1

Introduction to Maximum Likelihood

1.1 INTRODUCTION TO MAXIMUM LIKELIHOOD

The method of maximum likelihood is more than a collection of statistical models or even an estimation procedure. It is a unified way of thinking about model construction, estimation, and evaluation. The study of maximum likelihood represents a transition in methodological training for social scientists. It marks the point at which we possess the conceptual, mathematical, and computational foundations for writing down our own statistical estimators that can be custom-designed for our own research questions. A solid understanding of the principles and properties of maximum likelihood is fundamental to more advanced study, whether self-directed or formally course-based.

To begin our introduction to the maximum likelihood approach we present a toy example involving the most hackneyed of statistics contrivances: coin flips. We undertake this example to illustrate the mechanics of the likelihood with maximal simplicity. We then move on to a more realistic problem: describing the degree of association between two continuous variables. Least squares regression – the portal through which nearly every researcher enters the realm of applied statistics – is a common tool for describing such a relationship. Our goal is to introduce the broader likelihood framework for statistical inference, showing that the familiar least squares estimator is, in fact, a special type of maximum likelihood estimator. We then provide a more general outline of the likelihood approach to model building, something we revisit in more mathematical and computational detail in the next three chapters.

1.2 COIN FLIPS AND MAXIMUM LIKELIHOOD

Three friends are trying to decide between two restaurants, an Ethiopian restaurant and a brewpub. Each is indifferent, since none of them has previously

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

eaten at either restaurant. They each flip a single coin, deciding that a heads will indicate a vote for the brewpub. The result is two heads and one tails. The friends deposit the coin in the parking meter and go to the brewpub.

We might wonder whether the coin was, in fact, fair. As a data analysis problem, these coin flips were not obtained in a traditional sampling framework, nor are we interested in making inferences about the general class of restaurant coin flips. Rather, the three flips of a single coin are all the data that exist, and we just want to know how the decision was taken. This is a binary outcomes problem. The data are described by the following set in which 1 represents heads: $\{1,1,0\}$. Call the probability of a flip in favor of eating at the brewpub θ; the probability of a flip in favor of eating Ethiopian is thereby $1-\theta$. In other words, we assume a Bernoulli distribution for a coin flip.

In case you were wondering ... 1.1 Bernoulli distribution

Let $Y \in \{0,1\}$. Suppose $\Pr(Y=1)=\theta$. We say that Y follows a Bernoulli distribution with parameter θ:

$$Y \sim f_B(y;\theta) = \begin{cases} \theta^y(1-\theta)^{1-y} & \forall \quad y \in \{0,1\}, \\ 0 & \text{otherwise} \end{cases}$$

with $\mathrm{E}[Y]=\theta$ and $\mathrm{var}(Y)=\theta(1-\theta)$.

What value of the parameter, θ, best describes the observed data? Prior experience may lead us to believe that coin flips are equiprobable; $\hat{\theta}=0.5$ seems a reasonable guess. Further, one might also reason that since there are three pieces of data, the probability of the joint outcome of three flips is $0.5^3=0.125$. This may be a reasonable summary of our prior expectations, but this calculation fails to take advantage of the actual data at hand to inform our estimate.

A simple tabulation reveals this insight more clearly. We know that in this example, θ is defined on the interval $[0,1]$, i.e., $0 \leq \theta \leq 1$. We also know that unconditional probabilities compound themselves so that the probability of a head on the first coin toss times the probability of a head on the second times the probability of tails on the third produces the joint probability of the observed data: $\theta \times \theta \times (1-\theta)$. Given this expression we can easily calculate the probability of getting the observed data for different values of θ. Computationally, the results are given by $\Pr(y_1 \mid \hat{\theta}) \times \Pr(y_2 \mid \hat{\theta}) \times \Pr(y_3 \mid \hat{\theta})$, where y_i is the value of each observation, $i \in \{1,2,3\}$ and $\mid \hat{\theta}$ is read, "given the proposed value of θ." Table 1.1 displays these calculations in increments of 0.1.

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

TABLE 1.1 *Choosing a restaurant with three flips of a fair coin?*

Observed Data			
y	$\hat{\theta}$	$\theta^{\mathbf{1s}} \times (1-\theta)^{\mathbf{0s}}$	$f_B(\mathbf{y} \mid \hat{\theta})$
$\{1,1,0\}$	0.00	$0.00^2 \times (1-0.00)^1$	0.000
$\{1,1,0\}$	0.10	$0.10^2 \times (1-0.10)^1$	0.009
$\{1,1,0\}$	0.20	$0.20^2 \times (1-0.20)^1$	0.032
$\{1,1,0\}$	0.30	$0.30^2 \times (1-0.30)^1$	0.063
$\{1,1,0\}$	0.40	$0.40^2 \times (1-0.40)^1$	0.096
$\{1,1,0\}$	0.50	$0.50^2 \times (1-0.50)^1$	0.125
$\{1,1,0\}$	0.60	$0.60^2 \times (1-0.60)^1$	0.144
$\{1,1,0\}$	0.67	$0.67^2 \times (1-0.67)^1$	0.148
$\{1,1,0\}$	0.70	$0.70^2 \times (1-0.70)^1$	0.147
$\{1,1,0\}$	0.80	$0.80^2 \times (1-0.80)^1$	0.128
$\{1,1,0\}$	0.90	$0.90^2 \times (1-0.90)^1$	0.081
$\{1,1,0\}$	1.00	$1.00^2 \times (1-0.00)^1$	0.000

The a priori guess of 0.5 turns out not to be the most likely to have generated these data. Rather, the value of $\frac{2}{3}$ is the most likely value for θ. It is not necessary to do all of this by guessing values of θ. This case can be solved analytically.

