Maximum Likelihood Estimation

Philip Leifeld

GV903: Advanced Research Methods, Week 11

University of Essex

1. Conceptual Introduction

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE)

- MLE is both an estimation strategy (like OLS, but more flexible) and a statistical theory of inference.

Maximum Likelihood Estimation (MLE)

- MLE is both an estimation strategy (like OLS, but more flexible) and a statistical theory of inference.
- Estimation: it produces the slopes (more generally, parameters) for a regression model.

Maximum Likelihood Estimation (MLE)

- MLE is both an estimation strategy (like OLS, but more flexible) and a statistical theory of inference.
- Estimation: it produces the slopes (more generally, parameters) for a regression model.
- Theory of inference: principled way to think about the data-generating process.

Maximum Likelihood Estimation (MLE)

- MLE is both an estimation strategy (like OLS, but more flexible) and a statistical theory of inference.
- Estimation: it produces the slopes (more generally, parameters) for a regression model.
- Theory of inference: principled way to think about the data-generating process.
- The linear model can be re-interpreted through MLE.

Maximum Likelihood Estimation (MLE)

- MLE is both an estimation strategy (like OLS, but more flexible) and a statistical theory of inference.
- Estimation: it produces the slopes (more generally, parameters) for a regression model.
- Theory of inference: principled way to think about the data-generating process.
- The linear model can be re-interpreted through MLE.
- We can come up with linear models for other outcome distributions, like binary or Poisson-distributed data and plug them in.

Maximum Likelihood Estimation (MLE)

- MLE is both an estimation strategy (like OLS, but more flexible) and a statistical theory of inference.
- Estimation: it produces the slopes (more generally, parameters) for a regression model.
- Theory of inference: principled way to think about the data-generating process.
- The linear model can be re-interpreted through MLE.
- We can come up with linear models for other outcome distributions, like binary or Poisson-distributed data and plug them in.
- In fact, we can use *any* distribution as an underlying population process.

Probabilities: Example of a Die

- A die has six possible outcomes, the number 1–6.

Probabilities: Example of a Die

- A die has six possible outcomes, the number 1–6.
- This is the *sample space* of the DGP.

Probabilities: Example of a Die

- A die has six possible outcomes, the number 1–6.
- This is the *sample space* of the DGP.
- We can come up with a probability for each element in the sample space.

Probabilities: Example of a Die

- A die has six possible outcomes, the number 1–6.
- This is the *sample space* of the DGP.
- We can come up with a probability for each element in the sample space.
- In this case, $P(Y = y) = \frac{1}{6}$.

Probabilities: Example of a Die

- A die has six possible outcomes, the number 1–6.
- This is the *sample space* of the DGP.
- We can come up with a probability for each element in the sample space.
- In this case, $P(Y = y) = \frac{1}{6}$.
- We can also compute the probability to retrieve a certain set of outcomes if we repeat the random experiment a certain number of times.

Probabilities: Example of a Die

- A die has six possible outcomes, the number 1–6.
- This is the *sample space* of the DGP.
- We can come up with a probability for each element in the sample space.
- In this case, $P(Y = y) = \frac{1}{6}$.
- We can also compute the probability to retrieve a certain set of outcomes if we repeat the random experiment a certain number of times.
- Easy: take the product of all individual probabilities because the probabilities are independent from each other.

Probabilities: Example of a Die

- A die has six possible outcomes, the number 1–6.
- This is the *sample space* of the DGP.
- We can come up with a probability for each element in the sample space.
- In this case, $P(Y = y) = \frac{1}{6}$.
- We can also compute the probability to retrieve a certain set of outcomes if we repeat the random experiment a certain number of times.
- Easy: take the product of all individual probabilities because the probabilities are independent from each other.
- We can leverage this idea for forming a likelihood function, too.

Probability versus Likelihood

- Probability: what is the chance of observing the outcome (e. g., any specific number) given that we know the parameter or the way this distribution works.

Probability versus Likelihood

- Probability: what is the chance of observing the outcome (e. g., any specific number) given that we know the parameter or the way this distribution works.
- $P(Y|\theta)$.

Probability versus Likelihood

- Probability: what is the chance of observing the outcome (e. g., any specific number) given that we know the parameter or the way this distribution works.
- $P(Y|\theta)$.
- Likelihood: what is the parameter given that we know the observed data?

Probability versus Likelihood

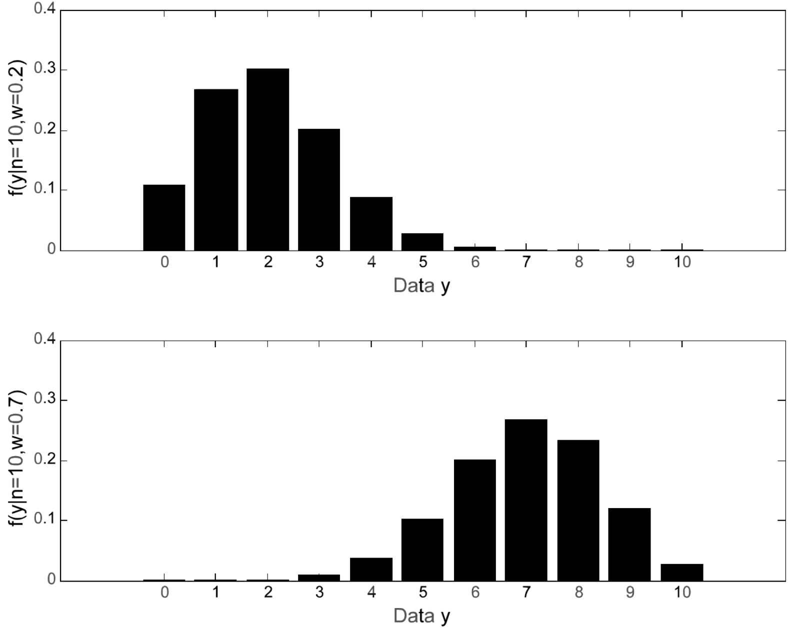
- Probability: what is the chance of observing the outcome (e. g., any specific number) given that we know the parameter or the way this distribution works.
- $P(Y|\theta)$.
- Likelihood: what is the parameter given that we know the observed data?
- E. g., given that we observe outcomes 1, 6, 4, 2, 2, 1, 5, 2, 3, what do we think is the data-generating process, or parameter?
- $\mathcal{L}(\theta|Y)$.

