

# Lectures in Quantitative International Relations

## Introduction to Maximum Likelihood

Kevin A. Clarke  
University of Rochester

Dublin City University, May 2007

# Lectures in this Series

# Lectures in this Series

## 1 Introduction to Maximum Likelihood Estimation

# Lectures in this Series

- 1 Introduction to Maximum Likelihood Estimation
- 2 Common ML Models Used in International Relations

# Lectures in this Series

- 1 Introduction to Maximum Likelihood Estimation
- 2 Common ML Models Used in International Relations
- 3 Comparative Theory Testing

# Lectures in this Series

- 1 Introduction to Maximum Likelihood Estimation
- 2 Common ML Models Used in International Relations
- 3 Comparative Theory Testing
- 4 Choosing a Specification

# Course Web Site

<http://www.rochester.edu/college/psc/clarke/Dublin/Dublin.html>

## Course Web Site

<http://www.rochester.edu/college/psc/clarke/Dublin/Dublin.html>

Includes the following:

- syllabus

## Course Web Site

<http://www.rochester.edu/college/psc/clarke/Dublin/Dublin.html>

Includes the following:

- syllabus
- exercises

## Course Web Site

<http://www.rochester.edu/college/psc/clarke/Dublin/Dublin.html>

Includes the following:

- syllabus
- exercises
- data sets

## Course Web Site

<http://www.rochester.edu/college/psc/clarke/Dublin/Dublin.html>

Includes the following:

- syllabus
- exercises
- data sets
- codebooks

## Course Web Site

<http://www.rochester.edu/college/psc/clarke/Dublin/Dublin.html>

Includes the following:

- syllabus
- exercises
- data sets
- codebooks
- relevant articles

## Course Web Site

<http://www.rochester.edu/college/psc/clarke/Dublin/Dublin.html>

Includes the following:

- syllabus
- exercises
- data sets
- codebooks
- relevant articles
- software resources

## Why the Focus on Maximum Likelihood?

- Fearon, Kasara, Laitin (2007). “Ethnic Minority Rule and Civil War Onset.” *APSR*.
- Cederman and Girardin (2007). “Beyond Fractionalization: Mapping Ethnicity onto Nationalist Insurgencies.” *APSR*.
- Crescenzi (2007). “Reputation and Interstate Conflict.” *AJPS*.
- Danilovic and Clare (2007). “The Kantian Liberal Peace (Revisited)” *AJPS*.
- Lektzian and Sprecher. (2007) “Sanctions, Signals, and Militarized Conflict.” *AJPS*.
- Gartzke (2007). “The Capitalist Peace.” *AJPS*.

# What are our Goals?

What should you get out of these lectures?

# What are our Goals?

What should you get out of these lectures?

- intuition

# What are our Goals?

What should you get out of these lectures?

- intuition
- an idea of what's out there

# What are our Goals?

What should you get out of these lectures?

- intuition
- an idea of what's out there
- a basis for learning new techniques

# What are our Goals?

What should you get out of these lectures?

- intuition
- an idea of what's out there
- a basis for learning new techniques
- help in reading the literature

# Overview of Lecture 1

- 1 The Logic of Political Survival
  - The Model
  - Explaining the Bias

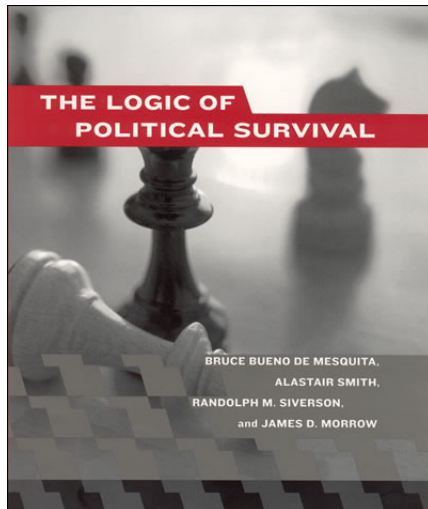
# Overview of Lecture 1

- 1 The Logic of Political Survival
  - The Model
  - Explaining the Bias
- 2 Maximum Likelihood Theory
  - Intuition
  - Fundamentals
  - MLE and Regression

# Overview of Lecture 1

- 1 The Logic of Political Survival
  - The Model
  - Explaining the Bias
- 2 Maximum Likelihood Theory
  - Intuition
  - Fundamentals
  - MLE and Regression
- 3 R

# The Logic of Political Survival



# The Selectorate Model

- Leaders want to survive.

# The Selectorate Model

- Leaders want to survive.
- Selectorate ( $S$ ) and minimum winning coalition ( $W$ ).

# The Selectorate Model

- Leaders want to survive.
- Selectorate ( $S$ ) and minimum winning coalition ( $W$ ).
- Public vs. private goods.

# The Selectorate Model

- Leaders want to survive.
- Selectorate ( $S$ ) and minimum winning coalition ( $W$ ).
- Public vs. private goods.
- Loyalty norm ( $W/S$ ).

# An Implication of the Model

Direct implication of the model:

- Institutions that call for larger winning coalitions should be associated with more effort to produce public goods.

Thus,  $W$  should be positively associated with public expenditures (measured as a proportion of GDP).

# Expenditures

Variable	Coef.	S.E.
W	2.32	(0.99)
S	3.10	(0.85)
ln(Population)	-1.90	(0.13)
Democracy residuals	2.83	(1.15)
ln(GDP) residuals	1.07	(0.20)
Constant	53.89	(2.19)

# Expenditures

Variable	Coef.	S.E.
W	2.32	(0.99)
S	3.10	(0.85)
ln(Population)	-1.90	(0.13)
Democracy residuals	2.83	(1.15)
ln(GDP) residuals	1.07	(0.20)
Constant	53.89	(2.19)

What are democracy residuals and ln(GDP) residuals?

Concerned about the correlation between  $W$  and *democracy* ( $r = 0.80$ ), BdM and coauthors regress *democracy* on  $W$  and substitute the residuals from this regression into their main regression.