When we have data on each of the trials (flips), the Bernoulli probability model, f_B, is a natural place to start. We will call the expression that describes the joint probability of the observed data as function of the parameters the *likelihood function*, denoted $\mathcal{L}(\mathbf{y};\theta)$. We can use the tools of differential calculus to solve for the maximum; we take the logarithm of the likelihood for computational convenience:

$$\begin{aligned}
\mathcal{L} &= \theta^2(1-\theta)^1 \\
\log \mathcal{L} &= 2\log\theta + 1\log(1-\theta) \\
\frac{\partial \log \mathcal{L}}{\partial \theta} &= \frac{2}{\theta} - \frac{1}{(1-\theta)} = 0 \\
\hat{\theta} &= \frac{2}{3}.
\end{aligned}$$

The value of θ that the maximizes the likelihood function is called the *maximum likelihood estimate*, or MLE.

It is clear that, in this case, it does not matter who gets heads and who gets tails. Only the number of heads out of three flips matters. When Bernoulli data are grouped in such a way, we can describe them equivalently with the closely related *binomial* distribution.

In case you were wondering ... 1.2 Binomial distribution

Let $Y \sim f_B(y; p)$ where $\Pr(Y = 1) = p$. Suppose we take n independent draws and let $X = \sum_{i=1}^{n} Y_i$. We say that X follows a binomial distribution with parameter $\boldsymbol{\theta} = (n, p)$:

$$X \sim f_b(x; n, p)$$

$$\Pr(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \forall \quad k \in \{0, \ldots, n\}, \\ 0 & \forall \quad k \notin \{0, \ldots, n\} \end{cases}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and with $\mathrm{E}[X] = np$ and $\mathrm{var}(X) = np(1-p)$. The Bernoulli distribution is a binomial distribution with $n = 1$.

Jacob Bernoulli was a Swiss mathematician who derived the law of large numbers, discovered the mathematical constant e, and formulated the eponymous Bernoulli and binomial distributions.

Analytically and numerically the MLE is equivalent whether derived using the Bernoulli or binomial distribution with known n. Figure 1.1 illustrates the likelihood function for Bernoulli/binomial data consisting of two heads and one tail. The maximum occurs at $\hat{\theta} = 2/3$.

1.3 SAMPLES AND SAMPLING DISTRIBUTIONS

Trying to decide whether the coin used to choose a restaurant is fair is a problem of statistical inference. Many inferential approaches are plausible; most catholic among them is the classical model based on *asymptotic* results obtained by imagining repeated, independent samples drawn from a fixed population.[1] As a result, we often conceptualize statistics calculated from samples as providing information on a population parameter. For example, suppose $x_1, \ldots, x_n$ comprise a random sample from a population with a mean of μ and a variance of σ^2. It follows that the mean of this sample is a random variable with a mean of μ and variance that is equal to σ^2/n. Why? This is true, since the expected value of the mean of an independent sample is the mean of the population from which the sample is drawn. The variance of the sample is, similarly, equal to the variance of the population divided by the size of the sample.[2] This is demonstrated graphically in Figure 1.2.

[1] In statistics, asymptotic analysis refers to theoretical results describing the limiting behavior of a function as a value, typically the sample size, tends to infinity.

[2] This is the most basic statement of the Central Limit Theorem. We state the theorem more formally in Section 2.2.2.