Probability versus Likelihood

- Probability: what is the chance of observing the outcome (e. g., any specific number) given that we know the parameter or the way this distribution works.
- $P(Y|\theta)$.
- Likelihood: what is the parameter given that we know the observed data?
- E. g., given that we observe outcomes 1, 6, 4, 2, 2, 1, 5, 2, 3, what do we think is the data-generating process, or parameter?
- $\mathcal{L}(\theta|Y)$.
- The two are in principle identical, except you ask the other way around. What is the model, given the observed data? Rather than what are the expected data, given a known model?

2. Example: Binomial Distribution

Binomial Sampling, $n = 10$ and $w = 0.2$ or $w = 0.7$



Binomial Probability Function

Multiplication of probabilities:

$$f(y = (y_1, y_2, \ldots, y_n) \mid w) = f_1(y_1 \mid w)\, f_2(y_2 \mid w) \cdots f_n(y_m \mid w).$$

Binomial Probability Function

Multiplication of probabilities:

$$f(y = (y_1, y_2, \ldots, y_n) \mid w) = f_1(y_1 \mid w) f_2(y_2 \mid w) \cdots f_n(y_m \mid w).$$

Binomial probability function:

$$f(y \mid n = 10, w = 0.2) = \frac{10!}{y!(10-y)!}(0.2)^y(0.8)^{10-y} \quad (y = 0, 1, \ldots, 10)$$

Binomial Probability Function

Multiplication of probabilities:

$$f(y = (y_1, y_2, \ldots, y_n) \mid w) = f_1(y_1 \mid w) f_2(y_2 \mid w) \cdots f_n(y_m \mid w).$$

Binomial probability function:

$$f(y \mid n = 10, w = 0.2) = \frac{10!}{y!(10-y)!}(0.2)^y(0.8)^{10-y} \quad (y = 0, 1, \ldots, 10)$$

Binomial probability function, more generally:

$$f(y|n, w) = \frac{n!}{y!(n-y)!}w^y(1-w)^{n-y} \quad (0 \leqslant w \leqslant 1; \;\; y = 0, 1, \ldots, n)$$

Binomial Probability Function

Multiplication of probabilities:

$$f(y = (y_1, y_2, \dots, y_n) \mid w) = f_1(y_1 \mid w)\, f_2(y_2 \mid w) \cdots f_n(y_m \mid w).$$

Binomial probability function:

$$f(y \mid n = 10, w = 0.2) = \frac{10!}{y!(10-y)!}(0.2)^y (0.8)^{10-y} \quad (y = 0, 1, \dots, 10)$$

Binomial probability function, more generally:

$$f(y|n, w) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y} \quad (0 \leqslant w \leqslant 1;\ \ y = 0, 1, \dots, n)$$

I. e., we insert the probability parameter first and then get the expected observations.

Binomial Likelihood Function

Likelihood function:

$L(w|y) = f(y|w).$

Binomial Likelihood Function

Likelihood function:

$L(w|y) = f(y|w)$.

Binomial likelihood function for $w = 0.7$:

$$L(w \mid n = 10, y = 7) = f(y = 7 \mid n = 10, w)$$
$$= \frac{10!}{7!3!} w^7 (1 - w)^3 \quad (0 \leqslant w \leqslant 1).$$

Binomial Likelihood Function

Likelihood function:

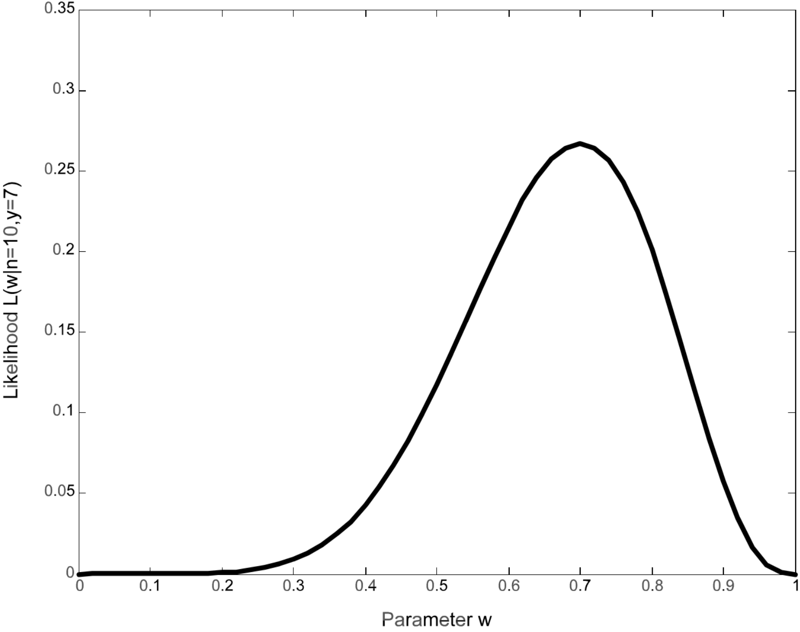
$L(w|y) = f(y|w)$.

Binomial likelihood function for $w = 0.7$:

$$L(w \mid n = 10, y = 7) = f(y = 7 \mid n = 10, w)$$
$$= \frac{10!}{7!3!} w^7 (1 - w)^3 \quad (0 \leqslant w \leqslant 1).$$

I. e., now we insert the observations first and then solve for the parameter w.

The Likelihood Function for $w = 0.7$



3. MLE: The General Procedure

MLE: Four-Step Procedure

MLE consists of four steps:

MLE: Four-Step Procedure

MLE consists of four steps:

1. Formulate the correct probability density or mass function. In this example, the PMF for the binomial distribution.

MLE: Four-Step Procedure

MLE consists of four steps:

1. Formulate the correct probability density or mass function. In this example, the PMF for the binomial distribution.
2. Write probability as likelihood. Take the logarithm and insert the numbers for the observations.

