Optimization applied 1: Maximum likelihood estimation

Marcin Lewandowski

November 26, 2024

MLE: Intuition (Bernoulli)

Let $Y \sim Bern(p)$.

- Suppose we observe 6 i.i.d coin flips (1 for heads, 0 for tails):
$$\mathbf{y} = \{1, 1, 1, 0, 1, 0\}$$
- Denote each flip as $y_1, y_2, \ldots, y_6$, where $y_1 = 1$ and $y_6 = 1$.

MLE: Intuition (Bernoulli)

Let $Y \sim Bern(p)$.

- Suppose we observe 6 i.i.d coin flips (1 for heads, 0 for tails):

$$\mathbf{y} = \{1, 1, 1, 0, 1, 0\}$$

- Denote each flip as $y_1, y_2, \ldots, y_6$, where $y_1 = 1$ and $y_6 = 1$.

Recall: the probability mass function (PMF) of a Bernoulli r.v. is:

$$f(y; p) = p^y (1-p)^{1-y}, \quad y \in \{0, 1\}$$

The probability of 1 is p, while the probability of 0 is $(1-p)$.

MLE: Intuition (Bernoulli)

Let $Y \sim Bern(p)$.

- Suppose we observe 6 i.i.d coin flips (1 for heads, 0 for tails):

$$\mathbf{y} = \{1, 1, 1, 0, 1, 0\}$$

- Denote each flip as $y_1, y_2, \ldots, y_6$, where $y_1 = 1$ and $y_6 = 1$.

Recall: the probability mass function (PMF) of a Bernoulli r.v. is:

$$f(y; p) = p^y (1 - p)^{1-y}, \quad y \in \{0, 1\}$$

The probability of 1 is p, while the probability of 0 is $(1 - p)$.

Overarching Goal: Estimate p given a vector of observed data.

MLE: Intuition (Bernoulli)

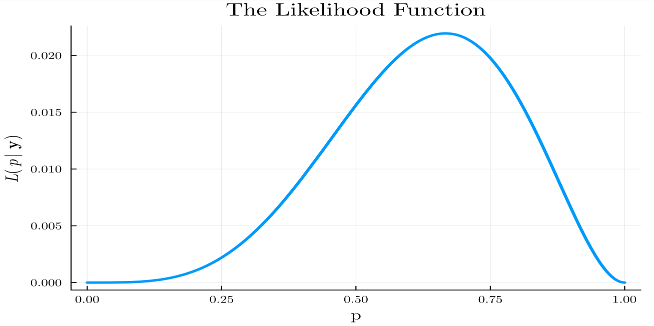
- y_i are independent
- The likelihood of observing $\mathbf{y} = \{1, 1, 1, 0, 1, 0\}$ is:

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

MLE: Intuition (Bernoulli)

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

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



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)

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

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

Necessary condition for maxima:

MLE: Log likelihood function (Bernoulli)

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

Necessary condition for maxima:

MLE: Log likelihood function (Bernoulli)

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

Necessary condition for maxima:

MLE: Log likelihood function (Bernoulli)

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

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

Necessary condition for maxima:

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

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:

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

MLE: Log likelihood function (Bernoulli) maximized

Note: In this simple example we can solve for p!

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

MLE: Log likelihood function (Bernoulli) maximized

Note: In this simple example we can solve for p!

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

MLE: Log likelihood function (Bernoulli) maximized

Note: In this simple example we can solve for p!

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

$$\frac{\sum_{i=1}^{n} y_i}{p} = \frac{\sum_{i=1}^{n}\left(1 - y_i\right)}{1 - p}$$

$$\text{Using: } y = \{1, 1, 1, 0, 1, 0\}$$

MLE: Log likelihood function (Bernoulli) maximized

Note: In this simple example we can solve for p!