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

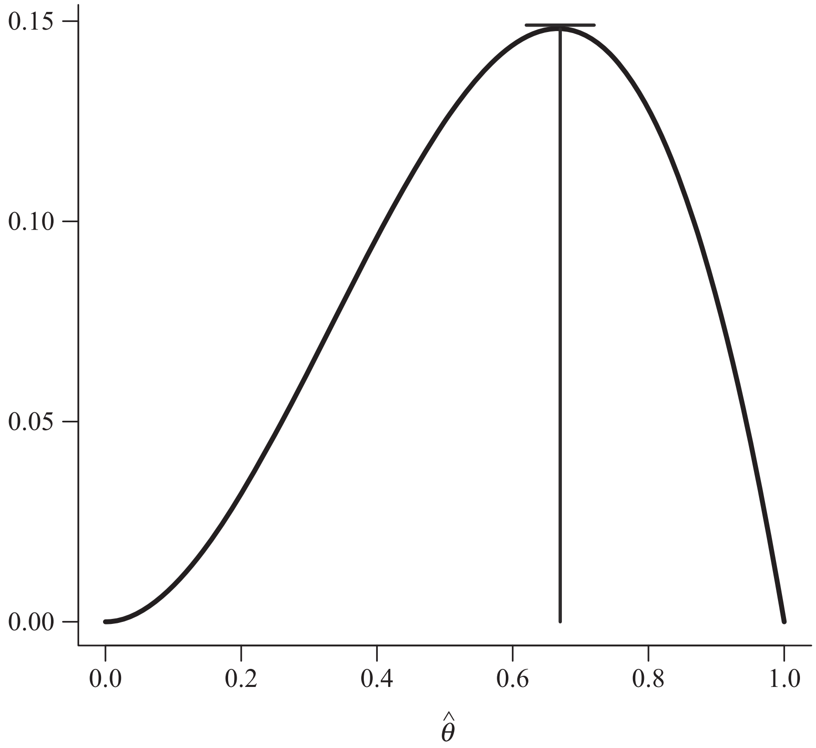

FIGURE 1.1 The likelihood/probability of getting two heads in three coin tosses, over various values of θ.

This basic result is often used to interpret output from statistical models as if the observed data are a sample from a population for which the mean and variance are *unknown*. We can use a random sample to calculate our best guesses as to what those population values are. If we have either a large enough random sample of the population or enough independent, random samples – as in the American National Election Study or the Eurobarometer surveys, for example – then we can retrieve good estimates of the population parameters of interest without having to actually conduct a census of the entire population. Indeed, most statistical procedures are based on the idea that they perform well in repeated samples, i.e., they have good sampling distribution properties.

Observed data in the social sciences frequently fail to conform to a "sample" in the classical sense. They instead consist of observations on a particular (nonrandom) selection of units, such as the 50 US states, all villages in Afghanistan safe for researchers to visit, all the terror events that newspapers choose to report, or, as in the example that follows, all the data available from the World Bank on GDP *and* CO_2 emissions from 2012. In such situations,

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

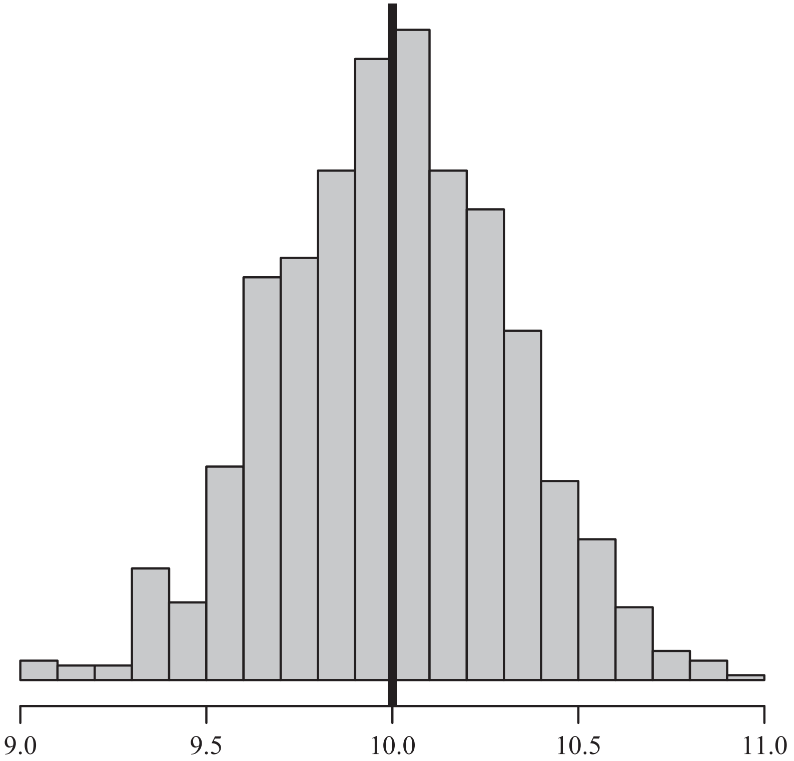

FIGURE 1.2 Illustrating the Central Limit Theorem with a histogram of the means of 1,000 random samples of size 10 drawn from a population with mean of 10 and variance of 1.

the basic sampling distribution result is more complicated to administer. To salvage the classical approach, some argue for a *conceptual* sampling perspective. This often takes the form of a hypothetical: you have data on all 234 countries in the world at present, but these are just a sample from all the *possible* worlds that might have existed. The implied conclusion is that you can treat this as a random sample and gain leverage from the basic results of sampling distributions and the asymptotic properties of the least squares estimators.

Part of the problem with this line of attack is that it sets up the expectation that we are using the observed data to learn about some larger, possibly hypothetical, population. Standard inference frequently relies on this conception, asking questions like "How likely are estimates at least as large as what we found if, in the larger population, the 'true value' is 0?" We speak of estimates as *(non)significant* by way of trying to demonstrate that they did not arise by chance and really do reflect accurately the unobserved, underlying population

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

parameters. In the same way, we argue that the estimators we employ are good if they produce *unbiased* estimates of population parameters. Thus we conceptualize the problem as having estimates that are shifted around by estimators and sample sizes.