MLE: Four-Step Procedure

MLE consists of four steps:

1. Formulate the correct probability density or mass function. In this example, the PMF for the binomial distribution.
2. Write probability as likelihood. Take the logarithm and insert the numbers for the observations.
3. Take the first derivative and set zero to find the maximum:

$$\frac{\partial \ln L(w|y)}{\partial w_i} = 0$$

MLE: Four-Step Procedure

MLE consists of four steps:

1. Formulate the correct probability density or mass function. In this example, the PMF for the binomial distribution.
2. Write probability as likelihood. Take the logarithm and insert the numbers for the observations.
3. Take the first derivative and set zero to find the maximum:

$$\frac{\partial \ln L(w|y)}{\partial w_i} = 0$$

4. Take the second derivative to verify this is a maximum, not a minimum:

$$\frac{\partial^2 \ln L(w|y)}{\partial w_i^2} < 0.$$

Example of the Four Steps: Binomial Likelihood

Example of the Four Steps: Binomial Likelihood

Step 1: formulate the correct PMF:

$$f(y|n,w) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y}$$

$$(0 \leqslant w \leqslant 1; \quad y = 0, 1, \ldots, n)$$

Example of the Four Steps: Binomial Likelihood

Step 1: formulate the correct PMF:

$$f(y|n,w) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y}$$
$$(0 \leqslant w \leqslant 1; \quad y = 0, 1, \ldots, n)$$

Step 2: Take logarithm and insert numbers:

$$\ln L(w \mid n = 10, y = 7) = \ln \frac{10!}{7!3!} + 7 \ln w + 3 \ln(1 - w)$$

Example of the Four Steps: Binomial Likelihood

Step 1: formulate the correct PMF:

$$f(y|n,w) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y}$$

$$(0 \leqslant w \leqslant 1; \quad y = 0, 1, \ldots, n)$$

Step 2: Take logarithm and insert numbers:

$$\ln L(w \mid n = 10, y = 7) = \ln \frac{10!}{7!3!} + 7 \ln w + 3 \ln(1 - w)$$

Step 3: First derivative:

$$\frac{d \ln L(w \mid n = 10, y = 7)}{dw} = \frac{7}{w} - \frac{3}{1 - w} = \frac{7 - 10w}{w(1 - w)}$$

Example of the Four Steps: Binomial Likelihood

Step 1: formulate the correct PMF:

$$f(y|n,w) = \frac{n!}{y!(n-y)!} w^y (1-w)^{n-y}$$

$$(0 \leqslant w \leqslant 1; \quad y = 0, 1, \ldots, n)$$

Step 2: Take logarithm and insert numbers:

$$\ln L(w \mid n = 10, y = 7) = \ln \frac{10!}{7!3!} + 7 \ln w + 3 \ln(1-w)$$

Step 3: First derivative:

$$\frac{d \ln L(w \mid n = 10, y = 7)}{dw} = \frac{7}{w} - \frac{3}{1-w} = \frac{7 - 10w}{w(1-w)}$$

Step 4: Second derivative:

$$\frac{d^2 \ln L(w \mid n = 10, y = 7)}{dw^2} = -\frac{7}{w^2} - \frac{3}{(1-w)^2}$$

$$= -47.62 < 0$$

Why the Logarithm?

Why the Logarithm?

- In Step 2, we take the log.

Why the Logarithm?

- In Step 2, we take the log.
- We do this because the log likelihood is a lot easier to deal with for any computations.

Why the Logarithm?

- In Step 2, we take the log.
- We do this because the log likelihood is a lot easier to deal with for any computations.
- In particular (see Wooldridge, Appendix A):

Why the Logarithm?

- In Step 2, we take the log.
- We do this because the log likelihood is a lot easier to deal with for any computations.
- In particular (see Wooldridge, Appendix A):

$$\begin{aligned}
\log(x_1 \cdot x_2) &= \log(x_1) + \log(x_2),\ x_1, x_2 > 0 \\
\log(x_1/x_2) &= \log(x_1) - \log(x_2),\ x_1, x_2 > 0 \\
\log(x^c) &= c\log(x),\ x > 0,\ c \text{ any number.}
\end{aligned}$$

Why the Logarithm?

- In Step 2, we take the log.
- We do this because the log likelihood is a lot easier to deal with for any computations.
- In particular (see Wooldridge, Appendix A):

$$\begin{aligned} \log(x_1 \cdot x_2) &= \log(x_1) + \log(x_2), x_1, x_2 > 0 \\ \log(x_1/x_2) &= \log(x_1) - \log(x_2), x_1, x_2 > 0 \\ \log(x^c) &= c \log(x), x > 0, c \text{ any number.} \end{aligned}$$

- This comes in handy for taking derivatives in Steps 3 and 4.

Why the Logarithm?

- In Step 2, we take the log.
- We do this because the log likelihood is a lot easier to deal with for any computations.
- In particular (see Wooldridge, Appendix A):

$$\begin{aligned}
\log(x_1 \cdot x_2) &= \log(x_1) + \log(x_2),\ x_1, x_2 > 0 \\
\log(x_1/x_2) &= \log(x_1) - \log(x_2),\ x_1, x_2 > 0 \\
\log(x^c) &= c\log(x),\ x > 0,\ c \text{ any number.}
\end{aligned}$$

- This comes in handy for taking derivatives in Steps 3 and 4.
- It is a valid transformation and has the maximum in the same spot because it is a monotonically increasing function.

Why the Logarithm?