Concerned about the correlation between  $W$  and *democracy* ( $r = 0.80$ ), BdM and coauthors regress *democracy* on  $W$  and substitute the residuals from this regression into their main regression.

This technique is known as “residualization,” and it has been used extensively in American politics in the ideology literature.

Concerned about the correlation between  $W$  and *democracy* ( $r = 0.80$ ), BdM and coauthors regress *democracy* on  $W$  and substitute the residuals from this regression into their main regression.

This technique is known as “residualization,” and it has been used extensively in American politics in the ideology literature.

What happens when we substitute the original variables back into the regression?

# Expenditures

Variable	Residuals		Controls	
	Coef.	S.E.	Coef.	S.E.
<b>W</b>	<b>2.32</b>	<b>(0.99)</b>	<b>-4.47</b>	<b>(1.74)</b>
S	3.10	(0.85)	4.35	(0.87)
ln(Population)	-1.90	(0.13)	-1.90	(0.13)
Democracy residuals	2.83	(1.15)		
ln(GDP) residuals	1.07	(0.20)		
Democracy			2.83	(1.15)
ln(GDP)			1.07	(0.20)
Constant	53.89	(2.19)	47.49	(2.69)

# The Problem

BdM and coauthors want to run:

$$\mathbf{y} = \beta_0 + \mathbf{w}\beta_1 + \mathbf{d}\beta_2 + \epsilon.$$

# The Problem

BdM and coauthors want to run:

$$\mathbf{y} = \beta_0 + \mathbf{w}\beta_1 + \mathbf{d}\beta_2 + \epsilon.$$

They actually run this instead:

$$\mathbf{y} = \beta_0 + \mathbf{w}\beta_1 + \hat{\nu}\beta_2 + \epsilon,$$

# The Problem

BdM and coauthors want to run:

$$\mathbf{y} = \beta_0 + \mathbf{w}\beta_1 + \mathbf{d}\beta_2 + \epsilon.$$

They actually run this instead:

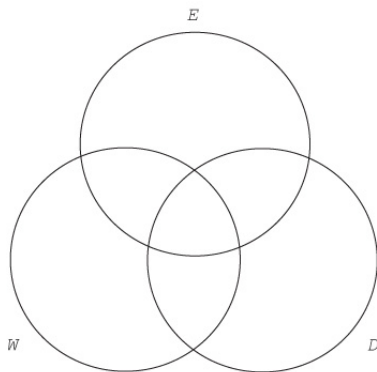
$$\mathbf{y} = \beta_0 + \mathbf{w}\beta_1 + \hat{\nu}\beta_2 + \epsilon,$$

where  $\hat{\nu}$  are the residuals from the auxiliary regression,

$$\mathbf{d} = \gamma_0 + \mathbf{w}\gamma_1 + \nu.$$

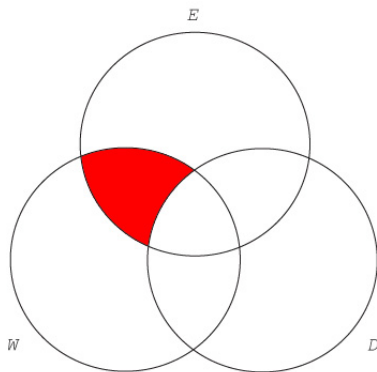
# Intuition

## Kennedy's Ballentine



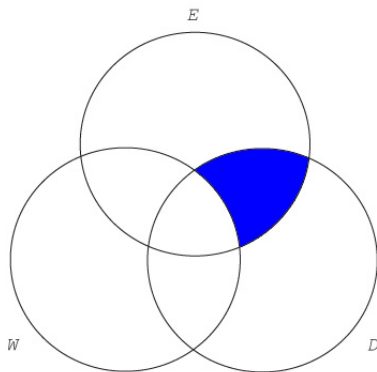
# Intuition

## Kennedy's Ballentine



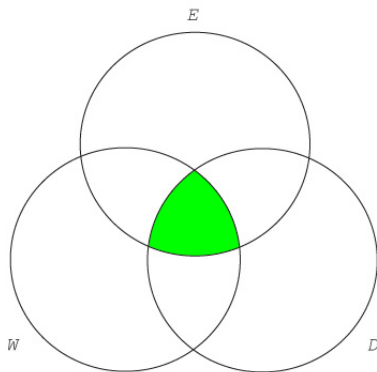
# Intuition

## Kennedy's Ballentine



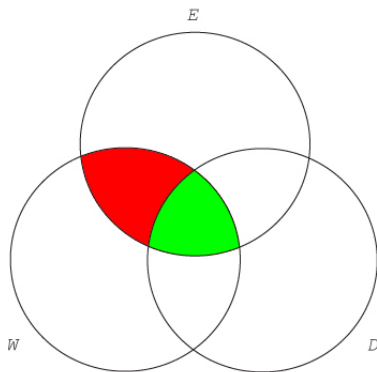
# Intuition

## Kennedy's Ballentine



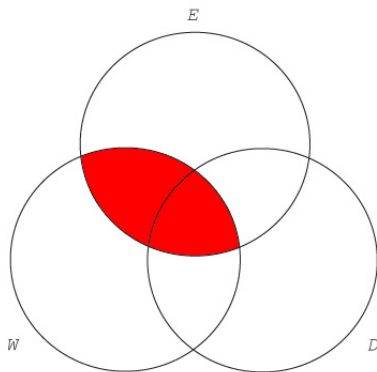
# Intuition

## Kennedy's Ballentine



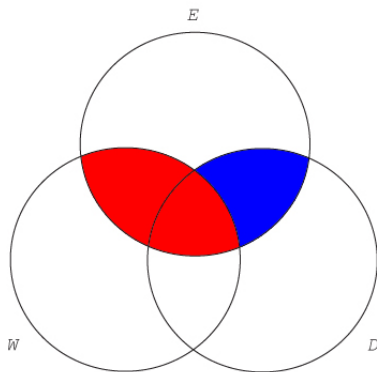
# Intuition

## Kennedy's Ballentine



# Intuition

## Kennedy's Ballentine



# The Estimated Coefficient on $W$

Given the regression

$$\mathbf{y} = \beta_0 + \mathbf{w}\beta_1 + \hat{\mathbf{v}}\beta_2 + \epsilon,$$

the expected value of this estimator, assuming  $E[\epsilon] = \mathbf{0}$ , is

$$E[\hat{\beta}_1] = \beta_1 + (\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\mathbf{d}\beta_2,$$