But there is a different way to think about all of this, a way that is not only completely different, but complementary at the same time.

In case you were wondering ... 1.3 Bias and mean squared error

A statistical estimator is simply a formula or algorithm for calculating some unknown quantity using observed data. Let $T(X)$ be an estimator for θ. The bias of $T(X)$, denoted bias(θ), is

$$\text{bias}(\theta) = \text{E}[T(X)] - \theta.$$

The mean squared error, MSE(θ), is given as

$$\begin{aligned}\text{MSE}(\theta) &= \text{E}[(T(X) - \theta)^2] \\ &= \text{var}(T(X)) + \text{bias}(\theta)^2.\end{aligned}$$

1.4 MAXIMUM LIKELIHOOD: AN OVERVIEW

The principle of maximum likelihood is based on the idea that the observed data (even if it is not a random sample) are more likely to have come about as a result of a particular set of parameters. Thus, we flip the problem on its head. *Rather than consider the data as random and the parameters as fixed, the principle of maximum likelihood treats the observed data as fixed and asks: "What parameter values are most likely to have generated the data?"* Thus, the parameters are random variables. More formally, in the likelihood framework we think of the joint probability of the data as a function of parameter values for a particular density or mass function. We call this particular conceptualization of the probability function the *likelihood*, since it is being maximized with respect to the parameters, not on the sample data. The MLEs are those that provide the density or mass function with the highest likelihood of generating the observed data.

1.4.1 Maximum Likelihood: Specific

The World Bank assembled data on gross domestic product and CO_2 emissions for many countries in 2012. These data are accessible directly from $\mathcal{R}$ via the library `WDI` (Arel-Bundock, 2013). If we believe that CO_2 pollution is a linear function of economic activity, then we might propose the simple model $Y =$

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

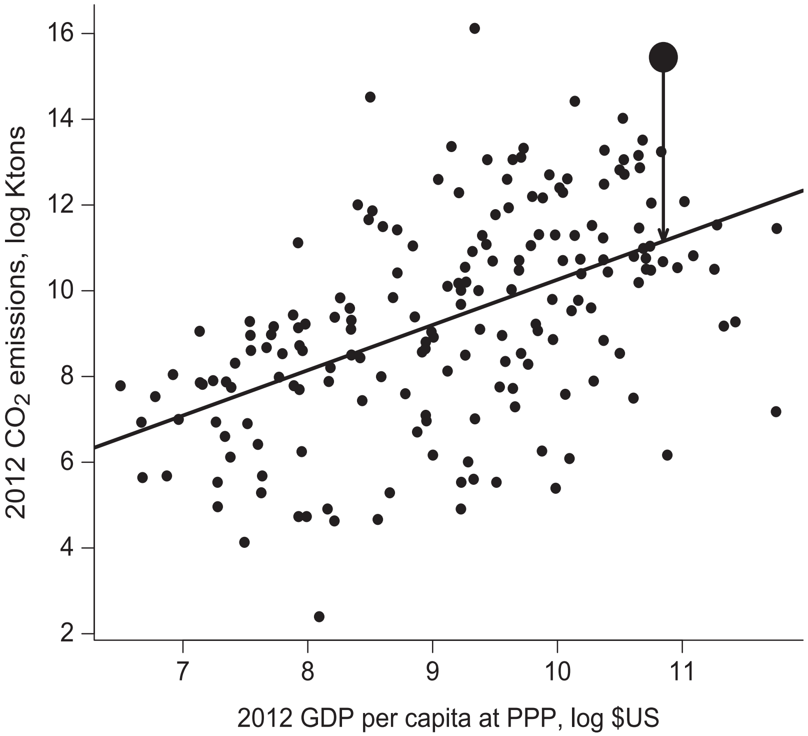

FIGURE 1.3 2012 GDP per capita and CO_2 emissions. The prediction equation is shown as a straight line, with intercept and slope as reported in Table 1.2. The large solid dot represents the United States and the length of the arrow is its residual value given the model.

$\beta_0 + \beta_1 X + \varepsilon$, where Y is the logged data on CO_2 emissions and X is the logged data on gross domestic product (GDP), both taken for 183 countries in the year 2012. The ε term represents the stochastic processes – sampling, measurement error, and other omitted factors – that cause a particular country's observed CO_2 emissions to deviate from the simple linear relationship.

A scatterplot of these data appear in Figure 1.3, with an estimate of the linear relationship included as a straight line. The United States is highlighted for its CO_2 emissions well in excess of what the linear relationship expects, given its per capita GDP. The vertical arrow highlights this positive *residual*.