- In Step 2, we take the log.
- We do this because the log likelihood is a lot easier to deal with for any computations.
- In particular (see Wooldridge, Appendix A):

$$\begin{aligned}
\log(x_1 \cdot x_2) &= \log(x_1) + \log(x_2), x_1, x_2 > 0 \\
\log(x_1/x_2) &= \log(x_1) - \log(x_2), x_1, x_2 > 0 \\
\log(x^c) &= c \log(x), x > 0, c \text{ any number.}
\end{aligned}$$

- This comes in handy for taking derivatives in Steps 3 and 4.
- It is a valid transformation and has the maximum in the same spot because it is a monotonically increasing function.

In most real-world applications, Steps 3 and 4 cannot be done analytically.

Why the Logarithm?

- In Step 2, we take the log.
- We do this because the log likelihood is a lot easier to deal with for any computations.
- In particular (see Wooldridge, Appendix A):

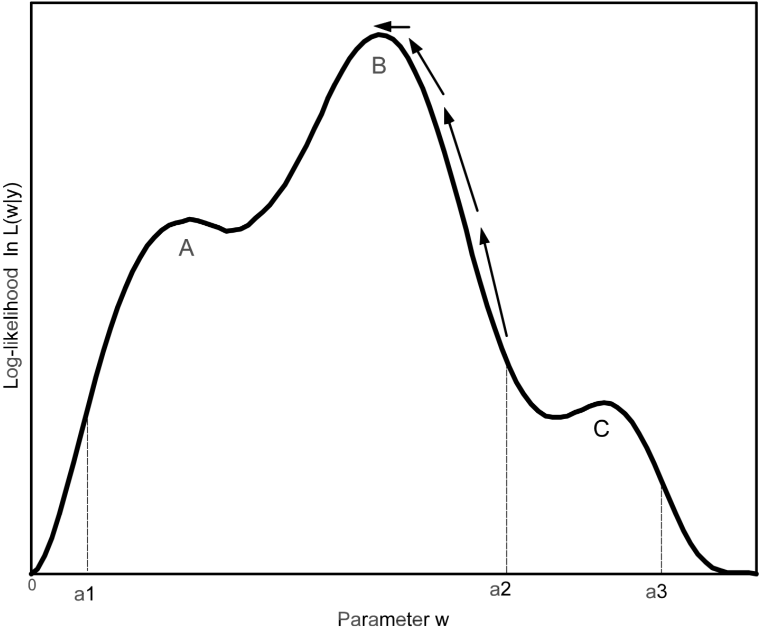
$$\begin{aligned}
\log(x_1 \cdot x_2) &= \log(x_1) + \log(x_2), x_1, x_2 > 0 \\
\log(x_1/x_2) &= \log(x_1) - \log(x_2), x_1, x_2 > 0 \\
\log(x^c) &= c \log(x), x > 0, c \text{ any number.}
\end{aligned}$$

- This comes in handy for taking derivatives in Steps 3 and 4.
- It is a valid transformation and has the maximum in the same spot because it is a monotonically increasing function.

In most real-world applications, Steps 3 and 4 cannot be done analytically.

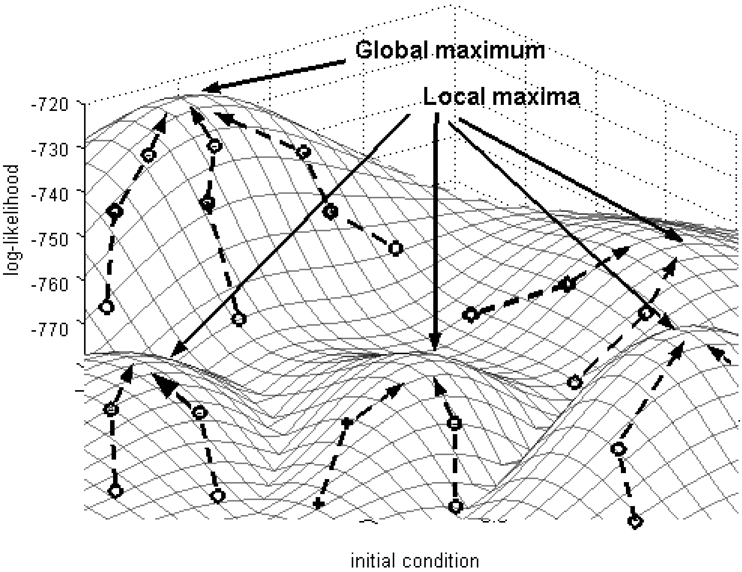
An optimization algorithm needs to find the maximum, for example Newton-Raphson or BFGS.

Iterative Optimization to Find the MLE

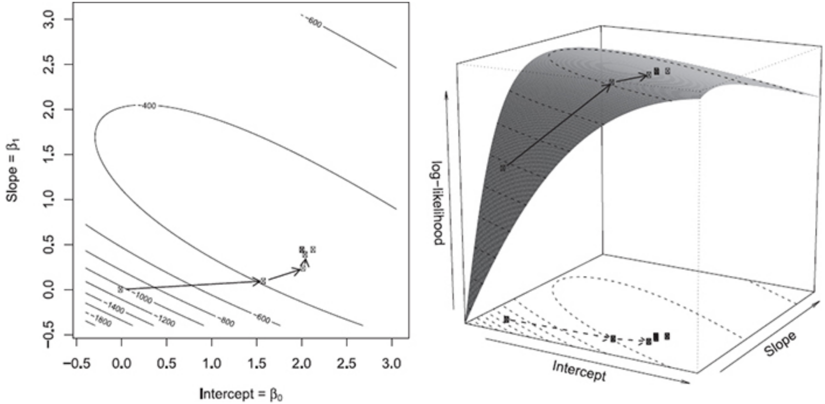


Multidimensional Likelihood and Local Maxima

Source: Akbar et al 2019: Statistical Analysis of Wireless Systems Using Markov Models



Newton-Raphson Optimization for Linear Models



4. Example: Sample Mean

You should know the PDF of the normal distribution (here with $\sigma = 1$ fixed):

$$f(y_i \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y_i - \mu)^2}{2} \right)$$

MLE of the Sample Mean

Following the exposition in Scott Long (1997)