# This **Is** Omitted Variable Bias

Compare the expectation from the last frame

$$E[\hat{\beta}_1] = \beta_1 + (\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\mathbf{d}\beta_2$$

to the standard omitted variable bias result

$$E[\hat{\beta}_1] = \beta_1 + (\mathbf{x}'_1\mathbf{x}_1)^{-1}\mathbf{x}'_1\mathbf{x}_2\beta_2.$$

# This **Is** Omitted Variable Bias

Compare the expectation from the last frame

$$E[\hat{\beta}_1] = \beta_1 + (\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\mathbf{d}\beta_2$$

to the standard omitted variable bias result

$$E[\hat{\beta}_1] = \beta_1 + (\mathbf{x}'_1\mathbf{x}_1)^{-1}\mathbf{x}'_1\mathbf{x}_2\beta_2.$$

In most of their regressions, BdM and coauthors estimated the effect of **w** as if **d** were not in the equations.

# Implications

- The coefficient on  $\mathbf{w}$  is biased and inconsistent.

# Implications

- The coefficient on  $\mathbf{w}$  is biased and inconsistent.
- Whether the bias is large or not depends on the effect of  $\mathbf{d}$ .

# Implications

- The coefficient on  $\mathbf{w}$  is biased and inconsistent.
- Whether the bias is large or not depends on the effect of  $\mathbf{d}$ .
- The coefficient on the residuals is unbiased.

# Implications

- The coefficient on  $\mathbf{w}$  is biased and inconsistent.
- Whether the bias is large or not depends on the effect of  $\mathbf{d}$ .
- The coefficient on the residuals is unbiased.
- The standard error on  $\mathbf{w}$  is attenuated.

# Implications

- The coefficient on  $\mathbf{w}$  is biased and inconsistent.
- Whether the bias is large or not depends on the effect of  $\mathbf{d}$ .
- The coefficient on the residuals is unbiased.
- The standard error on  $\mathbf{w}$  is attenuated.
- When more than one set of residuals is included, the effect of the  $\mathbf{w}$  coefficient is a sum of the biases.

# Implications

- The coefficient on  $\mathbf{w}$  is biased and inconsistent.
- Whether the bias is large or not depends on the effect of  $\mathbf{d}$ .
- The coefficient on the residuals is unbiased.
- The standard error on  $\mathbf{w}$  is attenuated.
- When more than one set of residuals is included, the effect of the  $\mathbf{w}$  coefficient is a sum of the biases.
- Last, but not least...

# Implications

- The coefficient on  $\mathbf{w}$  is biased and inconsistent.
- Whether the bias is large or not depends on the effect of  $\mathbf{d}$ .
- The coefficient on the residuals is unbiased.
- The standard error on  $\mathbf{w}$  is attenuated.
- When more than one set of residuals is included, the effect of the  $\mathbf{w}$  coefficient is a sum of the biases.
- Last, but not least...
- Know what you are doing.

# Maximum Likelihood Theory

With that in mind, let's turn to learning maximum likelihood theory. Remember, if some of the math is unfamiliar to you, try and get the larger picture.

# Maximum Likelihood Theory

With that in mind, let's turn to learning maximum likelihood theory. Remember, if some of the math is unfamiliar to you, try and get the larger picture.

Let's start with some intuition.

# The Binomial Probability

Consider a sample of great power militarized disputes. If we know the probability of escalation, the binomial formula tells us the probability of, say, 3 disputes out of 10 escalating.

# The Binomial Probability

Consider a sample of great power militarized disputes. If we know the probability of escalation, the binomial formula tells us the probability of, say, 3 disputes out of 10 escalating.

If the probability of escalation is  $\theta = 0.3$ ,

# The Binomial Probability

Consider a sample of great power militarized disputes. If we know the probability of escalation, the binomial formula tells us the probability of, say, 3 disputes out of 10 escalating.

If the probability of escalation is  $\theta = 0.3$ ,

$$\begin{aligned}\Pr(x|\theta, N) &= \binom{N}{x} \theta^x (1 - \theta)^{N-x} \\ &= \binom{10}{3} 0.3^3 (1 - 0.3)^{10-3} \\ &= 0.27\end{aligned}$$

# The Binomial Probability

How would we go about estimating the probability of dispute escalation if we didn't know it?

# The Binomial Probability

How would we go about estimating the probability of dispute escalation if we didn't know it?

One possibility is choosing the value of  $\theta$  that gives us the largest probability.

# The Binomial Probability

How would we go about estimating the probability of dispute escalation if we didn't know it?

One possibility is choosing the value of  $\theta$  that gives us the largest probability.

- If  $\theta = 0.1$ ,  $\Pr(3|0.1, 10) = 0.057$

# The Binomial Probability

How would we go about estimating the probability of dispute escalation if we didn't know it?

One possibility is choosing the value of  $\theta$  that gives us the largest probability.

- If  $\theta = 0.1$ ,  $\Pr(3|0.1, 10) = 0.057$
- If  $\theta = 0.3$ ,  $\Pr(3|0.3, 10) = 0.27$

# The Binomial Probability

How would we go about estimating the probability of dispute escalation if we didn't know it?

One possibility is choosing the value of  $\theta$  that gives us the largest probability.