How can we choose the parameters for the prediction line using maximum likelihood? The first step in constructing any likelihood is the specification of a probability distribution describing the outcome, Y_i. Here we will turn to the Gaussian distribution. If we assume that observations are independently and

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

identically distributed (iid) – they follow the same distribution and contain no dependencies – then we write

$$Y_i \overset{iid}{\sim} \mathcal{N}(\mu_i, \sigma^2). \tag{1.1}$$

Equation 1.1 reads as "Y_i is distributed iid normal with mean μ_i and variance σ^2." When used as a part of a likelihood model, we will adopt the following notational convention:

$$Y_i \sim f_{\mathcal{N}}(y_i; \mu_i, \sigma^2).$$

In case you were wondering ... 1.4 Gaussian (normal) distribution

We say that the random variable $Y \in \mathbb{R}$ follows a Gaussian (or normal) distribution with parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)$ if the probability distribution function can be written as

$$Y \sim f_{\mathcal{N}}(y; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right], \tag{1.2}$$

with $\mathrm{E}[Y] = \mu$ and $\mathrm{var}(Y) = \sigma^2$. The special case in which $\mu = 0$ and $\sigma = 1$ is called the standard normal distribution. The standard normal density and distribution functions are written as $\phi(\cdot)$ and $\Phi(\cdot)$, respectively.

The normal distribution was first derived by Karl Freidrich Gauss and published in his 1810 monograph on celestial mechanics. In the same volume, Gauss derived the least squares estimator and alluded to the principle of maximum likelihood. Gauss, a child prodigy, has long been lauded as the foremost mathematical mind since Newton. Gauss's image along with the formula and graph of the normal distribution appeared on the German 10 mark banknote from 1989 until the mark was superseded by the Euro.

The Marquis de Laplace first proved the Central Limit Theorem in which the mean of repeated random samples follows a Gaussian distribution, paving the way for the distribution's ubiquity in probability and statistics.

Next, we develop a model for the expected outcome – the mean – as a function of covariates. We assume a linear relationship between (log) per capita GDP and (log) CO_2 emissions: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. This implies that $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$. As a result, assuming that Y_i is normal with $\mu = \beta_0 + \beta_1 x_i$ is equivalent to assuming that $\varepsilon_i \sim f_{\mathcal{N}}(\varepsilon_i; 0, \sigma^2)$. That is, assuming Y is iid normal

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

is equivalent to assuming that the errors come from a normal distribution with mean zero and a fixed, constant variance.

Setting aside, for the moment, the connection between ε_i and the independent variable (log per capita GDP), we can further specify the probability distribution for the outcome variable from Equation 1.2:

$$f_{\mathcal{N}}(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(\varepsilon_i)^2}{2\sigma^2}\right].$$

This yields:

$$f_{\mathcal{N}}(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right].$$

Crucially, we need to transform this density function for a single observation into a likelihood for the complete sample. The likelihood is simply a formula for the joint probability distribution of the sample data. The joint probability of independent events A and B, which could represent two outcomes in two separate cases, is simply $\Pr(A) \times \Pr(B)$. The probability of the entire set of observed data is the product of each observation's marginal probability – under this assumption of independence. Since we have assumed that the errors are all independent of one another, the joint probability is just a product of marginal probabilities. Similarly, the likelihood will be a product of the $f_{\mathcal{N}}(\varepsilon_i)$ for each observation, i, in the sample. Thus, the likelihood is

$$\begin{aligned}
&\mathcal{L}(\beta_0, \beta_1, \sigma \mid \{y_1, \ldots, y_n\}, \{x_1, \ldots, x_n\}) \\
&\quad = (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} \exp\left\{\frac{-1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \\
&\quad = (2\pi\sigma^2)^{-n/2} \exp\left\{\sum_{i=1}^{n} \frac{-1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}.
\end{aligned}$$

The likelihood is a function of the parameters (here, $\boldsymbol{\beta}$ and σ) and all of the data on the dependent and independent variables, i.e., it is the formula for the joint probability distribution of the sample. With a likelihood function and the data we can now use the tools of optimization to find the set of parameter values that maximize the value of this likelihood function.

Before doing this, however, the likelihood function can be simplified quite a bit. First, because we are interested in maximizing the function, any monotonically increasing function of the likelihood can serve as the maximand. Since the logarithmic transformation is monotonic and sums are easier to manage than products, we take the natural log. Thus, the *log-likelihood* is:

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

$$
\begin{aligned}
\log \mathcal{L} &= \log \left\{ (2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right] \right\} \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
&= -\frac{1}{2} n \log(2\pi) - n \log \sigma - \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}.
\end{aligned}
$$

Terms that are not functions of parameters to be estimated may be dropped, since these terms only scale the likelihood while leaving it proportional to the original. The maximizing arguments are unchanged. Dropping $\frac{1}{2} n \log(2\pi)$ we have,

$$\log \mathcal{L} = -n \log \sigma - \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}, \tag{1.3}$$

To further simplify matters and because many computer optimization programs default to minimization, we often use $-2 \log \mathcal{L}$,

$$-2 \log \mathcal{L} = n \log \sigma^2 + \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2},$$

and for a fixed or known σ this simplifies to a quantity proportional to the sum of squared errors:

$$-2 \log \mathcal{L} \propto \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{1.4}$$

In practice, there are two ways to solve the likelihood maximization problem. First, if an analytic solution is apparent, then we can solve the likelihood for its extrema directly by taking partial derivatives with respect to each parameter and setting them to zero, as we did with the coin-flipping example. In many instances, however, the derivatives of the likelihood function do not have nice, closed-form solutions. Maximization occurs via numerical techniques described briefly later here and in more detail in Chapter 4.