You should know the PDF of the normal distribution (here with $\sigma = 1$ fixed):

$$f(y_i \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y_i - \mu)^2}{2} \right)$$

Let's swap the parameters around to convert this into a likelihood function:

$$L(\mu \mid y_i,\ \sigma = 1) = f(y_i \mid \mu,\ \sigma = 1)$$

MLE of the Sample Mean

Following the exposition in Scott Long (1997)

You should know the PDF of the normal distribution (here with $\sigma = 1$ fixed):

$$f(y_i \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y_i - \mu)^2}{2} \right)$$

Let's swap the parameters around to convert this into a likelihood function:

$$L(\mu \mid y_i, \ \sigma = 1) = f(y_i \mid \mu, \ \sigma = 1)$$

Let's do this for an example with three observed values. Then we have to take the product of the likelihood function for the three values:

MLE of the Sample Mean

Following the exposition in Scott Long (1997)

You should know the PDF of the normal distribution (here with $\sigma = 1$ fixed):

$$f(y_i \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y_i - \mu)^2}{2}\right)$$

Let's swap the parameters around to convert this into a likelihood function:

$$L(\mu \mid y_i,\ \sigma = 1) = f(y_i \mid \mu,\ \sigma = 1)$$

Let's do this for an example with three observed values. Then we have to take the product of the likelihood function for the three values:

$$L(\mu \mid \mathbf{y},\ \sigma = 1) = \prod_{i=1}^{3} L(\mu \mid y_i,\ \sigma = 1) = \prod_{i=1}^{3} f(y_i \mid \mu,\ \sigma = 1)$$

Next, take the log because that's easier to compute:

$$\ln L(\mu \mid \mathbf{y},\ \sigma = 1) = \sum_{i=1}^{3} \ln L(\mu \mid y_i,\ \sigma = 1) = \sum_{i=1}^{3} \ln f(y_i \mid \mu,\ \sigma = 1)$$

MLE of the Sample Mean

Following the exposition in Scott Long (1997)

Next, take the log because that's easier to compute:

$$\ln L(\mu \mid \mathbf{y},\ \sigma = 1) = \sum_{i=1}^{3} \ln L(\mu \mid y_i,\ \sigma = 1) = \sum_{i=1}^{3} \ln f(y_i \mid \mu,\ \sigma = 1)$$

The ML estimate is the value $\hat{\mu}$ that maximizes this equation.

MLE of the Sample Mean

Following the exposition in Scott Long (1997)

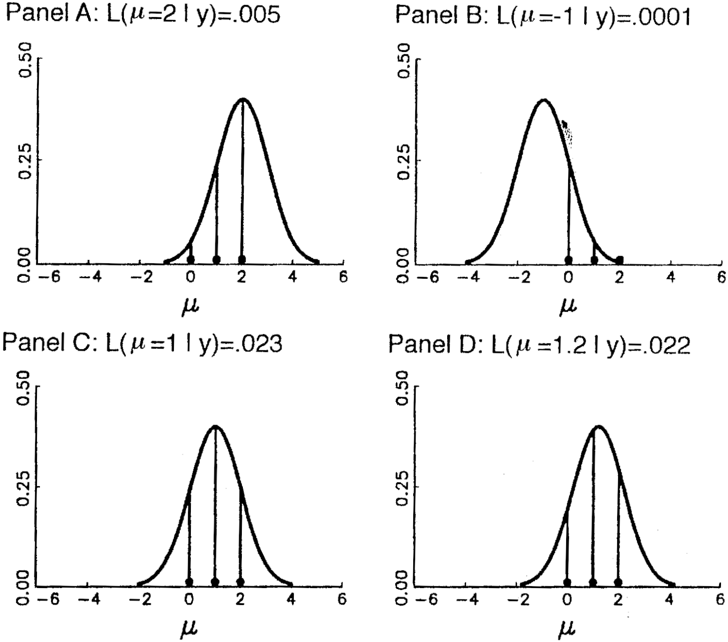
Next, take the log because that's easier to compute:

$$\ln L(\mu \mid \mathbf{y},\ \sigma = 1) = \sum_{i=1}^{3} \ln L(\mu \mid y_i,\ \sigma = 1) = \sum_{i=1}^{3} \ln f(y_i \mid \mu,\ \sigma = 1)$$

The ML estimate is the value $\hat{\mu}$ that maximizes this equation.

Let's say the three observed values are 0, 1, and 2. We can take a few "guesses" of μ and evaluate the likelihood...

Four Guesses of μ: $\mu = 1$ Maximizes the Likelihood



5. Re-Interpreting the Linear Model

MLE for the Linear Model

Now let's do this with the linear model. Plug the SSE into the normal distribution (because we want normally distributed errors):

MLE for the Linear Model

Following the exposition in Scott Long (1997)

Now let's do this with the linear model. Plug the SSE into the normal distribution (because we want normally distributed errors):

$$f(y_i \mid \alpha + \beta x_i, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{[y_i - (\alpha + \beta x_i)]^2}{\sigma^2}\right)$$

MLE for the Linear Model

Following the exposition in Scott Long (1997)

Now let's do this with the linear model. Plug the SSE into the normal distribution (because we want normally distributed errors):

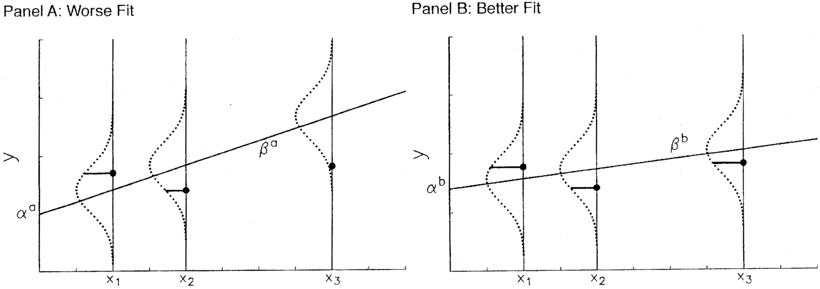
$$f(y_i \mid \alpha + \beta x_i, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{[y_i - (\alpha + \beta x_i)]^2}{\sigma^2} \right)$$