- If  $\theta = 0.1$ ,  $\Pr(3|0.1, 10) = 0.057$
- If  $\theta = 0.3$ ,  $\Pr(3|0.3, 10) = 0.27$
- If  $\theta = 0.5$ ,  $\Pr(3|0.5, 10) = 0.12$

# The Binomial Probability

How would we go about estimating the probability of dispute escalation if we didn't know it?

One possibility is choosing the value of  $\theta$  that gives us the largest probability.

- If  $\theta = 0.1$ ,  $\Pr(3|0.1, 10) = 0.057$
- If  $\theta = 0.3$ ,  $\Pr(3|0.3, 10) = 0.27$
- If  $\theta = 0.5$ ,  $\Pr(3|0.5, 10) = 0.12$
- If  $\theta = 0.7$ ,  $\Pr(3|0.7, 10) = 0.009$

# The Binomial Probability

How would we go about estimating the probability of dispute escalation if we didn't know it?

One possibility is choosing the value of  $\theta$  that gives us the largest probability.

- If  $\theta = 0.1$ ,  $\Pr(3|0.1, 10) = 0.057$
- If  $\theta = 0.3$ ,  $\Pr(3|0.3, 10) = 0.27$
- If  $\theta = 0.5$ ,  $\Pr(3|0.5, 10) = 0.12$
- If  $\theta = 0.7$ ,  $\Pr(3|0.7, 10) = 0.009$
- If  $\theta = 0.9$ ,  $\Pr(3|0.9, 10) \approx 0$

# The Maximum Likelihood Estimator

$\theta = 0.3$  is the value that makes observing 3 escalations out of 10 militarized disputes most likely.

# The Maximum Likelihood Estimator

$\theta = 0.3$  is the value that makes observing 3 escalations out of 10 militarized disputes most likely.

That makes 0.3 the *maximum likelihood estimate*.

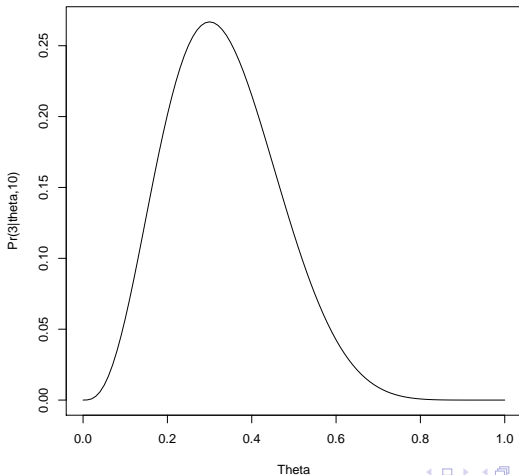
# The Maximum Likelihood Estimator

$\theta = 0.3$  is the value that makes observing 3 escalations out of 10 militarized disputes most likely.

That makes 0.3 the *maximum likelihood estimate*.

The ML estimate is the value of the parameter that makes the observed data most likely.

# The Likelihood Function



# Definitions and Notation

# Definitions and Notation

- **Random variable:**  $X$

# Definitions and Notation

- **Random variable:**  $X$
- **Realization:**  $X = x$

# Definitions and Notation

- **Random variable:**  $X$
- **Realization:**  $X = x$
- **Sample:**  $\mathbf{X} = (X_1, X_2, \dots, X_n)$

# Definitions and Notation

- **Random variable:**  $X$
- **Realization:**  $X = x$
- **Sample:**  $\mathbf{X} = (X_1, X_2, \dots, X_n)$
- **Realization:**  $\mathbf{X} = \mathbf{x}$

# Definitions and Notation

- **Random variable:**  $X$
- **Realization:**  $X = x$
- **Sample:**  $\mathbf{X} = (X_1, X_2, \dots, X_n)$
- **Realization:**  $\mathbf{X} = \mathbf{x}$
- **Random sample:** a sample where the random variables are independent and identically distributed.

# Definitions and Notation

- **Random variable:**  $X$
- **Realization:**  $X = x$
- **Sample:**  $\mathbf{X} = (X_1, X_2, \dots, X_n)$
- **Realization:**  $\mathbf{X} = \mathbf{x}$
- **Random sample:** a sample where the random variables are independent and identically distributed.
- **Joint density of the sample:** given a random sample, it is the product of the individual densities,

$$f(\mathbf{x}; \theta) = f(x_1, x_2, \dots, x_n; \theta) = \prod_{k=1}^n f_k(x_k; \theta).$$

# The Likelihood Function

Let  $f(\mathbf{x}; \theta)$  be the joint density of the sample.

The *likelihood function* is

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta).$$

Isn't that the joint density of the sample? What's the difference?

# The Likelihood Function

Let  $f(\mathbf{x}; \theta)$  be the joint density of the sample.

The *likelihood function* is

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta).$$

Isn't that the joint density of the sample? What's the difference?

Which variable we consider fixed.

# The Likelihood Function

Let  $f(\mathbf{x}; \theta)$  be the joint density of the sample.

The *likelihood function* is

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta).$$

Isn't that the joint density of the sample? What's the difference?

Which variable we consider fixed.

$f(\mathbf{x}; \theta) \rightarrow \theta$  is fixed and  $\mathbf{x}$  is variable.

# The Likelihood Function

Let  $f(\mathbf{x}; \theta)$  be the joint density of the sample.

The *likelihood function* is

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta).$$

Isn't that the joint density of the sample? What's the difference?

Which variable we consider fixed.

$f(\mathbf{x}; \theta) \rightarrow \theta$  is fixed and  $\mathbf{x}$  is variable.

$L(\theta; \mathbf{x}) \rightarrow \mathbf{x}$  is the observed sample point and  $\theta$  varies over all possible parameter values.

# Comparing Likelihoods

If  $\mathbf{X}$  is discrete,  $L(\theta; \mathbf{x}) = \Pr(\mathbf{X} = \mathbf{x})$ .

# Comparing Likelihoods

If  $\mathbf{X}$  is discrete,  $L(\theta; \mathbf{x}) = \Pr(\mathbf{X} = \mathbf{x})$ .