In the linear-normal case we can find the MLE using analytical methods. Nevertheless, it can be useful to plot the likelihood as a function of parameter values. Figure 1.4 does just that, displaying the likelihood surface for β_1. We can see that the maximum occurs near 1.06.

Using R to Maximize the Likelihood

In this book we emphasize turning mathematical statements and principles into executable computer code in the $\mathcal{R}$ statistical computing environment. For the CO_2 emissions example we can assemble the data directly in $\mathcal{R}$.

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

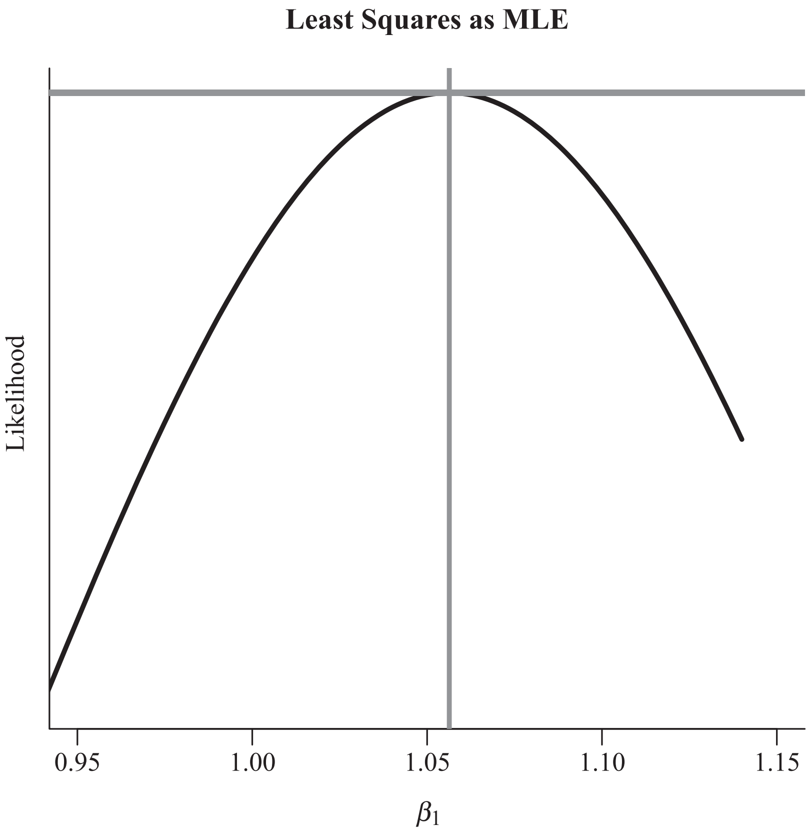

FIGURE 1.4 Likelihood as a function of possible values for β_1.

$\mathcal{R}$ Code Example 1.1 *Assembling the data*

```
library(WDI) #library for retrieving WDI data
wdi<-WDI(country = "all",
  indicator = c("EN.POP.DNST",           #pop density
    "EN.ATM.CO2E.KT",                    #CO2 emissions
    "NY.GDP.PCAP.PP.CD"),                #GDPpcPPP
  start = 2012, end = 2012, extra = TRUE, cache = NULL)
wdi<-subset(wdi, region !="Aggregates") #removing country aggregates
wdi<-na.omit(wdi) # Omit cases with NA; See chapter on missing data
names(wdi)[4:7]<-c("pop.den", "co2.kt","gdp.pc.ppp","wb.code")
attach(wdi)
```

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

It is easy to estimate the linear-normal regression in $\mathcal{R}$ using a maximum likelihood. In subsequent chapters, we will explain the heuristics used to interpret the MLE results and use $\mathcal{R}$'s many built-in functions to maximize the likelihood. But here we demonstrate how we might program the likelihood function directly and then pass it to one of $\mathcal{R}$'s numerical optimization routines. This requires a few steps:

Step 1: Set up data matrices.

$\mathcal{R}$ **Code Example 1.2** ***Step 1: Set up data matrices***

```
x.mat <- cbind(1,log(gdp.pc.ppp))  #note the column of 1s
y.vec <- log(co2.kt)
```

Step 2: Define the log-likelihood function, which can have several different parameterizations. Here we use $\log \mathcal{L}$ as expressed in Equation 1.3.

$\mathcal{R}$ **Code Example 1.3** ***Step 2: Log-likelihood function***

```
loglik.my <- function(par,X,y) {
  y <- y.vec
  X <- x.mat
  k <- ncol(X); n <- nrow(X)
  xbeta <- X%*%par[1:k]                          # matrix notation
  sigma <- sqrt(1/(n-k)*sum((y-xbeta)^2))        # assumed known here
  sum(-log(sigma)-(1/(2*sigma^2))*(y-xbeta)^2)   # log-likelihood
 }
```

Step 3: Pass the likelihood function to an optimizer, along with a vector of initial guesses for parameter values. Practically speaking, least squares estimates are often good starting guesses, but here we supply arbitrary initial values. For any iterative, numerical procedure it is a good idea to include a convergence check in your code, as shown. It is also important to store the Hessian matrix (see Chapter 2), if possible, for post regression analyses.