The principle is the same: We vary the α and β parameters given some observed data points to find the maximum likelihood:

Now let's do this with the linear model. Plug the SSE into the normal distribution (because we want normally distributed errors):

$$f(y_i \mid \alpha + \beta x_i, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{[y_i - (\alpha + \beta x_i)]^2}{\sigma^2}\right)$$

The principle is the same: We vary the α and β parameters given some observed data points to find the maximum likelihood:



MLE for the Linear Model

To make computations easier, we can use the standard normal distribution (z transformation) instead of the normal distribution:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

MLE for the Linear Model

To make computations easier, we can use the standard normal distribution (z transformation) instead of the normal distribution:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

This yields the following PDF for the linear model for a single observation:

$$f(y_i \mid \alpha + \beta x_i, \sigma) = \frac{1}{\sigma}\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)^2}{2}\right)\right]$$

$$= \frac{1}{\sigma}\,\phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)$$

MLE for the Linear Model

To make computations easier, we can use the standard normal distribution (z transformation) instead of the normal distribution:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

This yields the following PDF for the linear model for a single observation:

$$f(y_i \mid \alpha + \beta x_i, \sigma) = \frac{1}{\sigma}\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)^2}{2}\right)\right]$$

$$= \frac{1}{\sigma}\,\phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)$$

The corresponding likelihood function (i. e., for all data):

$$L(\alpha, \beta, \sigma \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \frac{1}{\sigma}\,\phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)$$

MLE for the Linear Model

Take the log (note how this replaces product by sum operator):

$$\ln L(\alpha, \beta, \sigma \,|\, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma}\, \phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)$$

MLE for the Linear Model

Take the log (note how this replaces product by sum operator):

$$\ln L(\alpha, \beta, \sigma \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma} \phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma}\right)$$

This extends seamlessly to the multiple regression case with more IVs (see Ward/Ahlquist p. 13 for additional simplification):

$$\ln L(\boldsymbol{\beta}, \sigma \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma} \phi\left(\frac{y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma}\right)$$

MLE for the Linear Model

Take the log (note how this replaces product by sum operator):

$$\ln L(\alpha, \beta, \sigma \,|\, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma} \phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma} \right)$$

This extends seamlessly to the multiple regression case with more IVs (see Ward/Ahlquist p. 13 for additional simplification):

$$\ln L(\boldsymbol{\beta}, \sigma \,|\, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma} \phi\left(\frac{y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma} \right)$$

We can now compute the first and second derivative to maximize this likelihood. This will yield $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$.

MLE for the Linear Model

Take the log (note how this replaces product by sum operator):

$$\ln L(\alpha, \beta, \sigma \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma} \phi\left(\frac{y_i - [\alpha + \beta x_i]}{\sigma} \right)$$

This extends seamlessly to the multiple regression case with more IVs (see Ward/Ahlquist p. 13 for additional simplification):

$$\ln L(\boldsymbol{\beta}, \sigma \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln \frac{1}{\sigma} \phi\left(\frac{y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma} \right)$$

We can now compute the first and second derivative to maximize this likelihood. This will yield $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$.

We can do this analytically for the linear model. (In most cases, we would need an optimization algorithm.)

The Hessian Matrix

$\boldsymbol{\theta}$ is a vector of all estimates ($\boldsymbol{\beta}$ and σ in this case).

The Hessian Matrix

$\boldsymbol{\theta}$ is a vector of all estimates ($\boldsymbol{\beta}$ and σ in this case).

Then the *Hessian matrix* is a matrix of second derivatives:

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}$$

The Hessian Matrix

$\boldsymbol{\theta}$ is a vector of all estimates ($\boldsymbol{\beta}$ and σ in this case).

Then the *Hessian matrix* is a matrix of second derivatives:

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}$$

This is a square, symmetric matrix:

$$\mathbf{H}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\alpha} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\beta} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\sigma} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\alpha} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\beta} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\sigma} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\alpha} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\beta} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\sigma} \end{pmatrix}$$

The Hessian Matrix

$\boldsymbol{\theta}$ is a vector of all estimates ($\boldsymbol{\beta}$ and σ in this case).

Then the *Hessian matrix* is a matrix of second derivatives:

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}$$

This is a square, symmetric matrix:

$$\mathbf{H}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\alpha} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\beta} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\sigma} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\alpha} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\beta} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\sigma} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\alpha} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\beta} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\sigma} \end{pmatrix}$$

Think of this as the multivariate version of the second derivative for testing for a maximum of the log likelihood.

Fisher Information Matrix and VCOV

A small value for the diagonal β entry, for example, indicates that the likelihood is changing slowly when β changes, which means the maximum is hard to find with respect to $\hat{\beta}$ and the variance is large.

Fisher Information Matrix and VCOV

A small value for the diagonal β entry, for example, indicates that the likelihood is changing slowly when β changes, which means the maximum is hard to find with respect to $\hat{\beta}$ and the variance is large. We can leverage this to find SEs for the estimates!

Fisher Information Matrix and VCOV

A small value for the diagonal β entry, for example, indicates that the likelihood is changing slowly when β changes, which means the maximum is hard to find with respect to $\hat{\beta}$ and the variance is large. We can leverage this to find SEs for the estimates!

The *Fisher information matrix* $\mathbf{I}$ is the negative of the expected value of the Hessian. The inverse of $\mathbf{I}$ is the VCOV matrix of the MLE:

Fisher Information Matrix and VCOV

A small value for the diagonal β entry, for example, indicates that the likelihood is changing slowly when β changes, which means the maximum is hard to find with respect to $\hat{\beta}$ and the variance is large. We can leverage this to find SEs for the estimates!

