we have 3 observations:

- $y = [1, 0, 1]$
- $x = [5, 2, 1]$

The likelihood function:

$$L\left(\beta_0, \beta_1 \mid \mathbf{y}, \mathbf{x}\right) = \prod_{i=1}^{N} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 x}}\right)^{1-y_i}$$

For example, suppose we have 3 observations:

- $y = [1, 0, 1]$
- $x = [5, 2, 1]$

$$L\left(\beta_0, \beta_1 \mid \mathbf{y}, \mathbf{x}\right) = \left(\frac{e^{\beta_0+5\beta_1}}{1+e^{\beta_0+5\beta_1}}\right)\left(\frac{1}{1+e^{\beta_0+2\beta_1}}\right)\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)$$

The log likelihood function:

Given:

$$L\left(\beta_0, \beta_1 \mid \mathbf{y}, \mathbf{x}\right) = \prod_{i=1}^{N} \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x}} \right)^{1 - y_i}$$

The log likelihood function is:

$$\log\left(L\left(\beta_0, \beta_1 \mid \mathbf{y}, \mathbf{x}\right)\right) =$$
$$\sum_{i=1}^{N} \left\{ y_i \log\left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) + (1 - y_i) \log\left(\frac{1}{1 + e^{\beta_0 + \beta_1 x}} \right) \right\}$$

The log likelihood function

Note that if needed one can split the log likelihood function into two parts:

$$\log\left(L\left(\beta_0, \beta_1 \mid \mathbf{y}, \mathbf{x}\right)\right) = \sum_{i=1}^{N} \underbrace{y_i \log\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)}_{\text{The first part}}$$

$$+ \sum_{i=1}^{N} \underbrace{(1-y_i) \log\left(\frac{1}{1+e^{\beta_0+\beta_1 x}}\right)}_{\text{The second part}}$$