If we compare the likelihood function at 2 different values of  $\theta$ ,  $\theta_1$  and  $\theta_2$ , we may find that

$$\Pr_{\theta_1}(\mathbf{X} = \mathbf{x}) = L(\theta_1; \mathbf{x}) > L(\theta_2; \mathbf{x}) = \Pr_{\theta_2}(\mathbf{X} = \mathbf{x}).$$

# Comparing Likelihoods

If  $\mathbf{X}$  is discrete,  $L(\theta; \mathbf{x}) = \Pr(\mathbf{X} = \mathbf{x})$ .

If we compare the likelihood function at 2 different values of  $\theta$ ,  $\theta_1$  and  $\theta_2$ , we may find that

$$\Pr_{\theta_1}(\mathbf{X} = \mathbf{x}) = L(\theta_1; \mathbf{x}) > L(\theta_2; \mathbf{x}) = \Pr_{\theta_2}(\mathbf{X} = \mathbf{x}).$$

What this tells us is that the sample actually observed is more likely to have occurred if  $\theta = \theta_1$ , than if  $\theta = \theta_2$ .

# Comparing Likelihoods

If  $\mathbf{X}$  is discrete,  $L(\theta; \mathbf{x}) = \Pr(\mathbf{X} = \mathbf{x})$ .

If we compare the likelihood function at 2 different values of  $\theta$ ,  $\theta_1$  and  $\theta_2$ , we may find that

$$\Pr_{\theta_1}(\mathbf{X} = \mathbf{x}) = L(\theta_1; \mathbf{x}) > L(\theta_2; \mathbf{x}) = \Pr_{\theta_2}(\mathbf{X} = \mathbf{x}).$$

What this tells us is that the sample actually observed is more likely to have occurred if  $\theta = \theta_1$ , than if  $\theta = \theta_2$ .

So  $\theta_1$  is a more plausible value for the true value of  $\theta$  than is  $\theta_2$ .

# Maximum Likelihood Estimators

# Maximum Likelihood Estimators

(Aside: what is an estimator?)

# Maximum Likelihood Estimators

(Aside: what is an estimator?)

For each sample point  $\mathbf{x}$ , let  $\hat{\theta}(\mathbf{x})$  be a parameter value at which  $L(\theta; \mathbf{x})$  attains its maximum as a function of  $\theta$ , with  $\mathbf{x}$  held fixed.

# Maximum Likelihood Estimators

(Aside: what is an estimator?)

For each sample point  $\mathbf{x}$ , let  $\hat{\theta}(\mathbf{x})$  be a parameter value at which  $L(\theta; \mathbf{x})$  attains its maximum as a function of  $\theta$ , with  $\mathbf{x}$  held fixed.

The maximum likelihood estimator is

$$L(\hat{\theta}; \mathbf{x}) = \max_{\hat{\theta}} L(\theta; \mathbf{x}).$$

# Maximum Likelihood Estimators

(Aside: what is an estimator?)

For each sample point  $\mathbf{x}$ , let  $\hat{\theta}(\mathbf{x})$  be a parameter value at which  $L(\theta; \mathbf{x})$  attains its maximum as a function of  $\theta$ , with  $\mathbf{x}$  held fixed.

The maximum likelihood estimator is

$$L(\hat{\theta}; \mathbf{x}) = \max_{\hat{\theta}} L(\theta; \mathbf{x}).$$

Remember the intuition. . .the MLE is the parameter point for which the observed sample is most likely.

# Finding the MLE

We find the MLE by finding the *maximum* of the likelihood function.

# Finding the MLE

We find the MLE by finding the *maximum* of the likelihood function.

We do this by taking the first derivative of the likelihood function and setting it equal to zero.

# Finding the MLE

We find the MLE by finding the *maximum* of the likelihood function.

We do this by taking the first derivative of the likelihood function and setting it equal to zero.

Why do we care about the derivative of the function?

# Finding the MLE

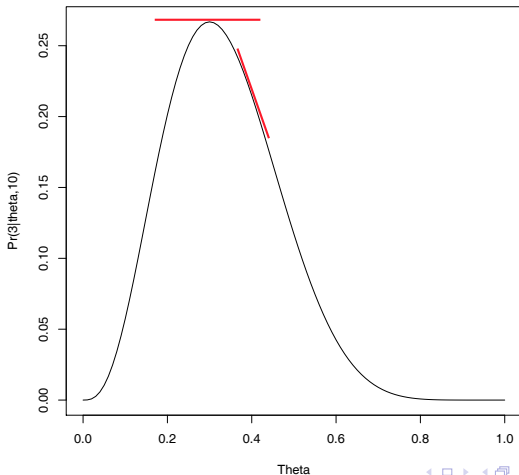
We find the MLE by finding the *maximum* of the likelihood function.

We do this by taking the first derivative of the likelihood function and setting it equal to zero.

Why do we care about the derivative of the function?

Because, loosely speaking, the derivative at a point on the curve is the slope of the line that is tangent to the curve at the point. And the line that is tangent to the maximum must have a particular slope....

# The Tangent Line



# Finding the MLE

So, the maximization problem to solve is

$$\max_{\theta} L(\theta; \mathbf{x}),$$

and we solve it by setting the derivative to zero:

$$\frac{dL(\theta; \mathbf{x})}{d\theta} = 0,$$

and solving for  $\theta$ .

## Example 1

Suppose that we have a random sample from a Bernoulli distribution (each value takes a 0 or 1) for which  $\theta$  is unknown.

# Example 1

Suppose that we have a random sample from a Bernoulli distribution (each value takes a 0 or 1) for which  $\theta$  is unknown.

The likelihood function is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}.$$

# Example 1

Suppose that we have a random sample from a Bernoulli distribution (each value takes a 0 or 1) for which  $\theta$  is unknown.