The *Fisher information matrix* $\mathbf{I}$ is the negative of the expected value of the Hessian. The inverse of $\mathbf{I}$ is the VCOV matrix of the MLE:

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = -E[\mathbf{H}(\boldsymbol{\theta})]^{-1}$$

Fisher Information Matrix and VCOV

A small value for the diagonal β entry, for example, indicates that the likelihood is changing slowly when β changes, which means the maximum is hard to find with respect to $\hat{\beta}$ and the variance is large. We can leverage this to find SEs for the estimates!

The *Fisher information matrix* $\mathbf{I}$ is the negative of the expected value of the Hessian. The inverse of $\mathbf{I}$ is the VCOV matrix of the MLE:

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = -E[\mathbf{H}(\boldsymbol{\theta})]^{-1}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\alpha}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\beta}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\alpha\partial\sigma}\right) \\ -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\alpha}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\beta}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\beta\partial\sigma}\right) \\ -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\alpha}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\beta}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\sigma\partial\sigma}\right) \end{pmatrix}^{-1}$$

6. MLE in `R`

MLE in R (see Monogan 2015)

Binomial likelihood function with y heads, n trials, π probability:
$L(\pi|n, y) = \pi^{y}(1 - \pi)^{n-y}$

MLE in R (see Monogan 2015)

Binomial likelihood function with y heads, n trials, π probability:
$L(\pi|n, y) = \pi^y(1-\pi)^{n-y}$

Corresponding log likelihood:
$\log L(\pi|n, y) = y \cdot \log(\pi) + (n-y)\log(1-\pi)$

MLE in R (see Monogan 2015)

Binomial likelihood function with y heads, n trials, π probability:
$L(\pi|n, y) = \pi^y(1-\pi)^{n-y}$

Corresponding log likelihood:
$\log L(\pi|n, y) = y \cdot \log(\pi) + (n-y)\log(1-\pi)$

Write this as an R function:

```
binomial.loglikelihood <- function(prob, y, n) {
   loglikelihood <- y*log(prob) + (n-y)*log(1-prob)
   return(loglikelihood)
}
```

MLE in R (see Monogan 2015)

Binomial likelihood function with y heads, n trials, π probability:
$L(\pi|n, y) = \pi^y(1-\pi)^{n-y}$

Corresponding log likelihood:
$\log L(\pi|n, y) = y \cdot \log(\pi) + (n-y)\log(1-\pi)$

Write this as an R function:

```
binomial.loglikelihood <- function(prob, y, n) {
   loglikelihood <- y*log(prob) + (n-y)*log(1-prob)
   return(loglikelihood)
}
```

Use the `optim` function to optimize the likelihood given fixed `y`, `n`:

```
test <- optim(c(.5),                # starting value for prob
   binomial.loglikelihood,          # the log-likelihood function
   method="BFGS",                   # optimization method
   hessian=TRUE,                    # return numerical Hessian
   control=list(fnscale=-1),        # maximize instead of minimize
   y=43, n=100)                     # the data
print(test)
```

MLE in R

This yields a list object with several slots:

```
$par
[1] 0.4300015

$value
[1] -68.33149

$counts
function gradient
      13        4

$convergence
[1] 0

$message
NULL

$hessian
          [,1]
[1,] -407.9996
```

MLE in R

This yields a list object with several slots:

```
$par
[1] 0.4300015

$value
[1] -68.33149

$counts
function gradient
      13        4

$convergence
[1] 0

$message
NULL

$hessian
          [,1]
[1,] -407.9996
```

par lists the estimated π parameter.

MLE in R

This yields a list object with several slots:

```
$par
[1] 0.4300015

$value
[1] -68.33149

$counts
function gradient
      13        4

$convergence
[1] 0

$message
NULL

$hessian
          [,1]
[1,] -407.9996
```

`value` shows the log-likelihood of our final solution.

MLE in R

This yields a list object with several slots:

```
$par
[1] 0.4300015

$value
[1] -68.33149

$counts
function gradient
      13        4

$convergence
[1] 0

$message
NULL

$hessian
          [,1]
[1,] -407.9996
```

`counts` shows how often `optim` called the function and gradient.

MLE in R

This yields a list object with several slots:

```
$par
[1] 0.4300015

$value
[1] -68.33149

$counts
function gradient
      13        4

$convergence
[1] 0

$message
NULL

$hessian
          [,1]
[1,] -407.9996
```

`convergence`: 0 means successfully converged to maximum.

This yields a list object with several slots:

```
$par
[1] 0.4300015

$value
[1] -68.33149

$counts
function gradient
      13        4

$convergence
[1] 0

$message
NULL

$hessian
          [,1]
[1,] -407.9996
```

`message`: Messages from the function during optimization.

MLE in R

This yields a list object with several slots:

```
$par
[1] 0.4300015

$value
[1] -68.33149

$counts
function gradient
      13        4

$convergence
[1] 0

$message
NULL

$hessian
          [,1]
[1,] -407.9996
```

`hessian`: Hessian matrix (here: only one parameter).

MLE in R

We can now proceed to compute the standard error(s):

```
sqrt(diag(solve(-test$hessian)))
```

MLE in R

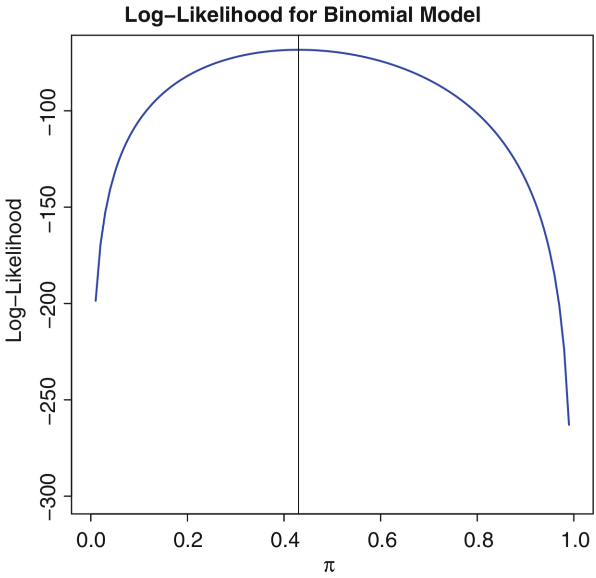
We can now proceed to compute the standard error(s):

```
sqrt(diag(solve(-test$hessian)))
```