$\mathcal{R}$ **Code Example 1.4** ***Step 3: Call optimizer***

```
mle.fit <- optim(
   par=c(5,5),          # starting values
   fn=loglik.my,        # function to minimize
   method = "BFGS",     # algorithm choice
   control = list(
      trace=TRUE,       # trace optimization progress
      maxit=1000,       # max 1000 iterations
      fnscale = -1),    # changes minimizer to maximizer;
   hessian = TRUE)      # return Hessian matrix (see ch. 2).
if(mle.fit$convergence!=0)
   print("WARNING: Convergence Problems; Try again!")
```

Step 4: Now, we can calculate all the standard diagnostics. We provide code for making a table of regression estimates, including standard errors, z-scores, and p-values. To preview the theory introduced in Chapter 2, inverting[3] the Hessian matrix will provide the variance-covariance matrix, the square root of the diagonal of which contains the estimated standard errors for the parameters. The ratio $\hat{\beta}_i/\sigma_{\hat{\beta}_i}$ is the z-score (or asymptotic t-score).

$\mathcal{R}$ **Code Example 1.5** ***Step 4: Post-estimation analysis***

```
  # Calculate standard BUTON output.
  stderrors <- sqrt(-diag(solve(mle.fit$hessian)))
  z <- mle.fit$par/stderrors
  p.z <- 2 * (1 - pnorm(abs(z))) #p-values
  out.table <- data.frame(Est=mle.fit$par,SE=stderrors,Z=z,pval=p.z)
  round(out.table,2)
      Est   SE     Z pval
1 -0.30 1.21 -0.25 0.80
2  1.06 0.13  8.08 0.00
```

These results imply that, in the set of countries examined, higher levels of GDP are associated with higher levels of CO_2 emissions. The estimated coefficient for GDP (β_1) is approximately 1.04 (with a standard error of 0.13, resulting in a t-ratio of 8). This means that for every order-of-magnitude increase in GDP per capita, there will be an order-of-magnitude increase in annual CO_2 emissions, on average.

1.4.2 The Least Squares Approach

Another way to choose values for β_0 and β_1 is the least squares approach; the "best" values are those that minimize the sum of squared deviations from the prediction line. As above, each error is $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$. We minimize the total squared error, also known as the sum of squared errors (SSE):

Definition 1.1 (Sum of squared errors (SSE)).

$$\text{SSE} = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Comparing this approach to Equation 1.4 yields an important result: under certain conditions, minimizing the SSE is equivalent to maximizing the log-likelihood. Specifically, if the model for the mean is linear in the parameters and the outcome is assumed to be distributed iid normal, then the least

[3] The $\mathcal{R}$ function `solve()` calculates the matrix inverse.

TABLE 1.2 *Standard output for the OLS regression of log CO_2 emissions on log per capita GDP. Data drawn from World Bank Data Repository, 15 January 2017.*

	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	t-Ratio	p-Value
Constant	−0.30	1.22	−0.25	0.80
log per capita GDP	1.06	0.13	8.04	0.00
$n = 184$				
$R^2 = 0.26$				
$F_{1,182} = 64.6$				

squares estimator can be justified by the more general principle of maximum likelihood.

Minimizing the SSE (or maximizing the likelihood) is a standard mathematical procedure that involves taking the derivative of the SSE separately with respect to β_0 and β_1, setting both equal to 0, and solving the resultant set of equations. This analytically produces least square estimates of β_0 and β_1:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Generalizing the model to include k predictor variables in an $n \times k$ *design matrix*, $\mathbf{X}$, we can write the least squares estimator in matrix form as

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}.$$

Regressing log CO_2 on log per capita GDP using OLS in $\mathcal{R}$ produces the quantities displayed Table 1.2. The OLS estimates are identical to the MLE results calculated above. This table reports the standard $\mathcal{R}$ summary output for a linear regression, including t- and F-statistics, and p-values. These quantities will be absent from virtually all of the tables you see in the remainder of the book. We will go into details about why in subsequent chapters.

1.5 MAXIMUM LIKELIHOOD: GENERAL

Having seen two examples, we can restate the process of applying the likelihood principle in more general terms. In the likelihood framework, modeling begins with two statements. The *stochastic* statement describes our assumptions about the probability distribution (or distributions, in the case of a mixture) that govern our data-generating process. The *systematic* statement describes our model for the parameters of the assumed probability distribution. For example,

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

in the case with one independent variable and assuming a linear relationship between the independent variables and the mean parameter, we get:

$$\begin{aligned}\text{stochastic component} &: Y_i \sim f(\theta_i)\\ \text{systematic component} &: \theta_i = \beta_0 + \beta_1 x_i.\end{aligned}$$

Here, f is some probability distribution or mass function. We might choose the Gaussian, but it could be a variety of others such as the binomial, Poisson, or Weibull that we explore in this book.