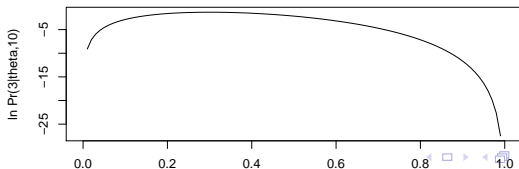
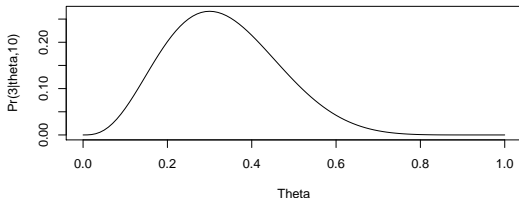
The likelihood function is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}.$$

Now, it turns out that it is easier to maximize likelihood functions after taking the log of the likelihood function.

# The Log-likelihood Function

We can do this because the log is a monotonic transformation, which has its maximum in the same place.



# The Log-likelihood Function

But why should we take the log?

# The Log-likelihood Function

But why should we take the log?

Remember that the likelihood function is a product.

# The Log-likelihood Function

But why should we take the log?

Remember that the likelihood function is a product.

- 1 It is easier to take expectations and variances of sums as opposed to products.

# The Log-likelihood Function

But why should we take the log?

Remember that the likelihood function is a product.

- 1 It is easier to take expectations and variances of sums as opposed to products.
- 2 It makes it possible for the computer to deal with large numbers of observations.

## Let's Explain Point 2

In a simple case, the likelihood is equal to a probability:

$$L(\theta; \mathbf{x}_j) = \Pr(\text{we would observe } \mathbf{x}_j).$$

## Let's Explain Point 2

In a simple case, the likelihood is equal to a probability:

$$L(\theta; \mathbf{x}_j) = \Pr(\text{we would observe } \mathbf{x}_j).$$

The likelihood for an entire data set would then be

$$\Pr(\text{data set}) = \Pr(\text{datum 1}) \times \Pr(\text{datum 2}) \times \cdots \times \Pr(\text{datum N}).$$

## Let's Explain Point 2

In a simple case, the likelihood is equal to a probability:

$$L(\theta; \mathbf{x}_j) = \Pr(\text{we would observe } \mathbf{x}_j).$$

The likelihood for an entire data set would then be

$$\Pr(\text{data set}) = \Pr(\text{datum 1}) \times \Pr(\text{datum 2}) \times \cdots \times \Pr(\text{datum N}).$$

Let's say we have 2000 observations, and each probability is around 0.5 The probability of the data set would be

$$.5^{2000} = 2 \times 10^{-603}$$

## Let's Explain Point 2

In a simple case, the likelihood is equal to a probability:

$$L(\theta; \mathbf{x}_j) = \Pr(\text{we would observe } \mathbf{x}_j).$$

The likelihood for an entire data set would then be

$$\Pr(\text{data set}) = \Pr(\text{datum 1}) \times \Pr(\text{datum 2}) \times \cdots \times \Pr(\text{datum N}).$$

Let's say we have 2000 observations, and each probability is around 0.5 The probability of the data set would be

$$.5^{2000} = 2 \times 10^{-603}$$

This value is too small to be computed by most computers. But if we take the log....

$$\ln(.5^{2000}) = 2000 \times \ln(.5) \approx 2000 \times -0.6931 = -1386.2.$$

# Example 1 continued

The likelihood function is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}.$$

# Example 1 continued

The likelihood function is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}.$$

Taking the log,

$$\begin{aligned} \ln L(\theta; \mathbf{x}) &= \sum_{i=1}^n [x_i \ln \theta + (1 - x_i) \ln(1 - \theta)] \\ &= \left( \sum_{i=1}^n x_i \right) \ln \theta + \left( \sum_{i=1}^n [1 - x_i] \right) \ln(1 - \theta) \end{aligned}$$

# Example 1 continued

Next we take the log and set it equal to 0,

$$\begin{aligned}\frac{d \ln L(\theta; \mathbf{x})}{d\theta} &= \left( \sum_{i=1}^n x_i \right) \frac{1}{\theta} - \left( \sum_{i=1}^n [1 - x_i] \right) \frac{1}{1 - \theta} = 0 \\ (1 - \theta) \sum_{i=1}^n x_i - \theta \left( \sum_{i=1}^n [1 - x_i] \right) &= 0 \\ \sum_{i=1}^n x_i - \theta \sum_{i=1}^n x_i - n\theta + \theta \sum_{i=1}^n x_i &= 0 \\ \hat{\theta} &= \frac{\sum_{i=1}^n x_i}{n}\end{aligned}$$

## Example 2

Suppose that we have a random sample from a Poisson distribution. The likelihood is

$$L(\lambda; \mathbf{x}) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}.$$

## Example 2

Suppose that we have a random sample from a Poisson distribution. The likelihood is

$$L(\lambda; \mathbf{x}) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}.$$

Taking the log,

$$\ln L(\lambda; \mathbf{x}) = \sum_{i=1}^n [-\lambda + x_i \ln \lambda - \ln x_i!].$$

## Example 2: the derivative

$$\begin{aligned}\frac{d \ln L(\lambda; \mathbf{x})}{d\lambda} &= \sum_{i=1}^n \left( -1 + \frac{x_i}{\lambda} \right) = 0 \\ \sum_{i=1}^n (-1) + \sum_{i=1}^n \frac{x_i}{\lambda} &= 0 \\ \sum_{i=1}^n \frac{x_i}{\lambda} &= n \\ \hat{\lambda} &= \frac{1}{n} \sum_{i=1}^n x_i\end{aligned}$$

## Example 3

Suppose that we have a random sample from the following density:

$$f(\mathbf{x}; \theta) = \begin{cases} \theta x_i^{\theta-1} & \text{for } 0 < x_i < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the MLE of  $\theta$  ( $\theta > 0$ ).