Only one parameter. It's easy to plot the likelihood function:

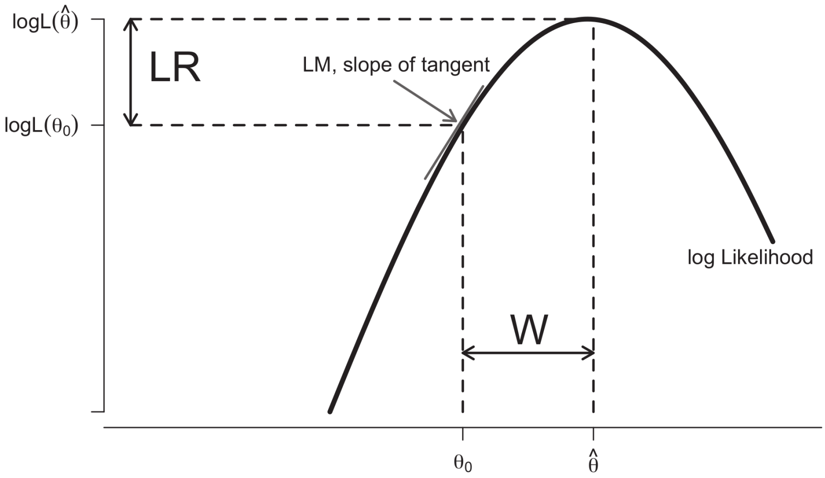
```
ruler <- seq(0,1,0.01)
loglikelihood <- binomial.loglikelihood(ruler, y=43, n=100)
plot(ruler, loglikelihood, type="l", lwd=2, col="blue",
   xlab=expression(pi),ylab="Log-Likelihood",ylim=c(-300,-70),
   main="Log-Likelihood for Binomial Model")
abline(v=.43)
```

MLE in R: The Likelihood Function



7. Diagnostics and Goodness of Fit

Likelihood Ratio Test, Wald Statistic, and LM Test



The Likelihood Ratio Test (LR Test)

This is the MLE analogue of the F test.

The Likelihood Ratio Test (LR Test)

This is the MLE analogue of the F test.

$$\mathrm{LR}(\boldsymbol{\theta}_R, \boldsymbol{\theta}_G \mid \mathbf{x}) = -2 \log \frac{\mathcal{L}(\boldsymbol{\theta}_R \mid \mathbf{x})}{\mathcal{L}(\boldsymbol{\theta}_G \mid \mathbf{x})}$$

The Likelihood Ratio Test (LR Test)

This is the MLE analogue of the F test.

$$\mathrm{LR}(\boldsymbol{\theta}_R, \boldsymbol{\theta}_G \mid \mathbf{x}) = -2 \log \frac{\mathcal{L}(\boldsymbol{\theta}_R \mid \mathbf{x})}{\mathcal{L}(\boldsymbol{\theta}_G \mid \mathbf{x})}$$

It follows a χ^2_{g-r} distribution with $g - r$ degrees of freedom.

The Likelihood Ratio Test (LR Test)

This is the MLE analogue of the F test.

$$\mathrm{LR}(\boldsymbol{\theta}_R, \boldsymbol{\theta}_G \mid \mathbf{x}) = -2 \log \frac{\mathcal{L}(\boldsymbol{\theta}_R \mid \mathbf{x})}{\mathcal{L}(\boldsymbol{\theta}_G \mid \mathbf{x})}$$

It follows a χ^2_{g-r} distribution with $g - r$ degrees of freedom.

A similar diagnostic is the *model deviance*, the log-likelihood difference between a model and a saturated model (with as many parameters as observations). It is sometimes used to compute the difference between residual deviance and null deviance.

AIC and BIC

The most common goodness-of-fit measures for MLE are AIC and BIC.

Akaike Information Criterion (AIC):

$$\mathsf{AIC} = -2 \log \hat{\mathcal{L}} + 2k$$

Bayesian Information Criterion (BIC):

$$\mathsf{BIC} = -2 \log \hat{\mathcal{L}} + k \log n$$

where k is the number of parameters and n is the number of observations.

Both measures penalise the inclusion of additional parameters.

Smaller values are "better". Only use for model comparison with the same data! Never interpret the absolute value of AIC or BIC!

Concluding Thoughts: Why Do We Need MLE?

Concluding Thoughts: Why Do We Need MLE?

- OLS is limited to normally distributed errors and quantitative DVs.

Concluding Thoughts: Why Do We Need MLE?

- OLS is limited to normally distributed errors and quantitative DVs.
- MLE is much more flexible.

Concluding Thoughts: Why Do We Need MLE?

- OLS is limited to normally distributed errors and quantitative DVs.
- MLE is much more flexible.
- Flexibility comes at a price: computational cost and complexity.

Concluding Thoughts: Why Do We Need MLE?

- OLS is limited to normally distributed errors and quantitative DVs.
- MLE is much more flexible.
- Flexibility comes at a price: computational cost and complexity.
- Proper theoretical foundation for inference.

Concluding Thoughts: Why Do We Need MLE?

- OLS is limited to normally distributed errors and quantitative DVs.
- MLE is much more flexible.
- Flexibility comes at a price: computational cost and complexity.
- Proper theoretical foundation for inference.
- MLE will allow us to look at models for various outcome distributions in the spring, such as binary, count, and ordinal DVs. This is called *generalized linear model* (GLM).

Concluding Thoughts: Why Do We Need MLE?

- OLS is limited to normally distributed errors and quantitative DVs.
- MLE is much more flexible.
- Flexibility comes at a price: computational cost and complexity.
- Proper theoretical foundation for inference.
- MLE will allow us to look at models for various outcome distributions in the spring, such as binary, count, and ordinal DVs. This is called *generalized linear model* (GLM).
- The MLE framework is also the starting point for Bayesian statistics (not covered in this module).