Once we specify the systematic and stochastic components, we can construct a likelihood for the data at hand.

Step 1: Express the joint probability of the data. For the case of independent data and distribution or mass function f with parameter(s) $\boldsymbol{\theta}$, we have:

$$\begin{aligned}\Pr(y_1 \mid \theta_1) &= f(y_1;\theta_1)\\ \Pr(y_1, y_2 \mid \theta_1, \theta_2) &= f(y_1;\theta_1) \times f(y_2;\theta_2)\\ &\vdots\\ \Pr(y_1, \ldots, y_n \mid \theta_1, \ldots, \theta_n) &= \prod_{i=1}^{n} f(y_i;\theta_i)\end{aligned}$$

Step 2: Convert the joint probability into a likelihood. Note the constant, $h(\mathbf{y})$, which reinforces the fact that the likelihood does not have a direct interpretation as a probability. A likelihood is defined only up to a multiplicative constant:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) &= h(\mathbf{y}) \times \Pr(\mathbf{y} \mid \boldsymbol{\theta})\\ &\propto \Pr(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \theta_i)\end{aligned}$$

Step 3: Use the chosen stochastic and systematic components to specify a probability model and functional form. For a simple linear regression model, let f be the Gaussian (normal) distribution, and we have:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) &= h(\mathbf{y}) \times \prod_{i=1}^{n} f_{\mathcal{N}}(y_i \mid \theta_i)\\ \theta_i &= (\mu_i, \sigma^2)\\ \mu_i &= \beta_0 + \beta_1 x_i\\ \mathcal{L}(\mu, \sigma^2 \mid \mathbf{y}) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right]\end{aligned}$$

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press

Step 4: Simplify the expression by first taking the log and then eliminating terms that do not depend on unknown parameters.

$$\begin{aligned}\log \mathcal{L}(\mu, \sigma^2 \mid \mathrm{y}) &= \sum_{i=1}^{n} \log \left\{ (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-(y_i - \mu_i)^2}{2\sigma^2} \right] \right\} \\ &= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) \right. \\ &\qquad \left. - \frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2 \right] \\ -2\log \mathcal{L}(\mu, \sigma^2 \mid \mathrm{y}) &= \sum_{i=1}^{n} \left[\log(2\pi) + \log(\sigma^2) + \frac{1}{\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2 \right]\end{aligned}$$

The first term, $\log(2\pi)$, is simply a constant, unrelated to any of the parameters to be estimated, and can be dropped.

$$-2\log \mathcal{L}(\mu, \sigma^2 \mid \mathrm{y}) = \sum_{i=1}^{n} \left[\log(\sigma^2) + \frac{1}{\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2 \right]$$

Step 5: Find the extrema of this expression either analytically or by writing a program that uses numerical tools to identify maxima and minima.

1.6 CONCLUSION

This chapter presented the core ideas for the likelihood approach to statistical modeling that we explore through the rest of the book. The key innovation in the likelihood framework is treating the observed data as fixed and asking what combination of probability model and parameter values are the most likely to have generated these specific data. We showed that the OLS estimator can be recast as a maximum likelihood estimator and introduced two examples, including one example of a likelihood programmed directly into the statistical package $\mathcal{R}$.

What are the advantages of the MLE approach in the case of ordinary least squares? None; they are equivalent, and the OLS estimator can be derived under weaker assumptions. But the OLS approach assumes a linear model and unbounded, continuous outcome. The linear model is a good one, but the world around can be both nonlinear and bounded. Indeed, we had to take logarithms of CO_2 and per capita GDP in order to make the example work, forcing linearity on our analysis that is not apparent in the untransformed data. It is a testament to "marketing" that the OLS model is called the ordinary model, because in many ways it is a restricted, special case. The maximum likelihood approach permits us to specify nonlinear models quite easily if warranted by either our theory or data. The flexibility to model categorical and bounded variables is a benefit of the maximum likelihood approach. In the rest of this book, we introduce such models.

1.7 FURTHER READING

Applications

A great strength of the likelihood approach is its flexibility. Researchers can derive and program their own likelihood functions that reflect the specific research problem at hand. Mebane and Sekhon (2002) and Carrubba and Clark (2012) are two examples of likelihoods customized or designed for specific empirical applications. Other custom applications in political science come from Curtis Signorino (1999; 2002).

Past Work

Several other texts describing likelihood principles applied to the social sciences have appeared in the past quarter-century. These include King (1989a), Fox (1997), and Long (1997).

Software Notes

For gentle introductions to the $\mathcal{R}$ language and regression, we recommend Faraway (2004) and more recently Fox and Weisberg (2011) and James E. Monogan, III (2015). Venables and Ripley (2002) is a canonical $\mathcal{R}$ reference accompanying the `MASS` library.

https://doi.org/10.1017/9781316888544.004 Published online by Cambridge University Press