## Example 3: solution

$$\ln L(\theta; \mathbf{x}) = \sum_{i=1}^n [\ln \theta + (\theta - 1) \ln x_i].$$

## Example 3: solution

$$\ln L(\theta; \mathbf{x}) = \sum_{i=1}^n [\ln \theta + (\theta - 1) \ln x_i].$$

$$\frac{d \ln L(\theta; \mathbf{x})}{d\theta} = \sum_{i=1}^n \left[ \frac{1}{\theta} + \ln x_i \right] = 0$$

$$\sum_{i=1}^n \frac{1}{\theta} = - \sum_{i=1}^n \ln x_i$$

$$\hat{\theta} = - \frac{n}{\sum_{i=1}^n \ln x_i}$$

## Technical Point

A first derivative that equals 0 is a *necessary*, but not *sufficient* condition for a maximum.

How can we distinguish between a maximum and a minimum?

## Technical Point

A first derivative that equals 0 is a *necessary*, but not *sufficient* condition for a maximum.

How can we distinguish between a maximum and a minimum?

A summit is reached when we ascend and then descend. Thus, the derivative must be declining, going from positive to zero to negative.

That is, the second derivative (the derivative of the derivative) must be negative.

## Example 3: A Maximum?

The log-likelihood function:

$$g(\theta) = \ln L(\theta; \mathbf{x}) = \sum_{i=1}^n [\ln \theta + (\theta - 1) \ln x_i].$$

## Example 3: A Maximum?

The log-likelihood function:

$$g(\theta) = \ln L(\theta; \mathbf{x}) = \sum_{i=1}^n [\ln \theta + (\theta - 1) \ln x_i].$$

The first derivative:

$$g'(\theta) = \sum_{i=1}^n \left[ \frac{1}{\theta} + \ln x_i \right]$$

## Example 3: A Maximum?

The log-likelihood function:

$$g(\theta) = \ln L(\theta; \mathbf{x}) = \sum_{i=1}^n [\ln \theta + (\theta - 1) \ln x_i].$$

The first derivative:

$$g'(\theta) = \sum_{i=1}^n \left[ \frac{1}{\theta} + \ln x_i \right]$$

The second derivative:

$$g''(\theta) = \sum_{i=1}^n \left[ -\frac{1}{\theta^2} \right]$$

# Simple Regression in MLE

The model is

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim IID N(0, \sigma^2).$$

# Simple Regression in MLE

The model is

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim \text{IID } N(0, \sigma^2).$$

Since the  $y_i$  are independently and normally distributed with means  $\alpha + \beta x_i$  and common variance,  $\sigma^2$ , the density of each observation follows a normal distribution:

$$f(x_i) = \left( \frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left[ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right]$$

or

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

# Simple Regression in MLE

The likelihood is therefore the product over the  $n$  observations:

$$L(\mathbf{y}) = \prod_{i=1}^n \left( \frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left[ -\frac{1}{2\sigma^2} (y_i - \alpha - \beta x_i)^2 \right],$$

and the log-likelihood is

$$\ln L = \sum_{i=1}^n \left[ -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_i - \alpha - \beta x_i)^2 \right].$$

# Simple Regression in MLE: Estimating $\alpha$

Let  $Q = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ .

$$\frac{\partial Q}{\partial \alpha} = \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-1) = 0$$

$$\sum_{i=1}^n y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^n x_i$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

# Simple Regression in MLE: Estimating $\beta$

Let  $Q = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ .

$$\frac{\partial Q}{\partial \beta} = \sum_{i=1}^n 2(y_i - \alpha - \beta x_i)(-x_i) = 0$$

$$\sum_{i=1}^n y_i x_i = \hat{\alpha} \sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i^2$$

Now substitute in for  $\hat{\alpha}$ .

Simple Regression in MLE: Estimating  $\beta$ 

$$\sum_{i=1}^n y_i x_i = \hat{\alpha} \sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i^2$$

$$\sum_{i=1}^n y_i x_i = (\bar{y} - \hat{\beta} \bar{x}) n \bar{x} + \hat{\beta} \sum_{i=1}^n x_i^2$$

$$\sum_{i=1}^n y_i x_i - n \bar{x} \bar{y} = \hat{\beta} \left( \sum_{i=1}^n x_i^2 - n \bar{x}^2 \right)$$

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

# Why We Like ML Estimators

# Why We Like ML Estimators

The major properties of MLEs are *large sample* or *asymptotic*.

# Why We Like ML Estimators

The major properties of MLEs are *large sample* or *asymptotic*.

- 1 They are consistent,  $\text{plim}(\hat{\theta}) = \theta$ .

# Why We Like ML Estimators

The major properties of MLEs are *large sample* or *asymptotic*.

- 1 They are consistent,  $\text{plim}(\hat{\theta}) = \theta$ .
- 2 They are asymptotically normal.

# Why We Like ML Estimators

The major properties of MLEs are *large sample* or *asymptotic*.

- 1 They are consistent,  $\text{plim}(\hat{\theta}) = \theta$ .
- 2 They are asymptotically normal.
- 3 They are asymptotically efficient.

# Why We Like ML Estimators

The major properties of MLEs are *large sample* or *asymptotic*.

- 1 They are consistent,  $\text{plim}(\hat{\theta}) = \theta$ .
- 2 They are asymptotically normal.
- 3 They are asymptotically efficient.
- 4 They are invariant to transformation: if  $\hat{\theta}$  is the MLE of  $\theta$  and  $g(\theta)$  is a continuous function of  $\theta$ , then  $g(\hat{\theta})$  is the MLE of  $g(\theta)$ .

# Where Do the Standard Errors Come From?

# Where Do the Standard Errors Come From?

(Aside: what is a standard error?)

# Where Do the Standard Errors Come From?

(Aside: what is a standard error?)