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

$$\frac{\sum_{i=1}^{n} y_i}{p} = \frac{\sum_{i=1}^{n}\left(1 - y_i\right)}{1-p}$$

$$\text{Using: } y = \{1, 1, 1, 0, 1, 0\}$$

$$\frac{4}{p} = \frac{2}{1-p}$$

MLE: Log likelihood function (Bernoulli) maximized

Note: In this simple example we can solve for p!

$$
\begin{aligned}
\frac{\partial \log\left(L\left(p \mid \mathbf{y}\right)\right)}{\partial p} &= 0 \\
\frac{\sum_{i=1}^{n} y_i}{p} &= \frac{\sum_{i=1}^{n}\left(1 - y_i\right)}{1-p} \\
\text{Using: } y &= \{1, 1, 1, 0, 1, 0\} \\
\frac{4}{p} &= \frac{2}{1-p} \\
4 - 4p &= 2p
\end{aligned}
$$

MLE: Log likelihood function (Bernoulli) maximized

Note: In this simple example we can solve for p!

$$\begin{aligned} \frac{\partial \log\left(L\left(p \mid \mathbf{y}\right)\right)}{\partial p} &= 0 \\ \frac{\sum_{i=1}^{n} y_i}{p} &= \frac{\sum_{i=1}^{n}\left(1 - y_i\right)}{1-p} \\ \text{Using: } y &= \{1,1,1,0,1,0\} \\ \frac{4}{p} &= \frac{2}{1-p} \\ 4 - 4p &= 2p \\ -6p &= -4 \end{aligned}$$

MLE: Log likelihood function (Bernoulli) maximized

Note: In this simple example we can solve for p!

$$
\begin{aligned}
\frac{\partial \log\left(L\left(p \mid \mathbf{y}\right)\right)}{\partial p} &= 0 \\
\frac{\sum_{i=1}^{n} y_i}{p} &= \frac{\sum_{i=1}^{n} \left(1 - y_i\right)}{1 - p} \\
\text{Using: } y &= \{1, 1, 1, 0, 1, 0\} \\
\frac{4}{p} &= \frac{2}{1 - p} \\
4 - 4p &= 2p \\
-6p &= -4 \\
\widehat{p} &= \frac{4}{6}
\end{aligned}
$$

MLE: Log likelihood function (Bernoulli) maximized

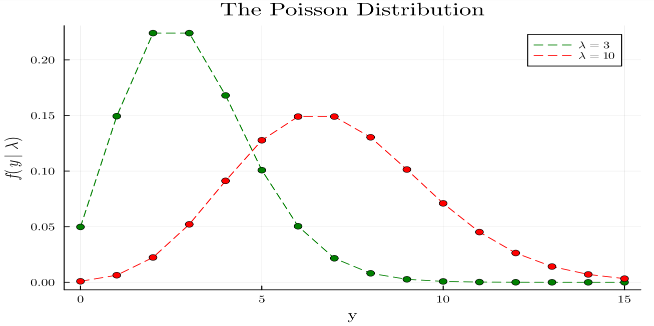
Note: In this simple example we can solve for p!

$$\begin{aligned}
\frac{\partial \log\left(L\left(p \mid \mathbf{y}\right)\right)}{\partial p} &= 0 \\
\frac{\sum_{i=1}^{n} y_i}{p} &= \frac{\sum_{i=1}^{n}\left(1 - y_i\right)}{1-p} \\
\text{Using: } y &= \{1, 1, 1, 0, 1, 0\} \\
\frac{4}{p} &= \frac{2}{1-p} \\
4 - 4p &= 2p \\
-6p &= -4 \\
\widehat{p} &= \frac{4}{6}
\end{aligned}$$

Let's see how to implement MLE in Julia!