The estimated variation of  $\hat{\theta}$  is given by  $-\mathbf{H}^{-1}$ , where  $\mathbf{H}$  is the matrix of second derivatives (better known as the Hessian),

$$\begin{aligned}\mathbf{H} &= \frac{\partial^2 \ln L(\hat{\theta}; \mathbf{X})}{\partial \hat{\theta} \partial \hat{\theta}'} \\ &= \frac{\partial^2 \ln \ell(\hat{\theta}; \mathbf{x}_i)}{\partial \hat{\theta} \partial \hat{\theta}'} + \dots + \frac{\partial^2 \ln \ell(\hat{\theta}; \mathbf{x}_N)}{\partial \hat{\theta} \partial \hat{\theta}'}\end{aligned}$$

# Where Do the Standard Errors Come From?

(Aside: what is a standard error?)

The estimated variation of  $\hat{\theta}$  is given by  $-\mathbf{H}^{-1}$ , where  $\mathbf{H}$  is the matrix of second derivatives (better known as the Hessian),

$$\begin{aligned}\mathbf{H} &= \frac{\partial^2 \ln L(\hat{\theta}; \mathbf{X})}{\partial \hat{\theta} \partial \hat{\theta}'} \\ &= \frac{\partial^2 \ln \ell(\hat{\theta}; \mathbf{x}_i)}{\partial \hat{\theta} \partial \hat{\theta}'} + \dots + \frac{\partial^2 \ln \ell(\hat{\theta}; \mathbf{x}_N)}{\partial \hat{\theta} \partial \hat{\theta}'}\end{aligned}$$

The estimated standard errors are the square roots of the diagonal of  $-\mathbf{H}^{-1}$ .

# How We Get SEs — Step 1

We stated earlier that ML estimates are asymptotically normal,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[\mathbf{0}, \text{Var}(\hat{\theta})].$$

# How We Get SEs — Step 1

We stated earlier that ML estimates are asymptotically normal,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[\mathbf{0}, \text{Var}(\hat{\theta})].$$

If  $\mathbf{g}(\theta)$  is some function of  $\theta$ , then it can be shown that

$$\text{Var}(\mathbf{g}(\hat{\theta})) = (D\mathbf{g}(\hat{\theta})) \text{Var}(\hat{\theta})(D\mathbf{g}(\hat{\theta}))',$$

where  $D = \frac{\partial}{\partial \theta}$ .

## How We Get SEs — Step 2

Let  $\mathbf{g}(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta}}$ , which is also known as the *score function*.

## How We Get SEs — Step 2

Let  $\mathbf{g}(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta}}$ , which is also known as the *score function*.

Therefore,

$$\begin{aligned}\text{Var}(\mathbf{g}(\hat{\theta})) &= (D\mathbf{g}(\hat{\theta})) \text{Var}(\hat{\theta})(D\mathbf{g}(\hat{\theta}))' \\ &= \left( \frac{\partial^2 \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta} \partial \hat{\theta}'} \right) \text{Var}(\hat{\theta}) \left( \frac{\partial^2 \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta} \partial \hat{\theta}'} \right)' \\ &= \mathbf{H} \text{Var}(\hat{\theta}) \mathbf{H}\end{aligned}$$

## How We Get SEs — Step 2

Let  $\mathbf{g}(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta}}$ , which is also known as the *score function*.

Therefore,

$$\begin{aligned}\text{Var}(\mathbf{g}(\hat{\theta})) &= (D\mathbf{g}(\hat{\theta})) \text{Var}(\hat{\theta})(D\mathbf{g}(\hat{\theta}))' \\ &= \left( \frac{\partial^2 \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta} \partial \hat{\theta}'} \right) \text{Var}(\hat{\theta}) \left( \frac{\partial^2 \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta} \partial \hat{\theta}'} \right)' \\ &= \mathbf{H} \text{Var}(\hat{\theta}) \mathbf{H}\end{aligned}$$

Thus,

$$\text{Var}(\hat{\theta}) = \mathbf{H}^{-1} \text{Var}(\mathbf{g}(\hat{\theta})) \mathbf{H}^{-1}.$$

## How We Get SEs — Step 3

What is  $\text{Var}(\mathbf{g}(\hat{\theta}))$ ?

## How We Get SEs — Step 3

What is  $\text{Var}(\mathbf{g}(\hat{\theta}))$ ?

Since the expected value of the score function is 0, the variance of the score function is the score function squared,

$$\begin{aligned}\text{Var}(\mathbf{g}(\hat{\theta})) &= E[\mathbf{g}(\hat{\theta})^2] \\ &= E\left[\left(\frac{\partial \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta}}\right)^2\right] \\ &= -E\left[\frac{\partial^2 \ln L(\hat{\theta}; \mathbf{x})}{\partial \hat{\theta} \partial \hat{\theta}'}\right] \\ &= -\mathbf{H}\end{aligned}$$

# How We Get SEs — Last Step

Since  $\text{Var}(\mathbf{g}(\hat{\theta})) = -\mathbf{H}$ ,

$$\begin{aligned}\text{Var}(\hat{\theta}) &= \mathbf{H}^{-1} \text{Var}(\mathbf{g}(\hat{\theta}))\mathbf{H}^{-1} \\ &= \mathbf{H}^{-1}(-\mathbf{H})\mathbf{H}^{-1} \\ &= -\mathbf{H}^{-1}\end{aligned}$$

# How We Get SEs — Last Step

Since  $\text{Var}(\mathbf{g}(\hat{\theta})) = -\mathbf{H}$ ,

$$\begin{aligned}\text{Var}(\hat{\theta}) &= \mathbf{H}^{-1} \text{Var}(\mathbf{g}(\hat{\theta}))\mathbf{H}^{-1} \\ &= \mathbf{H}^{-1}(-\mathbf{H})\mathbf{H}^{-1} \\ &= -\mathbf{H}^{-1}\end{aligned}$$

Other names for this result:

- Inverse of the Fisher information matrix
- Cramér-Rao Lower Bound

# R

Why *R*?

# R

## Why *R*?

- it has no limitations

## Why *R*?

- it has no limitations
- the calculations are correct

# R

## Why *R*?

- it has no limitations
- the calculations are correct
- it's free!

# Enough?

How about a coffee break?