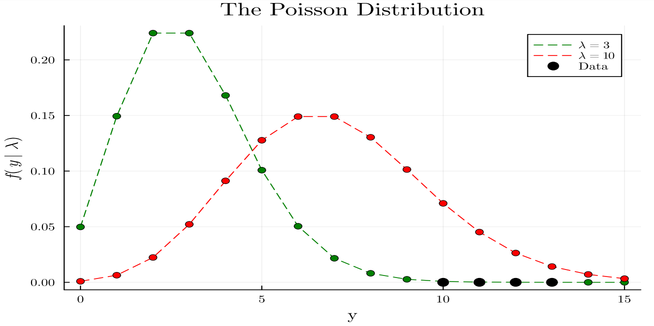
A task for you: the Poisson Distribution

$$f\left(y \mid \lambda\right) = \frac{e^{-\lambda}\lambda^{y}}{y!} \; ; \;\; y \in \{0, 1, 2, 3, \dots\}$$



The Poisson Distribution

$$f\left(y \mid \lambda\right) = \frac{e^{-\lambda}\lambda^{y}}{y!} \;\; ; \;\; y \in \{0, 1, 2, 3, \ldots\}$$



Poisson Distribution: Log-Likelihood Function

$$L\left(\lambda \mid \mathbf{y}\right) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

Poisson Distribution: Log-Likelihood Function

$$L(\lambda \mid \mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}$$

$$\log(L(\lambda \mid \mathbf{y})) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda} \lambda^{y_i}}{y_i!}\right)$$

Poisson Distribution: Log-Likelihood Function

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

Poisson Distribution: Log-Likelihood Function

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

Poisson Distribution: Log-Likelihood Function

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

Note: $\sum_{i=1}^{n} \log(y_i!)$ does not depend on λ

Poisson Distribution: Log-Likelihood Function

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

Note: $\sum_{i=1}^{n} \log(y_i!)$ does not depend on λ

$$
\frac{\partial \log(L(\lambda \mid \mathbf{y}))}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} y_i}{\lambda}
$$

Poisson Distribution: Log-Likelihood Function

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

Note: $\sum_{i=1}^{n} \log\left(y_i!\right)$ does not depend on λ

$$
\begin{aligned}
\frac{\partial \log\left(L\left(\lambda \mid \mathbf{y}\right)\right)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^{n} y_i}{\lambda} \\
\frac{\partial^2 \log\left(L\left(\lambda \mid \mathbf{y}\right)\right)}{\partial \lambda^2} &= -\frac{\sum_{i=1}^{n} y_i}{\lambda^2}
\end{aligned}
$$

The variance of the MLE estimator

The estimator of the MLE variance is:

$$\widehat{Var}\left(\widehat{\lambda}\right) = \left\{-\frac{\partial^2 \log\left(L\left(\lambda \mid \mathbf{y}\right)\right)}{\partial \lambda^2}\right\}^{-1}$$
$$= \left\{\frac{\sum_{i=1}^{n} y_i}{\lambda^2}\right\}^{-1}$$

The Logit Model:

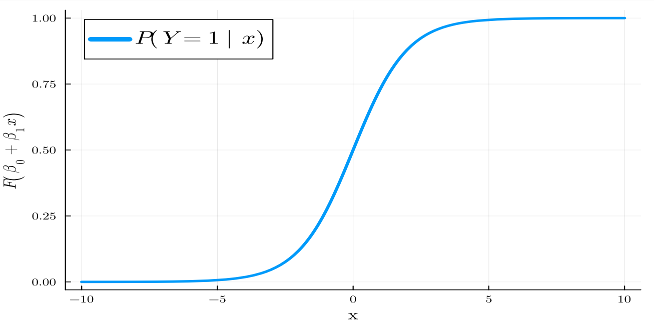
We want to model the probability of a binary outcome, $Y \in \{0, 1\}$

- Assume that $P(Y = y_i \mid x)$ depends on a single predictor, x.
- Thus, $P(Y = y_i \mid x_i) = F(\beta_0 + \beta_1 x)$

The Logit Model:

We want to model the probability of a binary outcome, $Y \in \{0, 1\}$

- Assume that $P(Y = y_i \mid x)$ depends on a single predictor, x.
- Thus, $P(Y = y_i \mid x_i) = F(\beta_0 + \beta_1 x)$



The Logit Model:

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

The Logit Model:

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

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

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:

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

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

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:

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